



TESIS DOCTORAL

2019

MULTIMODAL PERCEPTION OF ACOUSTIC PROMINENCE IN SPANISH

MIGUEL JIMÉNEZ-BRAVO BONILLA

**PROGRAMA DE DOCTORADO EN FILOLOGÍA. ESTUDIOS
LINGÜÍSTICOS Y LITERARIOS: TEORÍA Y APLICACIONES
DIRECTOR/A: DR./DRA. VICTORIA MARRERO AGUIAR**

U.N.E.D.

DEPARTAMENTO DE LENGUA ESPAÑOLA Y LINGÜÍSTICA GENERAL
FACULTAD DE FILOLOGÍA

**Multimodal perception of acoustic
prominence in Spanish**

MIGUEL JIMÉNEZ-BRAVO BONILLA

DIRECTORA: VICTORIA MARRERO AGUIAR

This dissertation was typeset using \LaTeX .

I, Miguel Jiménez-Bravo Bonilla, declare that the work presented in this thesis has been composed solely by myself and that it has not been submitted, in whole or in part, in any previous application for a degree. Except where stated otherwise by reference or acknowledgment, the work presented is entirely my own.

Agradecimientos

Hay ocasiones en que, ante situaciones difíciles, la constancia sobresale por encima de otras virtudes. Esto es lo que este proyecto me ha demostrado y es, quizá, lo más valioso de todo lo que he aprendido durante el transcurso de mi investigación. Mi familia lo ha vivido conmigo y a ellos les quiero agradecer su apoyo incondicional, pues saben bien los esfuerzos que ha conllevado llevar a buen término esta tesis doctoral. Muchas otras personas han tenido que soportar conmigo el peso de este trabajo y les doy las gracias por el esfuerzo compartido. Hoy se me figura, al mirar atrás, el cambio que ha operado en mí desde el comienzo. Ahora soy otro. Muchos otros cambios, algunos muy sutiles, han ido sucediendo y acumulándose. En la conciencia de cómo se han producido hay un gran aprendizaje y un disfrute sincero. Vendrán más cambios, sin duda, algunos tal vez como consecuencia de todo este trabajo, no lo sé; aunque en verdad no me inquieta en absoluto en qué dirección me lleven.

Quiero agradecer, también, de forma especial a Raúl Rojas por su autenticidad y porque él estuvo presente cuando dio comienzo este proyecto. A mi grandísimo amigo José Riva por su capacidad única de comprenderme, por su generosidad y porque vio cómo se gestaba el manuscrito final de esta tesis mientras compartíamos los días en Casa Santina. A Gea Gómez Pablos por la música inimitable que es y con la que sueña y por todos los momentos que tuvieron, y algún día tengan, un once sobre diez. A Óscar Esquivias y Rafael Eguílaz, cuya

amistad me inspira siempre tanto y con quienes deseo vivir una “vida futura” que sea hoy. A Yanina Carchak, por la lista de palabras donde la última entrada es “vínculo”. A Manu Sanz, con quien tengo la confianza de quienes comparten un mismo cielo. A Yasmina Alcántara, por el descubrimiento perpetuo. A Pepe Mas, porque su curiosidad es mi curiosidad. A mi Mai, con quien tengo la libertad de reírme de todo.

Igualmente me gustaría dar las gracias todas las personas que me han ayudado en algún momento de mi investigación, empezando por Victoria Marrero, José María Lahoz, Pilar Prieto y Jaydene Elvin; al departamento de lingüística del TiCC de la universidad de Tilburg, muy especialmente a Emiel Krahmer, Marc Swerts, Martijn Goudbeek, Ingrid Masson y Lieke van Maastricht; y finalmente también a Marieke Hoetjes y Louis ten Bosch por su tiempo y generosidad.

Contents

1	Introduction	1
2	Literature review: prominence and language	9
2.1	Acoustic correlates of prominence	9
2.1.1	Introduction	10
2.1.2	Prominence: lexical stress and pitch accents	11
2.1.3	The initial importance of f_0	13
2.1.4	An alternative: articulatory effort and intensity	15
2.1.5	Duration	17
2.1.6	Disentangling stress from accent	19
2.1.7	Acoustic correlates of prominence revisited	21
2.1.7.1	Reconsidering the role of intensity	21
2.1.7.2	Reconsidering the role of f_0	23
2.1.7.3	Reconsidering the role of duration	25
2.1.7.4	Reconsidering the perceptual threshold of prominence	28
2.1.8	Correlates of prominence in Castilian Spanish	29
2.1.8.1	Cross-linguistic differences	29
2.1.8.2	The confusion of the Spanish <i>acento</i>	31

Contents

2.1.8.3	The initial debate over the role of intensity	34
2.1.8.4	The fundamental role of f_0 observed in subsequent research	35
2.1.8.5	Different insights from other methodologies	38
2.1.8.6	Disentangling stress and accent in Spanish	41
2.1.8.7	Support for Navarro Tomás's <i>acento de intensidad</i>	42
2.1.8.8	Later research into the correlates of prominence in Spanish	43
2.1.9	Summary	46
2.2	Linguistic correlates of prominence	50
2.2.1	Introduction	50
2.2.2	Semantics, pragmatics and information structure	52
2.2.2.1	Prosodic prominence and information structure	53
2.2.2.2	Accessibility: repetition, frequency, and probability	60
2.2.3	Syntactic structures and prominence	63
2.2.4	Text-to-speech synthesis and prominence	71
2.2.5	Cross-linguistic differences	73
2.2.6	Summary	79
2.3	Gestural correlates of prominence	82
2.3.1	Introduction	83
2.3.2	The Kendon's continuum	83
2.3.3	The study of gestures until the 20 th century	87
2.3.3.1	Antiquity	87
2.3.3.2	The Renaissance and the Enlightenment	88
2.3.3.3	The 19 th century	90
2.3.4	Gesture studies in the 20 th century	93

Contents

2.3.4.1	The importance of spontaneous gestures	93
2.3.4.2	Kinesics and linguistics	95
2.3.4.3	Redefinition of gesture categories	96
2.3.4.4	Micro-analyses of gestures in relation to speech	98
2.3.4.5	Related hierarchy of gesture and speech	99
2.3.4.6	Gesture and the cognitive foundations of language	103
2.3.5	Categorising gestures	108
2.3.6	Head movements and facial expressions	114
2.3.7	Temporal coordination of gesture and speech	118
2.3.7.1	Precedence of gesture	119
2.3.7.2	More accurate measurements	120
2.3.7.3	On the applicability of the Phonological Synchrony Rule (PSR)	122
2.3.7.4	Redefining alignment landmarks	123
2.3.8	Audiovisual prosody	127
2.3.8.1	Multimodal prominence signalling	129
2.3.8.2	Studies using an animated agent	130
2.3.8.3	Studies in experimental settings	138
2.3.8.4	Studies using spontaneous speech	148
2.3.8.5	Gender differences in the audiovisual perception of speech	151
2.3.9	Summary	152
3	Methodology	156
3.1	Previous methodological approaches	156
3.1.1	Procedures	158

Contents

3.1.2	Speech materials and type of stimuli	161
3.1.3	Summary	163
3.2	Review of statistical methods	164
3.2.1	From ANOVAs towards Linear Mixed Models (LMMs)	165
3.2.2	LMMs and non-normal distributions: Generalised LMMs	175
3.2.3	Parameters estimation and statistical inference	179
3.2.4	Akaike Information Criterion and model selection	183
3.2.5	Summary	191
3.3	Methodology used in this study	194
3.3.1	Rationale	194
3.3.2	Speech material	197
3.3.3	Stimuli creation	199
3.3.4	Participants	201
3.3.5	Gesture annotation	202
3.3.6	Data analysis	204
3.3.7	Summary	207
4	Experiment I	209
4.1	Introduction	209
4.2	Methodology	211
4.2.1	Participants	211
4.2.2	Stimuli	211
4.2.3	Experiment design	213
4.2.4	Hypotheses	213
4.2.5	Procedure	214

Contents

4.3	Results	216
4.3.1	Descriptive statistics	216
4.3.1.1	Prominence marks	216
4.3.1.2	Inter-rater agreement	217
4.3.1.3	Prominence and acoustic cues	218
4.3.1.4	Prominence and visual cues	219
4.3.2	Inferential statistics	224
4.3.2.1	Correlation	224
4.3.2.2	Number of prominence marks	224
4.3.2.3	Model building and selection	228
4.3.2.4	Details of minimal adequate model <i>M18</i>	235
4.4	Discussion	242
5	Experiment II	251
5.1	Introduction	251
5.2	Methodology	253
5.2.1	Participants	253
5.2.2	Stimuli	256
5.2.3	Experiment design	256
5.2.4	Hypotheses	257
5.2.5	Procedure	258
5.3	Results	261
5.3.1	Descriptive statistics	261
5.3.1.1	Prominence marks	261
5.3.1.2	Inter-rater agreement	261

Contents

5.3.1.3	Prominence and acoustic cues	263
5.3.1.4	Prominence and visual cues	263
5.3.1.5	Prominence per sentence: P-score	265
5.3.2	Inferential statistics	272
5.3.2.1	Number of prominence marks	272
5.3.2.2	Model building and model selection procedure	275
5.3.2.3	Analysis A: Control vs. experimental conditions	275
5.3.2.4	Analysis B: Control condition ‘Exp0’	287
5.3.2.5	Analysis C: First condition ‘Exp1’ (intensity and duration)	295
5.3.2.6	Analysis D: Second condition ‘Exp2’ (f_0 and duration)	305
5.3.2.7	Analysis E: Third condition ‘Exp3’ (duration)	315
5.4	Discussion	328
6	General discussion and conclusions	343
	Appendices	358
A	Experiment I	358
B	Experiment II	372

List of Figures

1	Kendon's continuum	5
2	Nuclear pitch accents	12
3	Patterns of acoustic correlates in Spanish	30
4	Kendon's continuum in its relation to speech	84
5	Kendon's continuum in its relation to conventionalisation	84
6	Example of gesture performed by a speaker	85
7	Kendon's continuum in its relation to semiosis	87
8	Example of a gesture annotated by Kendon	101
9	Gestural and intonational hierarchies	103
10	Precision grip gestures described by Kendon	110
11	Possible head movements	116
12	Results of De Ruiter (1998)	122
13	Results of Esteve-Gibert and Prieto (2013)	126
14	Animated agent in Granström et al. (1999)	130
15	Animated agent in House et al. (2001)	131
16	Animated agent in Krahmer et al. (2002a,b)	132
17	Animated agent in Munhall et al. (2004)	134
18	Animated agent in Al Moubayed and Beskow (2009)	136
19	Results of Al Moubayed and Beskow (2009)	136
20	Animated agent in Prieto et al. (2011; 2015)	138
21	Stills of the recorded speakers in Swerts and Krahmer (2004)	139
22	Results of Krahmer and Swerts (2007, experiment 3)	142

List of Figures

23	Stimuli presented in Swerts and Krahmer (2008, experiment 2)	144
24	Stimuli presented in Foxton et al. (2010)	147
25	P-scores calculated in Mo (2008a)	160
26	Different possible structures for random effects	173
27	Statistical methods in psycholinguistics and corpus linguistics	175
28	Normal density curve	176
29	Logistic regression	178
30	Manipulation of f_0	200
31	Sample screen of ELAN annotation	203
32	I. Sample screen of experimental task	216
33	I. Occurrences of gestures according to articulator	220
34	I. Sample of gesture sequence	221
35	I. Correlation between prominence marks and sentence length	224
36	I. Significant differences per modality and condition	226
37	I. Significant differences per gesture phase	227
38	I. Odds ratios in model $M18$	236
39	I. Main effects in model $M18$	239
40	I. Interactions in model $M18$	241
41	II. Details of participants: age	254
42	II. Details of participants: musical training	255
43	II. Sample screens of experimental tasks	260
44	II. P-scores for target sentences	268
45	II. Significant differences per modality and condition	273
46	II. Significant differences per gesture phase	274
47	II. Odds ratios in model $G17$	280
48	II. Main effects of minimal adequate model $G17$	283
49	II. Odds ratios comparison between Experiment I and Experiment II	286
50	II. Odds ratios in model g_020	290
51	II. Main effects and interactions in model g_020	292
52	II. Odds ratios for competitive model g_016	293

List of Figures

53	II. Odds ratios in model $g_1 13$	298
54	II. Main effects and interactions in model $g_1 13$	301
55	II. Odds ratios for competitive model $g_1 12$	303
56	II. Odds ratios in model $g_2 15$	308
57	II. Main effects in model $g_2 15$	310
58	II. Interactions in model $g_2 15$	311
59	II. Odds ratios for competitive model $g_2 16$	312
60	II. Participants according to gender in Exp2	314
61	II. Odds ratios in model $g_3 18$	319
62	II. Main effects in model $g_3 18$	320
63	II. Interaction in model $g_3 18$	321
64	II. Odds ratios for competitive model $g_3 19$	322
65	II. Participants according to gender in Exp3	325
66	II. Odds ratios comparison of Experiment II	326
66	II. Odds ratios comparison of Experiment II	327
A1	I. Histograms of f_0	367
A2	I. Histograms of intensity	368
A3	I. Histograms of duration	369
A4	I. Plot of residuals in $M18$	370
A5	I. Informed consent form	371
B1	II. Sample screens with the instructions for the online experiment	377
B2	II. Histograms of f_0	380
B3	II. Histograms of intensity	381
B4	II. Histograms of duration	382
B5	II. Plots of residuals in models $G17$, $g_0 20$, $g_1 13$, $g_2 15$, and $g_3 18$	383

List of Tables

1	Acoustic correlates of prominence and perceptual cues	3
2	Results of Mahrt et al. (2012)	29
3	Results of Pamies (1997)	40
4	Results of Vogel et al. (2016)	46
5	Linguistic features in automatic detection of prominence	73
6	Summary of gesture classifications	113
7	Results of Ambrazaitis and House (2017)	150
8	Summary of previous methodologies	157
9	Fictitious data from a study on word recognition	167
10	Summary of estimation methods in GLMMs	181
11	References for Δ_i values	187
12	Hypothetical ranking of a set of models	188
13	Acoustic measurements	204
14	I. Experimental design	214
15	I. Cross-table of prominence marks and inter-rater agreement	218
16	I. Distribution of marks of prominence	218
17	I. Prominent words according to gesture phase	222
18	I. Random effects for $M9$ and $M10$	231
19	I. Summary of AIC results for models	234
20	I. Results of fixed effects in model $M18$	237
21	I. Results of fixed effects in model $M19$	241

List of Tables

22	II. Details of inter-rater agreement	262
23	II. Distribution of prominence marks	263
24	II. Prominent words according to gesture phase	264
25	II. Example of cumulative prominence marks	266
26	II. Summary of <i>AIC</i> results for models	279
27	II. Results of fixed effects in model <i>G17</i>	282
28	II. Summary of <i>AIC</i> results for models	289
29	II. Results of fixed effects in model <i>g₀20</i> (Exp0)	291
30	II. Results of fixed effects in model <i>g₀16</i> (Exp0)	293
31	II. Summary of <i>AIC</i> results for models	297
32	II. Results of fixed effects in model <i>g₁13</i> (Exp1)	299
33	II. Results of fixed effects in model <i>g₁12</i> (Exp1)	302
34	II. Summary of <i>AIC</i> results for models	307
35	II. Results of fixed effects in model <i>g₂15</i> (Exp2)	309
36	II. Results of fixed effects in model <i>g₂16</i> (Exp2)	312
37	II. Summary of <i>AIC</i> results for models	317
38	II. Results of fixed effects in model <i>g₃18</i> (Exp3)	318
39	II. Results of fixed effects in model <i>g₃19</i> (Exp3)	322
A1	I. Details of the 50 sentences from the corpus	363
A2	I. Details of prominence marks per speaker and listener	364
A3	I. Details of prominence marks for words and word categories	365
A4	I. Details of gestures per articulator	366
B1	II. Details of participants	375
B2	II. Details of target sentences used	378
B3	II. Details of prominence marks for words and word categories	379

Introduction

While speaking, humans do not only communicate through speech, but they also nuance, enhance, or disambiguate the information they convey by means of gestures. It is this interplay between both modalities, the auditory and the audiovisual, that gives language a fuller and more expressive dimension.

Gesture and speech are tightly integrated and both modalities belong to a single system of communication (Kendon, 2004). The synchrony of gesture and speech at a semantic, pragmatic, and phonological level is proof of this close connection (McNeill, 1992), and the deep roots that both modalities share might go back to early stages in the evolution of human language (Armstrong et al., 1995; Burling, 1993; Cartmill et al., 2012; Gentilucci & Corballis, 2006; McNeill, 2012; Pika et al., 2007).

It has also been shown that gestures have much in common with prosody in their potential for adding non-discrete nuances, thus serving interactive functions and facilitating comprehension, for which the term audiovisual prosody has been adopted (for a review see Krahmer & Swerts, 2009). Furthermore, gestures have also been found to possess similar prominence-increasing effects to those of speech (Swerts & Krahmer, 2008), and prominence marking is one of the many possible interactions between audiovisual prosody and verbal prosody.

The main topic of this investigation is prominence. As a perceptual concept, prominence is a bottom-up phenomenon whose perception by the listener depends on acoustic-phonetic, linguistic and contextual factors. The concept of prominence is extensively present in the literature, but unfortunately it is not possible to find a consistent definition of it. The word *prominence* often appears as a synonym of a great variety of other terms such as emphasis, lexical stress, nuclear accent, prosodic focus, pitch accent, intensity peak, etc., depending on the perspective and the research framework under which it is invoked. As a first attempt, the definition given by Terken and Hermes (2000, p. 89) might serve as a generic template from which it is possible to further build a more precise definition of prominence: “a *linguistic entity* is *prosodically* prominent when *it stands out* from *its environment* by virtue of *its prosodic characteristics*”. Thus, this template might be fine-tuned within a certain research framework by replacing the place-holders in italics with more precisely defined terms (Wagner et al., 2015). Within the phonetic perspective used in this research, the term prominence is equated with acoustic perceptual salience, so henceforth a word is said to be prosodically prominent when it is acoustically salient within a sentence (also known as phrasal stress, prosodic stress, or accent) by virtue of the interplay between the acoustic correlates involved in its production and perception: fundamental frequency (f_0), intensity, and duration. Nevertheless, as described in the literature, prominence can also refer to lexical stress—the acoustic perceptual salience of a syllable within a word—, and research into prominence as phrasal stress has often been addressed in relation to, and on occasions confounded with, lexical stress.

The acoustic correlates of prominence just mentioned correspond to the perceptual cues of pitch, loudness, and length, respectively, which are perceived by

the listener from the speech signal (Table 1). However, other elements can also be taken into account in an attempt to bring the complexity of the acoustic features of prominence to a more tangible and measurable reality, e.g. excursion and shape of f_0 , spectral tilt, etc. Throughout this work, if possible, a distinction will be made between the terms ‘correlate’ and ‘cue’, the former referring to production, the latter referring to perception. In the literature, however, the acoustic cues of prominence are often named by the respective correlates involved in their production, and instead of pitch, loudness, and length, the terms f_0 , intensity, and duration are preferred.

Perceptual cue	Acoustic correlate to be measured	Measure unit
Pitch	Fundamental frequency (f_0)	Hertz (Hz)
Loudness	Intensity	Decibels (dB)
Length	Duration	Seconds (s)

Table 1: *Relation between the acoustic correlates of prominence and their corresponding perceptual cues.*

The production of prominence by the speaker through a speech signal that contains different acoustic features is a bottom-up phenomenon whose perception by the listener is never categorical but continuous. Thus, the acoustic correlates are the physical realisation of prominence. The listener combines the information that the correlates provide with his or her expectations based on the knowledge they have of the language in order to decide which parts of the sentence are of special interest in the communicative process and to consider them as prominent.

From a linguistic-functional perspective, research on prominence mostly focuses on the linguistic functions it encompasses, e.g. information structure, contextual givenness, word order, etc. (e.g. [Baumann & Roth, 2014](#); [Bocci & Avesani,](#)

2011; D'Imperio, 1998; Vainio & Jarvikivi, 2006; Watson et al., 2008). In this case, prominence is conveyed by the speaker according to their syntactic and lexical knowledge and/or to their semantic and pragmatic intentions, and it is a common recourse used in many languages such as Spanish, English, or Dutch, for example, to structure information and disambiguate a message (e.g. Gundel & Fretheim, 2004; Féry & Krifka, 2008; Vallduví & Engdahl, 1996). However, in many other languages a different word order or an alternative formulation is preferred (e.g. Donati & Nespors, 2003; Leonetti & Escandell-Vidal, 2009; Szaszák & Beke, 2017; Szendrői, 2001). So in Spanish, for example, where clitic pronouns (e.g. *me*, *te*, *se*, etc.) do not often receive prominence, the sentence **Me vio* ('She saw me') is rather rendered by a different structure: *Me vio a mí / Fue a mí a quien vio* ('It was me she saw').

Apart from the literature on the acoustic correlates just mentioned, there exists a growing body of research on the visual correlates of prominence (e.g. Al Moubayed & Beskow, 2009; Granström et al., 1999; De Ruiter, 1998; Dohen & Løevenbruck, 2009; Krahmer et al., 2002a,b; Krahmer & Swerts, 2007; Leonard & Cummins, 2010; Loehr, 2007; Munhall et al., 2004; Prieto et al., 2011). Not only is the perception of prominence mediated by the interplay of acoustic correlates such as f_0 , intensity, and duration, but multimodal information in the form of eyebrow movements, head nods, and manual co-speech gestures also provide important cues to detect prominence.

Gesture is at our disposal, next to speech, as a medium of expression. However, the wide range of its expressive capacity cannot be easily pinned down to a fixed typology. Recent interest in the subject has resulted in an extensive criteria for classifying and dividing the phenomenon of gesture into different

types. In the literature, gesture is usually defined as spontaneous, often unwitting body movements accompanying speech and performed with fingers, hands, arms, eyes, eyebrows, face, head, or trunk. These movements are also known as *gesticulation* (Kendon, 1982) and differ in fundamental ways from the gestures performed in pantomime, from those known as emblems, and from the gestures used sign languages. According to their increasing degree of conventionalisation body movements can be classified in the so-called Kendon's continuum (McNeill, 1992)¹ (Figure 1). The Kendon's continuum will be dealt with in more detail in § 2.3.2.

Gesticulation → Emblems → Pantomime → Sign Language

Figure 1: *Kendon's continuum*.

At the left end, gestures appear to be holistic in their mode of expression and users rarely produce them consciously. At the right end, gestures of sign languages show compositionality and lexical structure, and users produce them conventionally to communicate. This study focuses on *gesticulation*, i.e. the spontaneous movements performed by speakers with different body parts while speaking. In keeping with the literature the most commonly used term of *gestures* will be used to refer to them. As for the term *co-speech gestures*, it is worth noting that it is predominantly used in reference to those gestures performed exclusively with the hands.

The perceptual effects of the acoustic correlates of prominence have tradi-

¹ With this name McNeill pays homage to Kendon, who first described this ordering of gestures in 1983 in a paper published in Kendon, 1988.

tionally been identified with those mostly produced by fundamental frequency (f_0) (e.g. Beckman, 1986; Fry, 1958; Gussenhoven et al., 1997; Pierrehumbert, 1980). However, some controversy exists around how the perceptual salience of a syllable or a word from their environment is produced by the participation of also intensity, and duration (e.g. Ortega-Llebaria & Prieto, 2011; Silipo & Greenberg, 2000).

This study aims at analysing how gestures and speech relate to each other in prominence perception. Recent research into the visual component of communication has started to cast light on the visual correlates of prominence and their interaction with verbal prosody (e.g. Al Moubayed et al., 2011; Beskow et al., 2006; Foxton et al., 2010; Granström et al., 1999; Kim et al., 2014; Krahmer & Swerts, 2007; Kushch & Prieto Vives, 2016; Prieto et al., 2015; Scarborough et al., 2009; Swerts & Krahmer, 2008). As a result, it has been observed that visual cues in the form of manual and facial gestures result in both stronger production and stronger perception of verbal prominence (e.g. Krahmer & Swerts, 2007; Swerts & Krahmer, 2008).

So, the questions addressed in this study are:

1. How do the different acoustic correlates relate to one another and to gestures in the perception of prominence?
2. How do gestures contribute to the perception of prominence?

This investigation also wishes to contribute in two more ways to the ongoing research conducted in the field. Firstly, studies on the multimodal perception of prominence exist for French (Dohen & Løevenbruck, 2009), Swedish

(House et al., 2001), Dutch (Krahmer & Swerts, 2004), and Catalan (Prieto et al., 2011, 2015), but research on Spanish is still pending. For this reason, this investigation has Castilian Spanish as its language of study. Secondly, most methods applied to date to the study of multimodal prominence perception have used animated agents (e.g. Krahmer et al., 2002a,b; Prieto et al., 2011) or experimental settings with controlled speech stimuli (e.g. Dohen & Løevenbruck, 2009; Krahmer & Swerts, 2007). Nevertheless, some limitations are inherent to both approaches, most notably, the ecological validity and the generalisation of results. In addition, perception studies with both animated agents and elicited gestures in controlled settings have limited themselves to analyse only certain gestures. Consequently, this might be the reason why the interaction of the acoustic correlates of prominence in multimodal perception is a research question that has not been addressed yet. Therefore, in order to complement the current state of the art and increase the ecological validity of experimental research, the present study use spontaneous speech material. Such speech material has been extracted from a television talent show and is employed as stimuli in two experiments involving prominence judgements by naïve listeners, i.e. listeners that have not been previously trained in the phonetics and phonology of Spanish prosody.

Previous studies on prominence perception have applied a binary prominence-marking task (prominent vs. non-prominent) for word pairs (House et al., 2001), short sentences with two target words (Krahmer & Swerts, 2007), or read-aloud sentences (Streefkerk et al., 1997). Similarly, some authors have conducted perception experiments with naïve listeners (e.g. Cole et al., 2010; Mo, 2008a,b). In this study on the multimodal perception of prominence in Spanish, naïve listeners conduct a marking task in which words are labelled in a binary scale (prominent vs. non-prominent). These words available for marking are presented in sentences uttered by different speakers engaged in a spontaneous conversation.

In summary, this study aims at understanding the role played by different acoustic correlates in the perception of prominence in Spanish, both in the presence and absence of visual cues in the form of gestures performed with hands, head, and eyebrows. The outline of this thesis is as follows: Chapter 2 reviews the literature on the acoustic, linguistic, and gestural correlates of prominence, respectively. Chapter 3 is devoted to review previous methodologies and statistical methods, and it describes both the preparation of the stimuli and the experimental task devised for the experimental part. Chapter 4 presents the pilot study on prominence perception conducted with naïve listeners in order to assess the stimuli and the methodology—its goal is to evaluate the validity of this new methodological approach and obtain some provisional results. Chapter 5 gives a detailed account of the second experiment, in which each experimental condition is analysed separately in order to answer the research questions. The conclusions are presented in chapter 6.

Literature review

Prominence and Language

2.1 Acoustic correlates of prominence

This section offers a brief review of some basic concepts stemming from prosody research and gives an account of the most significant acoustic landmarks that characterise the linguistic phenomenon of prominence. The initial importance ascribed to f_0 competed with the alternative view that it was rather vocal effort and intensity the correlates that were responsible for the acoustic realisation of prominence. The role of duration, as the third main correlate of prominence, is also discussed. This is followed by a review of studies that reconsider the relative role of f_0 , intensity, and duration, as a result of the partial confusion between the correlates of lexical stress and phrasal stress originated in prior experimental research. Finally, the section closes with a description of the most relevant studies dealing with prosodic prominence in Castilian Spanish.

2.1.1 Introduction

Prominence perception is a complex process that results from the interaction of several factors. Among them, prominence is perceived through the acoustic information of the speech signal produced by the speaker. The listener combines the information present in the speech signal with his or her expectations. The knowledge of the language by the listener allows them to decide which parts of the sentence are of special interest in the communicative process and can be considered prominent.

In order to study acoustic prominence a correlation must be found between its physical variables and the articulatory and perceptual criteria able to be measured (Table 1). In addition, experimental research must choose the appropriate speech element to conduct the measurements on: whether word, syllable, syllabic group, or phoneme; elements that are always perceived as prominent respect to their environment.

Research into the acoustic correlates of prominence has mainly focused on the acoustic correlates of fundamental frequency (f_0), duration, intensity, and to a lesser extent, spectral characteristics, and vowel quality. In addition, these acoustic correlates of prominence involved in the realisation of prominence have sometimes been assumed to hardly vary in the case of prominence perception. As Heldner puts it: “The reliability of acoustic correlates is not the same as the reliability of perceptual cues” (2003, p. 57). Differences in the production and perception of prominence are important, as are also important the differences between lexical stress and phrasal/sentential stress, two phenomena that have been confounded on certain occasions (e.g. Huss, 1978; Ortega-Llebaria & Prieto, 2007; Sluijter et al., 1997).

2.1.2 Prominence: lexical stress and pitch accents

In intonational languages, and differently from pitch-accent languages, where lexical differences rely on pitch contrasts, prominence occurs at least at a word level as well as at a phrase/sentence level. In the first case, stress is a phonological characteristic of the lexical item and marks the relative prominence of syllables within the word. For this reason, ‘stress’ is also known in the literature as ‘lexical stress’ or ‘lexical prominence’. Early work on the acoustic correlates of prominence focused on the acoustic differences of words such as *object* (noun) and *object* (verb), two words contrasting in meaning according to their different stress patterns. Such stress contrasts also exist in Spanish, for example, in verbs, which encode tense, person, and mode (e.g. paroxytone vs. oxytone words: *llevo* vs. *llevó*, or even proparoxytone vs. paroxytone vs. oxytone: *límite* vs. *limite* vs. *limité*)¹.

In the case of phrase/sentence stress, the stressed syllable of a word may also carry a pitch accent by virtue of either the position the word occupies or by possessing some special semantic or pragmatic value. A pitch change mostly involves a maximum or a minimum in f_0 . This phenomenon is also known as ‘accent’, ‘sentence stress’, ‘prosodic stress’, or ‘prosodic prominence’. For example, certain words can bear a pitch accent to emphasise their relative importance in the sentence in a contrastive way, e.g. *She was wearing YELLOW trousers* vs. *She was wearing yellow TROUSERS*. Both words *yellow* and *trousers* can bear a prominent pitch accent on their respective first syllables depending on which word is the object of focus.

Additionally, another realisation of prominence is produced by f_0 excursions,

¹ *Llevo* (1st person sg. present indicative, ‘to carry’) vs. *llevó* (3rd sg. person past indicative, ‘to carry’). *Límite* (noun, ‘limit’) vs. *limite* (1st / 3rd person sg. present subjunctive, ‘to limit’) vs. *limité* (1st person sg. past, ‘to limit’).

either as a rise or a fall, or as a combination of the two, associated to pitch accents in phrase boundaries (e.g. Beckman & Pierrehumbert, 1986; Pierrehumbert, 1980). The most common cross-linguistic pattern for f_0 excursions is to fall on the primary stressed syllable in the rightmost content word of an intonational unit, which if sentence-final, distinguishes between different intonation patterns, i.e. declaratives, yes-no questions, etc. For example, the two sentences *Julia is coming* (declarative) and *Julia is coming?* (surprise question) are different in their realisation of the final pitch contour: a final pitch fall in the first case, and a final sudden pitch rise in the second (Figure 2). When this pattern occurs, pitch accents, together with a boundary tone, are known as *nuclear pitch accents* and are obligatory components of intonational units.

In the study of the acoustic correlates of prominence the methodological distinction between word level and phrase/sentence level was not initially addressed (e.g. Fry, 1955, 1958; Bolinger, 1958; Huss, 1975; Nakatani & Aston, 1978; Sluijter et al., 1997; Ortega-Llebaria, 2006; Ortega-Llebaria & Prieto, 2011). The initial use of minimal noun/verb pairs (e.g. *object* vs. *object*) made the one-word-sentences of experimental tasks become the bearers of a pitch accent, which confounded the acoustic correlates involved in both types of prominence.



Figure 2: Same sentence as a declarative sentence (a) with an final pitch fall and as a surprise question (b) with a final pitch rise.

2.1.3 The initial importance of f_0

Fundamental frequency reflects the fundamental periodicity of the sound wave of speech in cycles per second, i.e. the lowest frequency at which the vocal cords vibrate, and it correspond to the first harmonic of the speech signal. It is usually abbreviated as f_0 and is measured in Hertz (Hz). Fundamental frequency is tightly related to the perception of pitch, and the melodic contour of an utterance can be set against measurable changes in the f_0 curve, yet there is not a linear relationship between f_0 and perceived pitch. It is also possible to conduct measurements using a musical scale and express Hertz in semitones. Nonetheless, other units of measurement are possible, such as equivalent rectangular bandwidth (ERB) and bark scale, for example, both reflecting actual auditory perception.

The influential research conducted by Fry in the 50's on lexical stress led him to establish that a higher f_0 was more relevant for the production and perception of lexical stress than other prosodic features as duration and intensity. In a first experiment, Fry (1955) measured duration and intensity of stressed vowels in minimal noun/verb pairs (e.g. *object* vs. *object*) and then analysed listeners' perception of synthesized words that varied along these parameters. Differences in responses given by listeners were used to measure the effectiveness of the varied parameters in cueing stress on the first syllable respect to the second one. Thus, Fry found that a great variation in intensity from paroxytone to oxytone words accounted for very few responses reporting stress on the first syllable. This was not the case, however, for duration, whose variation made listeners identify stress on the first syllable more consistently.

In a second experiment, Fry (1958) repeated the same procedure, but this time he compared f_0 and duration. By combining different patterns of duration and f_0 , he concluded that f_0 had an 'all-or-none effect', so that the syllable coinciding

with a higher peak f_0 or a f_0 movement was consistently perceived as stressed, while syllables with lower pitch accents were not. Since stressed syllables in stress languages usually bear pitch accents to which listeners are highly sensitive, Fry considered that “sentence intonation is an over-riding factor in determining the perception of stress and that in this sense the fundamental frequency may outweigh the duration cue” (1958, p. 151). Thus, he considered intensity as a weaker cue of stress than duration, and duration as a weaker cue than f_0 .

Additionally, several studies conducted by Bolinger (1955; 1958) following Fry’s work reinforced the idea that stress was mostly perceived as a result of different f_0 configurations and also considered intensity and duration as non-relevant cues for the perception of prominence. Bolinger used the term ‘pitch accent’ instead of ‘stress’ in order to offer a more comprehensive account of the phenomenon he studied. Bolinger developed a *pitch accent theory* and held that three main tonal configurations were responsible for perceived prominence. For him, differently from tenets held in the generative field (Chomsky & Halle, 1968), pitch accents were a property of intonation.

Similar ideas for the main role of f_0 were supported also by other authors (e.g. Beckman, 1986; Morton & Jassem, 1965; Pierrehumbert, 1980). Conversely, in the discussion initially sparked by Fry, studies by Lehiste and Peterson (1959) and Lieberman (1960) insisted on the interplay of f_0 , intensity, and duration in cueing lexical stress. For example, Lieberman defended that all three correlates had to be considered together. In a study using similar methodologies as those used earlier by Fry, Lieberman concluded that there was not possible to identify a single acoustic correlate of stress, but it was consistently rendered by greater duration and higher average of both f_0 and intensity.

2.1.4 An alternative: articulatory effort and intensity

Although intensity and duration were generally considered as less relevant than f_0 in the perception of prominence, the confusion between the correlates of lexical stress and phrasal stress contributed to give f_0 a preponderant role. However, around the same time Fry published his results, an alternative view was put forward, which defended that stressed syllables were distinguished from unstressed ones by means of the perceived articulatory effort their production involved (Fónagy, 1958; Ladefoged et al., 1958; Lehiste & Peterson, 1959; Navarro Tomás, 1964)². This view was reminiscent of a linguistic tradition dating back to the early 20th century, when the concept of *force accent* (also known in French as *accent d'insistence*) was associated with physiological force and was opposed to that of *melodic accent* (Sievers, 1901; Stetson, 1928; Sweet, 1877). In this sense, little after the publication of Fry's studies, Lehiste and Peterson (1959), conducted an experiment using vowels recorded both with the same vocal effort and with different vocal effort but with the same intensity. The researchers asked participants to judge the relative loudness of the vowels rather than their stress, and they concluded that listeners identified those vowels produced with a greater amount of vocal effort as louder than those with greater intensity. Similarly, Mol and Uhlenbeck (1956) found that reversing the intensity values of the stressed and unstressed syllables of a word did not affect listeners' stress perception.

Initially, the role of vocal effort competed with the importance given to f_0 by Fry and Bolinger. Although the perceptual effects of intensity were not deemed important enough at the time, more recent studies have reconsidered its role in the perception of prominence, and different researchers have insisted on it being

² Strictly speaking, this correlate is not acoustic but physiological and is responsible for transforming aerodynamic energy into acoustic energy. Its importance in this debate and its close relationship with the acoustic correlate of intensity justify its inclusion in this section.

a reliable correlate of stress (Beckman, 1986; Kochanski et al., 2005; Kohler, 2005; Lea, 1977; Ortega-Llebaria, 2006; Sluijter et al., 1997; Tamburini, 2003; Terken, 1991; Terken & Hermes, 2000; Turk & Sawusch, 1996).

Intensity, which is associated to variations in speech loudness, can also be seen as amplitude changes in the speech signal. Thus, *intensity* corresponds to the amount of energy present in a sound resulting from variations in the pressure of air coming from the mouth either at a pulmonic, glottal, or articulatory stage. Intensity is measured in decibels (dB), and different ways to capture its perceptual effect have been proposed: e.g. intensity maxima, corresponding to peaks of amplitude in the speech signal; or overall intensity, which combined peak amplitude and duration across the syllable (Beckman, 1986), which was later criticized for confounding both correlates (Sluijter et al., 1997).

Despite the fact that listeners have been observed to be able to perceive small changes in amplitude (e.g. Sorin, 1981), it has also been pointed out that intensity seems too vulnerable to noise and other environmental factors for it to have communicative significance. The conditions and the quality of the recording, the position of the speaker in relation to the recording microphone, and even the emotional content of the utterance have been claimed to affect intensity perception (Sluijter et al., 1997). The initial idea put forward by several authors that articulatory effort is more a reliable correlate of stress was reconsidered again in the 90's. For example, Sluijter and her colleagues (Sluijter & van Heuven, 1996a,b; Sluijter et al., 1997) suggested that it is actually the different distribution of intensity along the spectrum of frequencies that better captures the variation of loudness in the production and perception of lexical stress. According to them, "intensity in the mid-frequency range contributes more to perceived loudness than intensity above 5 kHz and, especially, below 0.5 kHz" (1996a, p. 2472), and reflects more realistically articulatory effort. Thus, *spectral tilt*, i.e. the distri-

bution of intensity throughout the spectrum (also known as ‘spectral slope’, or ‘spectral balance’), is the relation existing between the intensity found in the higher frequencies of the spectrum respect to the intensity found in the lower frequencies. Similarly, the term *spectral emphasis* is used with a very similar meaning, i.e. the difference between the overall intensity and the intensity in a low-pass filtered signal (Eriksson & Heldner, 2015; Eriksson et al., 2016; Heldner, 2001, 2003).

2.1.5 Duration

As previously seen, duration was considered by Fry (1955; 1958) as second in importance after f_0 but as a more reliable cue than intensity in signalling lexical stress. Conversely, some authors considered duration as the most robust cue of lexical stress (e.g. Cutler & Darwin, 1981; Isenberg & Gay, 1978).

Duration is tightly linked to the concept of *segmental length* (also ‘segmental quantity’) in metrical phonology theory, so that in many languages lexical stress depends mostly on the segmental composition of the syllable rhyme (nucleus, and optionally coda) (e.g. Fudge, 1969; Halle & Vergnaud, 1987; van der Hulst, 1985; Prince & Smolensky, 1993). In stress-timed languages, such as English (Pike, 1945), the sequence of a series of strong syllables together with weak syllables is called a *metrical foot* (Halle & Vergnaud, 1987; Liberman & Prince, 1977). Conversely, in syllabic-timed languages, such as Spanish, where there are no phonological differences in quantity/weight, this alternation does not exist, and each metrical foot has the same length. In addition, in English, for example, the unaccented vowels of weak syllables are subject to quantity reduction, which can eventually lead to their partial or full deletion (Delattre, 1966).

Therefore, how different language rhythms are realized as well as other features of a language’s phonological system influences how duration exerts an ef-

fect on both lexical and phrasal stress (e.g. Bagdasarian & Vanyan, 2011; Prieto et al., 2012; Ramus et al., 1999). For example, Prieto et al. (2012) compared English (a stress-timed language) with Spanish (a syllable-timed language) and also included Catalan, which is considered to possess rhythmic properties of both languages. In their study, after controlling for syllable structure, they found that differences in duration patterns for lexical stress were partly due to the rhythmic differences of each language. A similar conclusion was reached by Bagdasarian and Vanyan (2011) for phrasal in a study comparing Armenian and English.

Despite the fact that in many languages stressed and accented syllables tend to be longer than unstressed ones, it has been made clear that cross-linguistic differences exist. For example, in Thai, a tone language, lexical stress is signalled uniquely by duration (Potisuk et al., 1996) and not by f_0 , as in Polish (Dogil, 1999). In addition, a study comparing segmental duration of stressed syllables in Swedish, English, and French found different patterns of segmental lengthening for each of these languages (Fant et al., 1991). In French, lengthening was observed in both the stressed vowel and the preceding consonant, while in Swedish lengthening affected rather the stressed vowel and the following consonants. In English, this lengthening was more balanced between the vowel and both the preceding and following consonants. Besides, the duration of syllables was observed to be more uniform in French, a syllable-timed language, due partly to a shorter effect of the lengthening on the stressed syllable than in English and Swedish.

Another effect of the rhythmic and phonological properties of languages on prominence is vowel reduction. In many cases, and typically in stress-timed languages, unstressed syllables suffer vowel reduction. This is an effect of duration and intensity affecting vowel quality, so that reduced vowels tend to be shorter

and quieter than stressed open vowels (e.g. van Bergem, 1993; Fry, 1965; Ritveld & Koopmans-van Beinum, 1987). In English, for example, vowel quality has proved a more reliable acoustic correlate of stress than in Dutch (Sluijter & van Heuven, 1996a,b; Sluijter et al., 1997), a language that also has vowel reduction. It has also been observed that in post-focal unaccented contexts, the small differences in duration between stressed and unstressed syllables do not allow English listeners to perceive stress if it is not in the presence of reduced vowels (Beckman & Edwards, 1994; Campbell & Beckman, 1997; Huss, 1978).

Duration can also interact with other correlates of prominence. For example, when complex pitch configurations coincide on a single syllable, some languages tend to truncate their pitch contour, while others opt for compressing it in order to fully realise it (Grabe et al., 2000). From these different strategies it has been observed that syllables bearing a complex pitch accent are longer than those bearing a simpler one in languages resorting to compressing their pitch contour (Gili-Fivela, 2006; Prieto & Ortega-Llebaria, 2009).

2.1.6 Disentangling stress from accent

At some point it was evident that studies on prominence perception had used target words occurring in the focal position of utterances, a position where both word-level stress (lexical stress) and phrase-level stress (pitch accents) coincide in languages typologically classified as intonational languages such as English, Dutch, or Spanish. In this prosodic context, stressed syllables tend to attract pitch accents, while unstressed syllables do not. This type of co-variation between lexical stress and pitch accent was observed to have an effect on stress perception, and which may vary across languages, depending, for example, on the pitch accent distribution of each specific language (Hellmuth, 2007). The consequence of this confusion was that research into the acoustic correlates of prominence

confounded word-level and phrasal-level stress.

Some studies in the late 70's addressed the issue by analyzing the correlates of lexical stress in unaccented post-focal contexts (Huss, 1975, 1978; Nakatani & Aston, 1978). For example, Nakatani and Aston (1978) used a similar experimental paradigm to that of Fry. However, in their case, they embedded disyllabic pseudo-words preceded by accented adjectives within meaningful sentences. They observed that listeners perceived the manipulated stress patterns of the target pseudo-words by means of duration rather than of f_0 , while intensity proved useless as a cue to stress. Similarly, Huss (1978) observed that f_0 lost its effect as a correlate of lexical stress in unaccented contexts. Later, Beckman and Edwards (1994) examined words with stressed and unstressed vowels both in focal and post-focal contexts and concluded that duration was a consistent correlate of stress at word level, while f_0 changes were mainly associated to phrasal stress.

The evidence gained with these studies were taken into account in subsequent research and were later extended. Sluijter and van Heuven (1996a), for example, addressed the issue and contested the theoretical position held by Beckman and Edwards (1994) by stating that in stress-accent languages:

“Stress is a structural, linguistic property of a word that specifies which are the potential docking sites for accent placement. They have an accent-lending pitch movement associated with them when they occur within a single word in a narrow focus. In our view, stress is therefore determined by the language system, and accent by language behavior” (Sluijter & van Heuven, 1996a, p. 2471).

Therefore, subsequent studies on lexical stress in unaccented contexts for other languages made that pitch movement was seen as a correlate of phrasal

stress rather than of lexical stress. It was confirmed that stressed syllables were consistently longer than unstressed ones in the absence of pitch accents, and duration became then a more important correlate in the realisation and perception of lexical stress (for German, [Dogil & Williams, 1999](#); for Spanish, [Ortega-Llebaria, 2006](#); for Romanian, [Manolescu et al., 2009](#); for Dutch, [Sluijter & van Heuven, 1996a](#); [Sluijter et al., 1997](#)).

In addition, it was also reported that pitch accents are not crucial for lexical stress perception in both accented and unaccented contexts ([Ortega-Llebaria & Prieto, 2011](#); [Sluijter et al., 1997](#); [Turk & Sawusch, 1996](#)). As a consequence, the role of f_0 , intensity, and duration as acoustic correlates of prominence—both lexical stress and phrasal stress—were reconsidered.

2.1.7 Acoustic correlates of prominence revisited

2.1.7.1 Reconsidering the role of intensity

In several studies, Sluijter and van Heuven ([1996a](#); [1996b](#); [1997](#)) investigated the hierarchical relations among various acoustic correlates of both lexical and phrasal stress in Dutch and compared them with those of American English. They showed that the stressed syllable of unaccented words was cued acoustically through differences in duration and spectral balance, while overall intensity was a stronger cue of accent rather than stress—together with f_0 —in both languages ([Sluijter & van Heuven, 1996b](#)). Similarly, on the perceptual side, unaccented stressed syllables in Dutch and American English were perceived by means of longer duration, greater spectral balance, and absence of vowel reduction, even without the presence of a pitch accent ([Sluijter et al., 1997](#)).

In their turn, Kochanski and his colleagues ([2005](#)) also studied perceived prominence in different varieties of British English. In their experiment, four

expert listeners marking for binary prominence at a syllable level without distinguishing between lexical and phrasal stress were more sensitive to intensity and duration as acoustic cues of prominence than to f_0 . Furthermore, one important finding made by Turk and Sawusch (1996) was not only that duration and intensity are processed as a unit in stress judgements by listeners, but also that minimal variations of duration have a larger effect on the perception of loudness than minimal variations of intensity in the perception of syllable length. In this sense, it was observed that duration, or duration and intensity, can be correlates of lexical stress but not intensity alone.

In this sense, the different possible measurements reflecting perceived loudness have more recently been examined. Since the first proposals of articulatory effort as an alternative to f_0 , and due to the fact that the role of intensity has always proved inconsistent, several authors proposed different measurements to operationalise perceived loudness. Beckman (1986), for example, measured intensity as a combination of both peak intensity and duration across the syllable, which was criticized for potentially confounding both intensity and duration (Sluijter et al., 1997). Instead, it was suggested that spectral tilt (or spectral balance), i.e. the degree of intensity in the higher frequency regions in relation to that in the lower frequency ones, is a better correlate of lexical stress, while overall intensity is a stronger correlate of accent (Sluijter & van Heuven, 1996b; Sluijter et al., 1997).

Later, Heldner (2003), in an experiment to automatically detect focal words in Swedish, confirmed that, next to overall intensity, spectral tilt was also an acoustic correlate of accent. He also pointed out that overall intensity positively co-varies with f_0 , so that when f_0 increases so does intensity and vice versa. On the other hand, in two studies to also automatically detect phrasal stress in American English conducted by Silipo and Greenberg (1999; 2000), even the

role of f_0 as an acoustic correlate of accent was questioned. Both authors used two linguistically trained listeners that marked words on a 3-point scale, and they concluded that intensity as well as duration of vowels were stronger cues of phrasal stress than f_0 and f_0 range.

2.1.7.2 Reconsidering the role of f_0

After the initial importance given to f_0 and the subsequent research on the role played by other acoustic correlates of prominence, a traditional view has consistently defended the more crucial role of this correlate in cueing prominence in English and other languages (e.g. for English and Japanese, Beckman, 1986; Pierrehumbert, 1980; for Dutch, Terken, 1991).

This view has gone hand in hand with studies using speech resynthesis, with research into the automatic extraction and labelling of prosodic features, and with the development of text-to-speech synthesis (e.g. Gussenhoven et al., 1997; Gussenhoven & Rietveld, 1998; Kießling et al., 1996; Portele & Heuft, 1997; Silipo & Greenberg, 1999; ten Bosch, 1993; Terken, 1996; van Kuijk & Boves, 1999; Wightman & Ostendorf, 1994; Kohler, 2008). In such a research context, the theoretical basis of a model of intonation was necessary for better understanding how prominence is realized and perceived. Thus, the relationship between pitch accents and f_0 was addressed from different theoretical approaches, e.g. IPO intonation grammar ('t Hart et al., 1990), auto-segmental metrical (AM) approach to intonation (Pierrehumbert, 1980; Silverman et al., 1992), the Kiel Intonation Model (KIM) (Kohler, 1991; Kohler et al., 1997; Kohler, 2006), the Mixdorff-Fujisaki Model of German Intonation (MFGI) (Mixdorff, 1998; Mixdorff & Widera, 2001).

As a result, the study of prominence continued focusing almost exclusively on the perceptual effects of changes in f_0 . One of the findings made in the context

of this research concerned the perceptual declination of f_0 . By varying the height of pitch accents, it was observed that the initial words in an utterance have to have larger peak heights if they are to be perceived with the same prominence as the words at the end (e.g. Cohen et al., 1982; Gussenhoven & Rietveld, 1988; Pierrehumbert, 1979; Terken, 1996).

Additionally, Gussenhoven and Rietveld (1988) and Gussenhoven et al. (1997) examined the relation between maxima and minima of f_0 and the perception of relative prominence of accent in Dutch. They observed that the perception of accent is influenced by the distance between each pitch maximum and the progressive degree of declination of the baseline. The perceptual effect of this phenomenon was compared by Terken (1991; 1994) with the relative magnitude of changes in f_0 . Terken's analyses revealed that neither the difference between f_0 maxima nor the f_0 baseline can by itself drive the perception of accented syllables in both sentence final and sentence non-final position, but that prominence is actually cued by the complex interplay of both dimensions of f_0 .

After the studies conducted in the 70's to disentangle lexical from phrasal stress (§ 2.1.6) and the contribution of Sluijter and van Heuven (1996a; 1996b; 1997) (§ 2.1.7.1), two experiments conducted by Heldner and Strangert (1997) questioned the role of f_0 in the perception of phrasal stress. Both studies involved the manipulation of the f_0 contour in naturally produced sentences in Swedish. In the first study, the f_0 rise on focused words in phrase-medial position was reduced; in the second one, the size of the f_0 rise of non-focused word in phrase-medial position increased. Target words in narrow focus were embedded in answer sentences prompted by questions. The results obtained showed that phrasal stress can be perceived in the absence of an f_0 rise, and, by the same token, that an f_0 rise can conversely be perceived as non-prominent.

Nevertheless, Kohler (2008) obtained results for the role of f_0 that were differ-

ent to those of e.g. Kochanski et al. (2005). Kohler conducted a study in German with a methodology that was reminiscent of that of Fry's minimal noun/verb pairs. He synthesized the syllable *ba* and duplicated it in the disyllabic pattern *baba*, whose levels of f_0 , duration, and overall intensity he manipulated. The perception experiment that he conducted revealed, in a similar way to Fry's experiments with isolated words, that f_0 was the most important cue to induce listeners to shift prominence from the second to the first syllable, with a minor role of duration and intensity. The role of the latter two correlates in cueing lexical stress was later analysed by Kohler (2012) in another perception experiment using the word pair *Kaffe/Café* in German in an unaccented context within a carrier sentence. Kohler concluded that no hierarchy can be determined as to what correlate has a preponderant role. For him, it is the prosodic context created by each segmental sequence what determines the interaction of correlates.

2.1.7.3 Reconsidering the role of duration

Besides, Kohler carried out another experiment (2005), in which he analysed the acoustic correlates of phrasal stress focusing on a theoretical difference between pitch accents and *force accents* using the Kiel Intonation Model (KIM) as a framework. According to this model, next to a pitch accent, it is necessary to include a 'force accent' category capturing the greater physiological and articulatory effort characteristic of emphatic and emotional speech (e.g. Kohler, 2003). Related research conducted mainly in German has reported that force accents show a difference in segmental duration respect to pitch accents, with longer onset segments in force accents but longer syllable nucleus in pitch accents (e.g. Peters, 2005). From the results of his perception experiment, Kohler concluded that:

“Force accents constitute a separate accent category with at least three phonetic features – onset duration, energy, and voice quality –

in speech production, and that they are equally relevant in perception, albeit only duration has been formally tested, the relevance of the other two being deduced from the results.” (Kohler, 2005, p. 119).

In two successive perception experiments, Mo (2008a; 2008b) analysed the acoustic correlates of phrasal stress (and phrase boundaries) in English as perceived by naïve listeners. In her studies two elements stand out that set them apart from previous research and that are also relevant for this investigation. First of all, Mo used a corpus of English spontaneous speech (Buckeye corpus, Pitt et al., 2007). Secondly, the perception of phrasal stress was operationalized as the pooled transcriptions done in real time by 74 naïve listeners (see Mo et al., 2008, for details), so that words marked as prominent in the sentence was assigned a probabilistic P(prominence)-score corresponding to the probability of its prosodic perceptual salience (see e.g. Cole et al., 2010; Swerts, 1997, for similar methods). After measuring the acoustic values of several correlates of prominence, Mo (2008a) observed that the acoustic measures most strongly correlated with prominent words were duration and spectral tilt. Next to duration, Mo (2008b) also reported the uneven cueing effect that overall intensity had on prominence perception in the case of certain vowels. Finally, an important conclusion resulting from both studies was that neither duration nor intensity sufficed to cue phrasal stress in all the analysed prosodic contexts (14 stressed vowels), but that prominence was perceived by means of the interaction of various acoustic cues.

Finally, a series of studies conducted by Ortega-Llebaria and Prieto analysed the acoustic correlates of lexical and phrasal stress in Spanish and Catalan (Ortega-Llebaria, 2006; Ortega-Llebaria & Prieto, 2007, 2009, 2011; Ortega-Llebaria et al., 2007; Prieto & Ortega-Llebaria, 2006). Although the details of their research will be discussed in more detail in the section devoted to the correlates of prominence

in Castilian Spanish (§ 2.1.8.7), their results indicate that duration is a consistent correlate in the realisation of lexical stress in both accented and unaccented contexts in Spanish. The researchers reported that accented stressed syllables were longer than unaccented stressed syllables; and the latter were longer than unaccented unstressed syllables. Besides, they observed that different phones had different lengthening effects in stressed syllables, which also conditioned listeners' perception of lexical stress (Ortega-Llebaria, 2006; Ortega-Llebaria et al., 2007).

Furthermore, vowel quality and spectral tilt were found to play a role in cueing lexical stress contrasts, while overall intensity was a reliable correlate of accent in Spanish (Prieto & Ortega-Llebaria, 2006). However, they pointed out that the effect of overall intensity was independent of f_0 , i.e. higher overall intensity does not result from a positive co-variation with f_0 in Spanish (Ortega-Llebaria & Prieto, 2007) as it has previously claimed (Heldner, 2003; Mo, 2008a). The role of f_0 was found to be consistent in the realisation of pitch accents, so that while unaccented stressed vowels showed a flat pitch contour—together with longer segments than in unstressed syllables—the presence of an accent changed f_0 to a rising trajectory and also lengthened the syllable (Ortega-Llebaria, 2006). A similar conclusion was later reached for the lengthening effect of f_0 on syllable duration through the realisation of a complex pitch pattern (Prieto & Ortega-Llebaria, 2009)³.

³ However, in a different study, Ortega-Llebaria and Prieto (2007) reported that the realisation of a pitch accent did not necessarily involve the lengthening of the stressed syllable, and they concluded that “while duration is a crucial acoustic cue to mark a lower level prominence contrast (stressed vs. unstressed), it is a secondary (and thus optional) acoustic marker of a higher-level prominence contrast (accented vs. unaccented)” (Ortega-Llebaria & Prieto, 2007, p. 172).

2.1.7.4 Reconsidering the perceptual threshold of prominence

Later, Mo's collaborators (Mahrt et al., 2011, 2012) used the prominence transcriptions previously obtained by Mo (2008) from the Buckeye Corpus (Pitt et al., 2007) to put to the test the traditional assumption that the different acoustic correlates involved in cueing phrasal stress are perceived in a binary way. The P-scores obtained previously, which show the degree of prominence of a given word, were found to usually cluster around a low P-score or a high P-score end of a continuum. More precisely, Mahrt et al. found that not all correlates had the same partition point along this P-score continuum, but some positively correlated with a low, while others did so with a high P-score threshold. Thus, having different P-score thresholds, some of the tested correlates were best explained by different Gaussian distributions, i.e. two-Gaussian distributions with either a low or a high P-score threshold. For example, intensity was observed to be perceived in a binary way with a predominance of low P-scores (in the same way as normalized $\log f_0$). Conversely, raw f_0 values had a high P-score, and word duration was best modelled by a single distribution along the continuum (Table 2). Following their results, Mahrt et al. argued that some listeners might be more attuned to different acoustic features, but if perception is assumed to be consistent across listeners, then a more logical conclusion is that prominence is not binary, but either gradient or at least having a three-way distinction. Finally, the authors further suggested that this fact might be due to a different realisation of prominence in the cases of contrastive focus and broad focus.

Acoustic measure	Threshold
Log word frequency	Low
Duration of the last vowel	Low
Max intensity of the last vowel	Low
Root Mean Square (RMS) intensity of the last vowel	Low
Min intensity of the last vowel	Low
Min intensity of the stressed vowel	Low
Log mean f_0 of the stressed vowel (z-scores)	Low
Log max f_0 of the stressed vowel (z-scores)	Low
Stressed vowel duration	High
Max f_0 of the stressed vowel (z-scores)	High
Mean f_0 of the stressed vowel (z-scores)	High
Word duration	None
Log word duration	None

Table 2: Summary of Mahrt et al.'s results (2012) showing two-Gaussian distributions for the tested acoustic measures, with either a low or a high P-score threshold.

2.1.8 Correlates of prominence in Castilian Spanish

2.1.8.1 Cross-linguistic differences

As mentioned in the previous section, the acoustic correlates of prominence present cross-linguistic differences. It is generally accepted that both higher f_0 and larger f_0 excursions are associated to a prominence increase of syllables and words in many languages (e.g. Astésano et al., 2004; Barbosa et al., 2013; Beckman & Pierrehumbert, 1986; Dogil, 1999; Gussenhoven et al., 1997; Pierrehumbert & Beckman, 1988; Terken, 1991, 1994). However, in other languages, as in Italian, f_0 is lower throughout stressed syllables (Eriksson et al., 2016). In the case of Castilian Spanish, similar to Italian, a flat pitch contour, together with longer duration and stronger intensity, characterise stressed syllables in unaccented contexts, while accented stressed syllables are associated to higher f_0 ,

larger f_0 excursions and increased overall intensity (e.g. Ortega-Llebaria, 2006) (Figure 3).

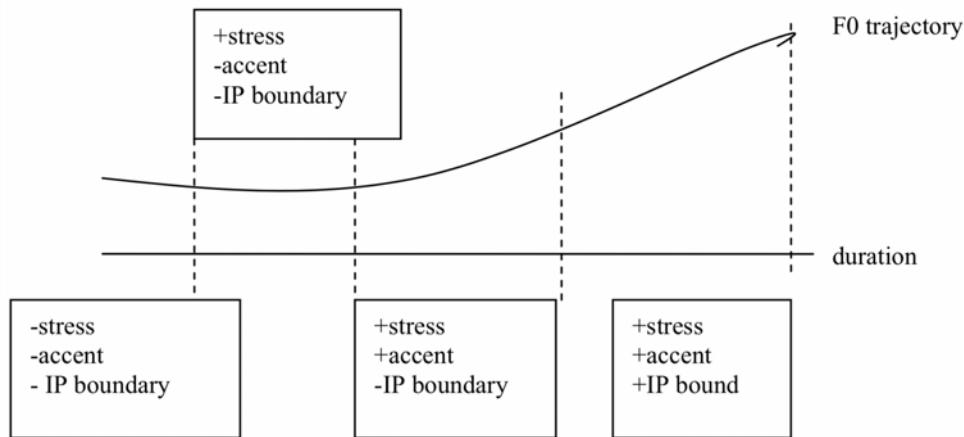


Figure 3: Trajectory of f_0 and patterns of duration lengthening for segments bearing stress, accent, and IP boundaries in oxytone words. Adapted from Ortega-Llebaria, 2006, p. 115.

Additionally, Spanish lacks the systematic reduction of unstressed vowels that is characteristic of English or Catalan (Ortega-Llebaria & Prieto, 2011), and although stressed syllables are typically longer than unstressed ones (Ortega-Llebaria & Prieto, 2011; Sluijter & van Heuven, 1996b), they do not contrast in length as much as in other languages, such as in Portuguese, another syllable-timed language (Ferreira, 2008). Although Spanish is characterised by the absence of vowel reduction in unstressed syllables, Ortega-Llebaria and Prieto (2007) found that unstressed [o] systematically becomes slightly more centralized than stressed [o], in agreement to previous results obtained by Quilis and Esgueva (1983). However, despite this weak use of vowel reduction in unstressed positions, Spanish do not use other acoustic correlates as a compensation in order to contrast between stressed and unstressed syllables (Ortega-Llebaria et al., 2007).

On the other hand, in the absence of the vowel reduction typically found in other languages, and when there are no pitch accents associated to the lexically-

stressed syllable in nuclear and pre-nuclear positions, lexical stress in Spanish is perceived by means of duration and intensity cues, similarly to Dutch. However, in perception Spanish speakers do not rely on spectral tilt like Dutch speakers do but on overall intensity and duration (Ortega-Llebaria et al., 2007; Sluijter et al., 1997). Nevertheless, the lengthening of duration cues by the effect of pitch accents seems controversial, since Ortega-Llebaria and Prieto (2007) claim that the presence of a pitch accent does not *consistently* trigger additive effects on the duration cues⁴, although Sluijter and van Heuven consider that it does in the case of English (1996a) and Dutch (1996b), while Beckman and Edwards (1994) pointed out that the lengthening pattern varied across speakers and speech rates.

2.1.8.2 The confusion of the Spanish *acento*

Descriptions of Spanish used to refer to the acoustic correlate responsible for cueing lexical stress with the term ‘acento de intensidad’, which was based on earlier theoretical accounts of the physiological force associated to articulatory effort. This was opposed to the contrastive effects resulting from pitch, which in Spanish was referred to as ‘acento de altura’, ‘acento melódico’, ‘acento tonal’, or even ‘acento contextual’. Researchers insisting on this difference followed closely the discussions held in other circles, as previously seen, in which ‘force accent’ was opposed to ‘melodic accent’ (e.g. Sievers, 1901; Stetson, 1928; Sweet, 1877). For example, Navarro Tomás (1918; 1964), along with some others (e.g. Cuervo, 1874; Gallinares, 1944), considered that an increase in local intensity was responsible for the Spanish ‘acento’, while other authors held that the ‘acento’ was realized by a slight lengthening of the stressed vowel, together with a small f_0 rise (Bello, 1847/1860; Real Academia Española, 1959). In the former

⁴ But see results of Ortega-Llebaria (2006) for oxytone words, which are summarised in Figure 3.

view, vowel lengthening and f_0 rise were often thought to be a consequence of the articulatory effort that the ‘acento de intensidad’ involved (Cuervo, 1874; Martínez Amador, 1954).

In addition, research into the acoustic correlates of prominence in Spanish has traditionally suffered from the same confusion between lexical stress and phrasal stress mentioned in the previous section (§ 2.1.6). Even more so since the Spanish word ‘acento’ could equally refer to *acento léxico* (lexical stress) or *acento tonal* (phrasal stress or pitch *accent*, a word with which ‘acento’ bears close resemblance)⁵. Thus, this confusion did not help determine more clearly the role played by the acoustic correlates of prominence in Spanish. For example, as late as the mid-80’s a more accurate definition of the Spanish term was called for so that research efforts become more effective:

“Está claro que un enfoque experimental del problema del acento requiere una formulación más rigurosa de la definición y utilización del término”⁶ (Solé, 1984, p. 138).

And a clarification of the term was offered in the subsequent lines, where it is equated with ‘lexical stress’:

“Por una parte empecemos por determinar que el acento es un fenómeno que se da en las sílabas, y que las sucesivas variaciones en la relación entre las sílabas acentuadas y no acentuadas constituyen el esquema rítmico de la frase, de la misma manera que las variaciones

⁵ Throughout this section, the Spanish word ‘acento’ is purposefully used to illustrate this ambiguity.

⁶ ‘It is without doubt that an experimental approach to the issue of the “acento” demands a more accurate formulation of its definition as well as a more precise use of the term.’

en las relaciones de tono forman el esquema entonativo”⁷ (Solé, 1984, p. 138-139).

Yet, although it was generally accepted that lexical stress and phrasal stress were two separate but interrelated phenomena, the precise nature of their relation was not clear:

“Cualquier intento de estudiar los factores en la percepción del acento sin tratar de contestar algunas preguntas sobre la entonación de las frases es incompleto. La pregunta más importante es si la entonación de frase es tan fuerte como para prevalecer sobre los otros factores en la percepción acentual”⁸ (Solé, 1984, p. 150).

A case in point of this confusion can be found in the first acoustic study on the correlates of prominence for Spanish, which was conducted by Bolinger and Hodapp (1961) following the work on English carried out by Bolinger himself a few years earlier (Bolinger, 1955, 1958). In their study, Bolinger and Hodapp set out to study the Spanish ‘accento’ by means of an experiment in which a series of sentences were recorded with different combinations of tone realisations and intensity levels for vowels occurring more than once in the sentence. In the experimental task, and according to the different acoustic cues applied in the stimuli, participants had to identify the narrow focus that disambiguated the

⁷ ‘For a start, let’s establish that “accento” (i.e. lexical stress) is a phenomenon of stressed syllables, and the continuous variations between stressed and unstressed syllables make up the rhythmic pattern of the sentence in the same way as variations in tone make up the intonational pattern.’

⁸ ‘Any attempt to analyse the elements involved in the perception of the “accento” (i.e. lexical stress) without trying to give an answer to some aspects of intonation is incomplete. The most important question is whether sentence intonation is so strong as to prevail over the elements responsible for stress perception.’

sentence they had been presented with. Bolinger and Hodapp, also with the help of spectrograms, concluded that the perceptual effects of f_0 (more precisely, changes in the f_0 trajectory) were a stronger correlate of ‘*acento*’ than intensity, and also duration. However, what Bolinger and Hodapp actually tested was the effects of the acoustic correlates of phrasal stress resulting from pitch accents.

More than thirty years later, when Solé (1984) conducted a study named “Experimentos sobre la percepción del *acento*” (‘Experiments of the perception of the *acento*’) (italics added), she surprisingly cited Bolinger and Hodapp’s study as the first acoustic study of this phenomenon in Spanish. However, differently from Bolinger and Hodapp, Solé analysed the perceptual effects of the acoustic correlates of *lexical stress*, for which she used synthesized pseudo-words in a similar experimental paradigm to that of Fry (1955; 1958).

2.1.8.3 The initial debate over the role of intensity

The work by Bolinger and Hodapp (1961) prompted a series of studies on the role of the acoustic correlates of prominence in Spanish. Later, Contreras (1963), drawing on Bolinger and Hodapp’s study, analysed the perception of lexical stress in Spanish in both isolated words and target words in short declarative sentences. Firstly, Contreras recorded three pairs of disyllabic words whose meanings depended on different levels of f_0 , intensity, and duration on the stressed syllable (e.g. *papa* vs. *papá*, *paro* vs. *paró*, and *pego* vs. *pegó*). After a first perception experiment, Contreras repeated the same procedure with the same word pairs, although this time she extracted the words from the carrier sentence they had been pronounced in. The results contradicted those of the first experiment and showed that the participants had mixed up some of the word pairs. For this reason, a third experiment was carried out in a similar way, but Contreras provided this time the context that the carrier sentences offered. In

her conclusion, Contreras made of f_0 the main correlate of lexical stress in Spanish, in line with Bolinger and Hodapp's initial, and she questioned the alternative view that the perceptual effects of the 'acento de intensidad' were more relevant than those of the 'acento melódico'.

Interestingly enough, the subsequent discussion over the role of f_0 and intensity was sparked off when one year later Navarro Tomás (1964) suggested that Contreras's results for the perceptual effects of f_0 actually reflected different intonation accents—which was replied shortly after by Contreras (1964)—, and he insisted on the idea that lexical stress in Spanish was actually mainly produced by an increase in local intensity.

2.1.8.4 The fundamental role of f_0 observed in subsequent research

After these initial studies on the acoustic correlates of the Spanish 'acento', subsequent research on the topic remained scarce for several decades. In this sense, the different views on the topic evolved together with the advances in phonetics. In Quilis's words:

“En una primera etapa [al acento] se le considera como un esfuerzo fisiológico y una impresión auditiva: ambos criterios en el marco de la fonética articulatoria (y auditiva). En el segundo estadio se buscan sus rasgos acústicos dentro de la fonética instrumental. Por último, se investiga sobre el papel de los índices acústicos controlando las variables que se puedan presentar por medio de la síntesis del lenguaje y juzgándolas en diversas pruebas de percepción: fonética experimental y psicofonética”⁹ (Quilis, 1981, p. 321).

⁹ 'In a first moment, [the Spanish "acento"] is considered a physiological effort and an auditory impression: both criteria belong to articulatory (and auditory) phonetics. Later, its acoustic features are sought within instrumental phonetics. Finally, research is conducted on its acoustic

Quilis himself addressed the acoustic realisation of lexical stress in Spanish with a first analysis of the vowel quantity in different contexts, including stressed and unstressed syllables (Quilis, 1965). In a later study on the production and perception of lexical stress, he focused on the acoustic values of stressed vowels using spectrograms, i.e. the first harmonic, vowel duration, intensity, and *área de intensidad* ('area of intensity')¹⁰ (Quilis, 1971). The results of these studies brought him closer to the views held by Bolinger and Hodapp (1961) and Contreras (1963; 1964), and he concluded that the most important correlate in both production and perception of the Spanish *acento* (in this case lexical stress) was f_0 , whether as higher f_0 or as a change in the f_0 trajectory, or even as the combination of both. Quilis, as Fry (1958) considered that “el índice más importante para la percepción del acento español es la frecuencia fundamental [...]. La duración sería el segundo componente”¹¹ (Quilis, 1981, p. 332).

Another perceptual study was conducted later by Solé (1984), as mentioned above. In a first experiment, she analysed the effects of duration and intensity in the perception of synthesized pseudo-words as either oxytone or paroxytone. With an experimental paradigm similar to Fry's (1955; 1958), Solé reported that

correlates—while controlling for the variables associated to speech synthesis—, which are analysed in different perception tests within experimental and functional phonetics.'

¹⁰ This measurement, which was also used in several other studies (e.g. Delgado-Martins, 1973; Santerre & Bothorel, 1969), consisted in editing the oscillogram on graph paper and counting the squared millimetres under all the amplitude peaks of a cycle. The rationale behind it intended to account for the perceptual differences that might arise by the interaction of different levels of intensity over time. A similar insight led later Beckman (1986) to put forward an alternative measure of intensity over duration that was criticized precisely for potentially confounding both correlates (Sluijter et al., 1997).

¹¹ 'The most important cue for the perception of the Spanish “acento” is fundamental frequency; duration is the second one.'

small differences in both acoustic correlates (only 5 dB in intensity and 10 ms in duration) were enough for listeners to change their judgment about stress from the first (paroxytone) to the second syllable (oxytone). However, she stated that both correlates, although sufficient for listeners to identify a change in stress pattern, were not decisive in cueing lexical stress, but others correlates not tested in this first experiment might also have played a role.

In order to gain more solid insight, Solé conducted a second experiment, in which, next to the previous two correlates, she also tested the perceptual effects of both f_0 levels and changes in the f_0 trajectory in vowels. Interestingly enough, based on work previously done by Delattre (1969), Solé considered as minimal the perceptual effects caused by vowel reduction in Spanish and did not explore further their role in cueing stress. She recorded three pairs of pseudo-words with different values for the correlates to be tested and administered them to the participants. Her results showed that f_0 —both as f_0 peaks and as changes in the f_0 trajectory—had a larger perceptual effect in cueing lexical stress when compared both one-to-one to the other tested correlates (intensity and duration) as well as in any possible combination with them.

Some subsequent studies also supported the role of f_0 as the main acoustic correlate of lexical stress. This is the case of a perceptual study conducted by Enríquez and his colleagues (1989) using synthetic speech. By manipulating f_0 , duration, and intensity in isolated words with different stress patterns (oxytone, paroxytone, and proparoxytone words), they reported that f_0 was the only correlate systematically used by listeners to identify stressed syllables. In addition, Enríquez and his colleagues concluded that stress identification was also partly due to the different stress patterns. They observed that duration was an inconsistent correlate of stress, especially since segment lengthening in oxytone words due to stress was perceived as a prepausal lengthening. Along these lines, it has

been suggested that listeners are sensitive to different stress patterns because stress perception in Spanish is partly driven by syllable weight (Face, 2000).

Similar results to those of Enríquez et al. (1989) for the preponderant role of f_0 were obtained by other authors (e.g. Figueras & Santiago, 1993a,b; Llisterri et al., 2003b, 2005). For example, Llisterri et al. (2003b; 2005) conducted a series of studies testing different stress patterns similar to the study of Enríquez and his colleagues, although Llisterri et al. used resynthesized words from natural speech instead of artificially synthesized words. They also observed that f_0 was the most robust cue, but only in combination with either duration or intensity.

2.1.8.5 Different insights from other methodologies

Nevertheless, the perceptual effect of f_0 in signalling lexical stress in Spanish was later questioned, especially in studies using methodologies different from those based on testing isolated words that differed in their stress pattern (e.g. Garrido et al., 1993; Pamies, 1997; Ruiz & Pereira, 2010). For example, Garrido and his colleagues (1993; 1995) compared the role of f_0 in two different contexts, namely, in read sentences presented both in isolation and within a paragraph. Their results revealed that f_0 peaks did not fall on the lexically stressed syllable, but their realisation was delayed, so that f_0 maxima were often aligned on the adjacent unstressed syllable:

“Los valores máximos de f_0 no parecen ser un correlato importante a la hora de marcar las sílabas tónicas en la lectura, puesto que un 80% de las sílabas tónicas no se corresponde con un valor máximo de f_0 . Cabe indicar también que los valores máximos de f_0 aparecen principalmente en la sílaba posterior a la tónica”¹² (Garrido et al., 1995, p. 189).

¹² ‘The maximum values of f_0 do not seem to have one specific correlate, since 80% of stressed

This phenomenon, known as peak delay (also f_0 shift or f_0 displacement), had been previously observed in Spanish (Navarro Tomás, 1944) and has been documented as well in other languages (e.g. Arvaniti et al., 1998; Gibbon, 1998; Jong, 1994; Xu, 1999)¹³. The insight that f_0 maxima are not consistently aligned with stressed syllables, together with the differences in duration they found between stressed and unstressed syllables, made Garrido and his colleagues (1993; 1995) conclude that duration was a more robust cue of lexical stress in Spanish than f_0 , a view that was also shared by other authors (Canellada & Madsen, 1987; Ríos, 1991).

Another view that questioned the preponderant role of f_0 in the Spanish ‘acento’ was the one held by Pamies (1997), who carried out acoustic measurements on the lexically stressed syllables of several utterances from a multilingual corpus that had been previously developed (Pamies, 1994). In his methodology, Pamies applied a coefficient of syllabic prominence resulting from comparing the three tested correlates measured in the stressed vowel with those of adjacent unstressed vowels. Thus, the positive coefficient obtained for each correlate—either alone or in combination—for a given syllable was interpreted as the proof that stress had been successfully realised.

syllables do not show a f_0 maximum. It is also worth mentioning that f_0 maxima mainly fall on the post-stressed syllable.’

¹³ This was first addressed in the tonal alignment literature partly due to the studies on intonation conducted by Bruce (1977) and Pierrehumbert (1980). In the resulting notation developed within the Auto-segmental Metrical (AM) framework and known as ToBI (Tone and Break Indices), peak delay was represented as the L*+H notation, in which the asterisk represents the alignment of a low tone with the stressed syllable, and the subsequent high tone represents the rising f_0 trajectory that occurs on the adjacent unaccented syllable. Although the realisation of f_0 maxima as a delayed peak is often caused by phonetic and contextual factors, cross-linguistic variation exists (see Gussenhoven, 2004; Prieto et al., 2005, for details).

Surprisingly enough, among the seven analysed languages, Pamies observed that negative coefficients for at least some of the correlates appeared associated with stressed vowels. This made difficult to consider just one correlate as the leading perceptual cue in the face of the other correlates, since each of them yielded negative coefficients at some point. For Spanish, stress was cued by only two correlates in 28% of the cases, and by just one of them in 44% of the cases. In addition, the no-realisation of stress was observed in 25% of the cases, i.e. the coefficient of syllabic prominence was negative for each acoustic correlate alone and for any possible combination of them. It is worth mentioning that Pamies excluded from his analysis those cases of stress realised both in sentence-final position and in positions adjacent to another stressed syllable. Finally, from his results he concluded that, even if Spanish stress is not the result of just one single acoustic correlate—but an *ad hoc* combination of more than one—, duration was found to be the most consistent of all three (Table 3).

Acoustic correlate	Alone	Combined
Fundamental frequency (f_0)	5%	14%
Intensity	10%	29%
Duration	28%	49%

Table 3: Summary of Pamies's (1997) results for the role of each tested correlate in cueing lexical stress in Spanish. Adapted from Pamies, 1997, p. 18.

Studies like these of Garrido (1993; 1995) and Pamies (1997) show that awareness of the interaction between stress and tonal peaks was slowly entering the debate over the correlates of the Spanish 'accento' (e.g. Pamies, 2003; Toledo et al., 2001), as it had previously occurred for other languages (e.g. Sluijter et al., 1997) (§ 2.1.6).

2.1.8.6 Disentangling stress and accent in Spanish

The different role played by the acoustic correlates of prominence both at lexical and phrasal levels was initially addressed in Spanish by Ortega-Llebaria (2006), who studied the production of lexical stress in three different contexts, namely, in unstressed words, in stressed unaccented words, and in stressed accented words. In her study, target pseudo-words were embedded in different carrier sentences that had to be repeated by five participants with the same intonation. In order to test the effect of accented and unaccented contexts, Ortega-Llebaria used declarative sentences with an Intermediate Phrase (IP) boundary occurring after the stressed accented syllable. She also used parenthetical sentences lacking any IP boundaries, in which the flat intonation and low pitch prevented stressed syllables from bearing a pitch accent. Ortega-Llebaria confirmed that in oxytone words lexical stress was produced by means of different acoustic cues and observed that a stressed syllable followed by an IP boundary suffered vowel lengthening and changed its f_0 level (Figure 3).

One important conclusion drawn from Ortega-Llebaria's study (2006), which puts in perspective several decades of research into the acoustic correlates of the Spanish 'accento', is that lexical stress should not be studied in isolation outside the prosodic structure where it occurs and which affects its realisation, especially because stressed syllables serve as anchor points for pitch accents. In this sense, it has been claimed that "typically (although not always), every lexically stressed syllable bears a pitch accent of some sort in a Spanish declarative sentence" (Hualde, 2002, p. 5), which contrasts with other languages such as European Portuguese (Frota, 2002).

Despite the fact that research on lexical stress in Spanish has benefitted from this insight (e.g. Llisterri et al., 2016), not many studies have further explored or tried to replicate Ortega-Llebaria's results (e.g. Ortega-Llebaria & Prieto, 2011;

Torreira et al., 2014; Vogel et al., 2016). For example, in their study, Llisterri and his colleagues (2016) complemented the results they had previously obtained for the acoustic correlates of lexical stress with isolated words (e.g. Llisterri et al., 2003b, 2005). This time they analysed listeners' perception of stress by comparing pseudo-words uttered in isolation with the same items extracted from the carrier sentences in which they had been uttered. The researchers observed that the perception of lexical stress largely depended both on the lexical stress pattern (oxytone, paroxytone, proparoxytone) and on the relationship between the acoustic parameters of the stressed vowel and those of adjacent unstressed vowels.

2.1.8.7 Support for Navarro Tomás's *acento de intensidad*

In addition to the results of Ortega-Llebaria (2006) already mentioned for Spanish, she also took up one of the questions concerning the perception of Spanish stress that had been raised in the early 60's, namely, the role played by intensity in signalling lexical stress. As previously discussed (§ 2.1.8.4), Bolinger and Hodapp's (1961) and Contreras (1963) supported f_0 as the main acoustic correlate of lexical stress. Little support had been found the idea of stress being prompted by an increase in overall local intensity, as proposed by Navarro Tomás (1918). However, Ortega-Llebaria and Prieto found partial support for Navarro Tomás's initial hypothesis (2007; 2011, Prieto & Ortega-Llebaria, 2006).

The follow-up study conducted by Ortega-Llebaria et al. (2006), actually offered evidence for the role of intensity in stress perception in Spanish. The researchers asked twenty listeners to identify oxytone words in unaccented contexts by means of the carrier sentence: *Hola —saluda (target word) contenta*¹⁴. In this case, the target words *mamá* and *mimí* varied along two oxytone-paroxytone

¹⁴ *Hello —says (target word) happily.*

continua: one of them using duration and overall intensity, and the other one, duration and spectral tilt as acoustic cues.

Ortega-Llebaria et al.'s findings (2007) showed that Spanish listeners ignored differences in spectral tilt in unaccented stressed syllables despite contrary results obtained in production studies: "Thus, in the absence of an accent, cues like duration and spectral tilt are crucial in the *production* of Spanish stress" (italics added) (Ortega-Llebaria & Prieto, 2007, p. 174). On the contrary, they observed that listeners detected stress contrasts on the basis of differences in duration and overall intensity between stressed syllables and unstressed adjacent syllables. This is not the case for Dutch, for example, where spectral tilt is preferred alongside longer duration to perceive lexical stress in unaccented contexts (Sluijter et al., 1997).

Furthermore, in the same study, Ortega-Llebaria and her colleagues also showed that vowel type influenced stress perception, so that oxytone words were more easily detected when the stressed vowel was [a] rather than [i]. Their interpretation was that duration and overall intensity are used in an additive manner to perceive stress contrast for vowel [a], whereas for vowel [i], duration is resorted to by listeners if overall intensity is not available.

2.1.8.8 Later research into the correlates of prominence in Spanish

The efforts to disentangle speech from accent in Spanish had led to use parenthetical sentences to provide unaccented contexts (e.g. *La masa del átomo es medible* —(determino / determino) *complacida*¹⁵, Ortega-Llebaria & Prieto, 2011). However, some doubts were raised about the effective control of intonation in such parenthetical phrases. On the one hand, it was argued that this kind of reporting sentences rarely appeared in conversational Spanish and, on the other

¹⁵ *The atom's mass is measurable* —I (confirm / confirmed) *satisfied*.

hand, the same unaccented context was better rendered by sentences where target words occupy a phrasal-medial position (Torreira et al., 2014; Vogel et al., 2016).

For example, in their study, Torreira et al. (2014) analysed whether the unaccented context provided in phrasal-medial position leads to the neutralisation of stress contrasts (oxytone vs. paroxytone) with sentences such as *Siempre que (miro / miró) la hora, ...*¹⁶. In a study that combined both production and perception experiments, they confirmed that lexical stress contrasts were maintained in the absence of a pitch accent, although they observed a considerable amount of phonetic overlap between stress patterns. Although longer duration and stronger intensity characterized the production of lexically stressed words as well as its perception, the available acoustic correlates were very similar between oxytone and paroxytone words:

“In production, roughly a quarter of the data could not be classified correctly by a model containing several features such as duration, intensity, voicing and F1; and, in the perception experiment, listeners made identification errors in a two-alternative forced-choice task in 37.1% of the trials” (Torreira et al., 2014, p. 200).

In this sense, Torreira and his colleagues concluded that phonetic contrasts signalling stress tend to be reduced in spontaneous speech, for which possibly contextual information is necessary to discriminate lexical stress patterns. The lack of robust cues that they observed was related to the predictability of stress, even if Spanish is a free-stress language.

Finally, a cross-linguistic study conducted by Vogel and her colleagues (2016) contradicted some of the previous studies on prominence production in Spanish.

¹⁶ *Every time* (I look up / she looked up) *the time, ...*

The researchers elicited a series of sentences in dialogues providing accented and unaccented contexts in order to analyse both lexical and phrasal stress. Measurements on a series of vowels (/i/, /a/, /u/) in target words included several acoustic measurements: mean f_0 , f_0 increase over the entire vowel, vowel centralization, mean duration, and mean intensity.

In the first dialogue the target vowel appeared the position of contrastive focus and therefore bore a pitch accent. In the second dialogue, by shifting phrasal stress to a post-focus position, the target vowel appeared in an unaccented context. In the examples provided here, the focused word is in boldface, and the target vowel, on which measurements were taken, is underlined¹⁷:

1. Phrasal stress [+accent / ± stress]:

Q: What did Maria say in the afternoon?

A: Maria said “CVCYCV” / “CVCYCV” in the afternoon.

2. Lexical stress [-accent / ± stress]:

Q: Did Maria say “CVCYCV” in the morning?

A: No, Maria said “CVCYCV” / “CYCV” in the **afternoon**, not in the morning.

Applying linear regression, Vogel and her colleagues reported the degree to which each specific context (i.e. ± accent / ± stress) was successfully predicted by their statistical model, and they specified the contribution of each of the correlates involved in the production of prominence (Table 4).

In the case of Spanish, lexical stress was found to be fundamentally cued by

¹⁷ Example adapted from [Vogel et al., 2016](#), p. 13. The actual Spanish dialogue was not provided by the researchers, who offered this English template as a model of the dialogues conducted in each of the tested languages.

Context	Predicted	Correlates
[+accent / +stress]	78%	Duration (73%), Intensity (65%), Vowel centralization (60%)
[+accent / -stress]	67%	Duration (60%), f_0 (60%), Vowel centralization (59%), Δf_0 (58%)
[-accent / +stress]	89%	f_0 (82%), Δf_0 (76%)
[-accent / -stress]	86%	f_0 (86%), Δf_0 (81%), Duration (55%)

Table 4: Summary of Vogel et al.'s (2016) results for both phrasal stress and lexical stress. Adapted from Vogel et al., 2016, pp. 17-18.

mean f_0 , which hardly varied between both accented and unaccented contexts, with a modest contribution of duration in the presence of an accent. In the case of phrasal stress in an unaccented context, duration was more important, with a small contribution of f_0 and vowel centralization, but reaching values slightly above chance level (50%) in the absence of a pitch accent. However, when a pitch accent was also present, f_0 seemed to play no role and prominence was cued by duration, intensity, and vowel centralization.

2.1.9 Summary

Acoustic correlates of prominence

For several decades research into the acoustic correlates of prominence has been linked to a very precise methodology that sought to determine what acoustic cues made listeners detect the stressed syllable within a word. Fry (1955; 1958) and Bolinger (1958) paved the way with their seminal experiments for subsequent research and stated that f_0 was more relevant for the production and perception of lexical stress than duration and intensity.

It was later observed that the target words used in such a methodological paradigm occurred at the focal position of utterances, so that lexical stress and the phrasal stress cued by pitch accents were confounded. Huss (1975; 1978)

and Nakatani and Aston (1978) observed that, in unaccented contexts, f_0 lost its perceptual effect, and duration was more relied upon by listeners to perceive lexical stress. The perceptual effects of f_0 were then considered to cue phrasal stress through the realisation of pitch accents. Studies in other languages different from English confirmed that duration was a more important correlate of lexical stress in the absence of pitch accents (e.g. Dogil & Williams, 1999; Ortega-Llebaria, 2006; Sluijter & van Heuven, 1996a; Sluijter et al., 1997).

Intensity was also observed to have a more fundamental role than initially thought, and it was reported to interact with duration in the perception of lexical stress in unaccented contexts (Turk & Sawusch, 1996; Sluijter et al., 1997). In addition, spectral balance—the intensity in the higher frequency regions respect to the that in the lower ones—was considered as a relevant acoustic correlate of prominence next to overall intensity and was observed to be crucial in cueing lexical stress in unaccented context together with duration in English and Dutch. In its turn, overall intensity cued accent (Prieto & Ortega-Llebaria, 2006; Sluijter & van Heuven, 1996b), together with f_0 , rather than stress. Similarly, on the perceptual side, unaccented stressed syllables were perceived by means of longer duration, greater spectral balance, and also by the absence of vowel reduction (Sluijter & van Heuven, 1996a,b; Sluijter et al., 1997).

These results were later confirmed by Kochanski et al. (2005), although in their study no difference was made between lexical and phrasal stress; and by Heldner (2003), who included spectral balance as a correlate cueing also phrasal stress. Subsequently, the role of f_0 was even questioned as an acoustic correlate of phrasal stress (Silipo & Greenberg, 2000), while other researchers using a methodology reminiscent of Fry's insisted on it being a reliable acoustic cue of lexical stress (Kohler, 2008).

Duration seemed to play a consistent role as a reliable correlate of lexical

stress in most studies; also its role as a correlate of phrasal stress was vindicated by several authors. For example, in sentences from a corpus of spontaneous speech, Mo (2008a) reported that duration and spectral tilt were the two acoustic correlates that drove prominence perception in naïve listeners. Ortega-Llebaria (2006), in her turn, observed that both f_0 and duration increased in the presence of a pitch accent in Spanish.

On a slightly different note, Kohler (2005) defended that duration is a strong correlate of what he dubbed as ‘force accents’, a separate category of pitch accents capturing the greater physiological and articulatory effort characteristic of emphatic and emotional speech. For Kohler, force accents are characterised by longer onset segments—but longer syllable nucleus in pitch accents—as well as by stronger energy and articulatory effort. In line with Kohler’s proposition, Mahrt et al. (2011; 2012) explored the possibility that in different contexts, prominence—e.g. contrastive focus, broad focus—might be produced and perceived through different acoustic correlates.

Correlates of prominence in Castilian Spanish

The acoustic correlates of prominence present cross-linguistic differences. In the case of Spanish, prominence is rendered by a flat pitch contour, together with longer duration and stronger intensity for unaccented stressed syllables; while accented stressed syllables are cued by longer duration, higher f_0 , larger f_0 excursions, and increased overall intensity (Ortega-Llebaria, 2006).

A similar confusion as that observed for other languages between lexical stress and phrasal stress determined the research into the acoustic correlates of prominence in the case of Spanish. For example, the important study by Bolinger and Hodapp (1961) aimed at analysing the correlates of *acento* (i.e. lexical stress), but it actually tested the perceptual effects of phrasal stress resulting from pitch

accents. Following into their footsteps, very much as in the debate conducted for English, most researchers defended the role of f_0 (e.g. Contreras, 1963, 1964; Martínez Amador, 1954; Quilis, 1971), while Navarro Tomás (e.g. 1964) supported the role of intensity, i.e. ‘acento de intensidad’ as the main correlate of stress.

More recent research has further explored the production and perception of lexical stress and similar conclusions in favour of f_0 have been made (e.g. Solé, 1984; Enríquez et al., 1989; Figueras & Santiago, 1993a,b; Llisterri et al., 2003b, 2005). Conversely, studies using a different methodology from previous research observed that: (a) f_0 peaks did not fall on the lexically stressed syllable, but were often aligned on the adjacent unstressed syllable (e.g. Garrido et al., 1993, 1995; Llisterri et al., 2003a), and (b) duration was a more important correlate of lexical stress than previously held (Garrido et al., 1993, 1995; Pamies, 1997).

The confusion between lexical stress and phrasal stress was addressed by Ortega-Llebaria (2006), who confirmed that lexical stress should not be studied in isolation outside the prosodic structure where it occurs. Following this insight, it was reported that Spanish listeners ignored differences in spectral tilt in unaccented stressed syllables (Ortega-Llebaria et al., 2007), differently from findings in English and Dutch (Campbell & Beckman, 1997; Sluijter et al., 1997) and despite contrary results obtained in production studies for Spanish (Prieto & Ortega-Llebaria, 2006). Rather, Spanish listeners detected contrasts in stress patterns using duration and overall intensity as cues of lexical stress (Ortega-Llebaria et al., 2007).

Subsequent research taking into account the prosodic context observed that the available acoustic correlates of duration and intensity were very similar between oxytone and paroxytone words in lexical stress contrasts, as observed in spontaneous speech (Torreira et al., 2014).

Finally, the preponderant role of duration, both as a correlate of lexical stress

and phrasal stress, was confirmed by Vogel et al. (2016), although they also observed that f_0 contributed to cue both stressed and unstressed syllables in the absence of pitch accents and that intensity helped duration in cueing accented stressed syllables, but not accented unstressed ones.

2.2 Linguistic correlates of prominence

This section firstly presents a summary of basic concepts from studies on discourse analysis and information structure and provides some context on the specific theoretical frameworks associated to them. Later, the role of prosodic prominence is discussed in relation to two distinctions commonly made in the literature: that of given vs. new information and that of (back)ground vs. focus. In addition, the concept of information ‘accessibility’ is presented as the third element affecting the acoustic realisation of prominence. The second part of this section deals with the way prosodic prominence is affected by the syntactic structures of utterances. Finally, cross-linguistic differences are presented with especial attention to the case of Spanish.

2.2.1 Introduction

Generally, some particular linguistic elements, such as syllables and words, stand out from their environment, and this phenomenon is determined by semantic and pragmatic factors and/or a syntactic and lexical component of the utterance.

Prosody has a perceptual effect on the listener’s processing of speech, so that it often conveys information about the pragmatic content and the meaning intended by the speaker. As previously seen, prosodic prominence signals the most salient element in the utterance, whose main role is generally accepted to be

marking the status of a constituent within the discourse.

It is necessary to mention that one of the key concepts on which the debate turns around in the literature on discourse analysis and pragmatics is *saliency* (e.g. [Chiarcos et al., 2011](#)). The theoretical basis of the term ‘saliency’ lies in the notion that those parts of the utterance that are perceived as being relevant in discourse planning and processing are more activated or accessible in memory than others. However, in phonetics, phonology, and prosody research, the term ‘prominence’ is more widely used to refer to a contextual relation between the prominent unit and its context. Thus, the term ‘prominence’ will be used throughout this section, even while discussing aspects related to discourse analysis, semantics and pragmatics.

In the case of syntax, the relationship between prosody and utterance structure has been addressed from different theoretical frameworks, as in the Auto-segmental Metrical model (e.g. [Ladd, 2008](#); [Pierrehumbert, 1980](#)). In this regard, initial research on the interaction between syntax and prosody, led Chomsky and Halle ([1968](#)) to claim that prominence was assigned to the sentence by means of phonological rules when its deep structure was transformed to a surface structure. Bolinger ([1972](#)), however, held the different view that there is no systematic relationship between the prosodic realisation of an utterance and its syntactic structure. He went on to argue that any theory attempting to deal with prominence assignment in purely syntactic terms was bound to make vacuous claims.

Nevertheless, syntactic and lexical correlates of prominence have been later used, most notably, in text-to-speech synthesis. In this line of research, one of the key concepts is that of ‘predictability’. Top-down expectations are based on the linguistic competence of the listener, and the syntactic marking of prominence is mainly predicted by word class, word length, and the position a word occupies in the utterance.

2.2.2 Semantics, pragmatics and information structure

Prosody is used differently across many languages to signal prominence, and it is structurally linked to the *information structure* of the utterance (also known as ‘information packaging’). In its turn, the information in an utterance is organised in relation to the *discourse*, i.e. any “coherent multi-utterance dialogue or monologue text” (Kruijff-Korbayová & Steedman, 2003, p. 249). In this sense, *information structure* can be defined as the way in which each new utterance relates to an existing discourse, e.g. by altering it or updating the already given information. In order to give a successful interpretation of an utterance it is not only necessary to interpret its logico-semantic meaning—propositional meaning—, but also its informational meaning—extrapropositional meaning. For example, alternative syntactic structures with different constituent order and/or intonational structure may express the same propositional content, but “they are not interpretively equivalent in absolute terms, but rather add some extrapropositional contribution to meaning” (Vallduví & Engdahl, 1996, p. 459).

- (1) a. Eva played the piano splendidly at the concert.
- b. EVA played the piano splendidly at the concert.
- c. It was Eva who played splendidly the piano at the concert.

For example, sentence (1a) is syntactically identical to (1b), but the realisation of a pitch accent on ‘Eva’ (as showed by the small caps)¹⁸ adds a different extrapropositional meaning, namely, that Eva, and no other person, played the piano splendidly at the concert. Similarly, the same effect is achieved in (1c) by means of an initial *it*-cleft clause.

¹⁸ Following the literature, in the following examples small caps reflect prosodic prominence through the realisation of a pitch accent.

The relationship and interaction of discourse structure and information structure is complex, and yet “a theory relating IS [information structure] and DS [discourse structure] is essential for accurate natural language processing” (Kruijff-Korbayová & Steedman, 2003, p. 250).

Despite the abundant approaches and the different terminology, all attempts to describe information structure make a basic distinction between a more informative part of an utterance and a less informative one. Basically, the nature of this split and the place where it occurs in the utterance are the matter of contention among different theories. According to Vallduví:

“[...] [I]t could be said that information is concentrated on a subpart of the sentence, while the remainder is licensed only as an anchoring vehicular frame for that informative part to guarantee an optimal entry into the hearer’s knowledge-store” (Vallduví, 1992, p. 35).

Different theoretical approaches to information structure have referred to this basic distinction by means of different opposing terms, e.g. *theme/rheme* (Firbas, 1964; Halliday, 1967b), *topic/comment* (Hockett, 1958; Gundel, 1974), *topic/focus* (Sgall & Hajičová, 1977, 1978), *focus/presupposition* (Jackendoff, 1972) or *focus/open-proposition* (Prince, 1981).

2.2.2.1 Prosodic prominence and information structure

It is generally accepted that information structure and discourse status interact, although there is a considerable debate over the exact nature of this interaction. Nonetheless, in many languages prosody plays a crucial role in this process by marking the information status of words and constituents in the utterance. For example, Bolinger (1958; 1961) had remarked that pitch accents serve

to mark semantic contrasts, and he introduced a distinction between a fall-rise accent and a falling accent as a way to contrast information in English.

- (2) {What about Andrea? Who did she come with?}
[_T ANDREA] [_C came with [_F LUCAS]].

This difference in intonation was related by other authors to topic and focus, respectively (e.g. Jackendoff, 1972), as in (2), where the topic ‘Andrea’ is uttered together with a fall-rise pitch accent, and ‘Lucas’ is pronounced with a falling accent for being the focused item within the comment.

Given/new distinction

Halliday (e.g. 1967a; 1967b) also drew attention to the role of pitch accents in signalling the basic division between given and new information in English. According to him, words carrying a pitch accent are commonly seen as referring to new information, whereas words that are perceived as information already given in the discourse are said to be ‘deaccented’. However, Halliday (1967b, p. 204) considered new information as that which “the speaker presents [...] as not being recoverable from the preceding discourse”, regardless of whether its referent has been mentioned before.

Experimental research has observed that speech processing was enhanced when new information was provided to listeners by a co-occurring pitch accent. For example, Cutler (1976) and Cutler and Foss (1977) confirmed that the processing speed of new information in a focal position was higher if it was accented. Additionally, Terken and Nootboom (1987) found that not only speed processing, but also comprehension improved when this accentuation pattern was applied. Similarly, a perception experiment on the evaluation and comprehension of sentences conducted by Birch and Clifton (1995) found that, for

simple question-answer pairs, listeners consistently considered more prosodically appropriate the pattern of accented new information and deaccented given information.

Nevertheless, it has also been observed that previous mention of a word in the discourse is not sufficient for already conveyed information to be deaccented, especially when the simple sentences used as examples in a theoretical framework are confronted to the spontaneous speech samples of more empirical approaches. For example, in a study analysing task-oriented speech, Brown (1983) reported that, although speakers tended to place pitch accents on new information, an item of given information may be re-introduced into the discourse after some digressions and be marked by a pitch accent.

Furthermore, in an analysis of a series of corpora for the development of text-to-speech systems conducted by Hirschberg (1993), it was also observed that, contrary to expected, previously mentioned information was often found to be highlighted through a pitch accent. Later, Terken and Hirschberg (1994) extended this observation by studying the deaccentuation of given information in relation to the syntactic position of an item in the utterance and its grammatical function.

They reported that speakers were likely to accent an item of given information if both grammatical function and syntactic position were different from the antecedent in the immediate context. Conversely, the information item was more likely to be deaccented if it occupied the same position and had the same grammatical function as its antecedent.

Another study by Dahan et al. (2002) exploited a visual paradigm using eye-tracking to analyse the perception and comprehension of pitch accents in the given/new distinction. In this paradigm, participants are usually presented with a series of items, and the verbal instructions they receive allow researchers to track the fixation of their eyes. These instructions are manipulated according to

the purpose of the study, so that this input is crucial for participants to complete the task.

In their case, Dahan and her colleagues used this paradigm to study on-line comprehension of accented information. They asked participants to move two different objects to different locations on a screen. Two of these objects shared their initial segments, e.g. ‘candy’ and ‘candle’. By manipulating the given/new status of one of the items in the instructions provided to participants (3), they observed how pitch accents were processed to resolve the introduced ambiguity.

- (3) a. Put the candle/candy below the triangle.
- b. Now put the CANDLE/candle above the square.

Dahan et al. observed that participants tended to fixate their eyes more often on a new referent when it was uttered with a pitch accent. Conversely, when the critical target word was deaccented, they fixed their eyes in the given referent instead. The measured eye reactions and the differences between both conditions revealed that pitch accents are rapidly detected and quickly integrated into the discourse representation.

Interestingly enough, deaccentuation of an utterance item that has the potential to carry a pitch accent can also be used by speakers on some occasions with an informative purpose. For example, it has been observed that information that has been previously mentioned or information that can be easily inferred—and is not contrastive—shows weaker prominence than expected, or it completely lacks prominence cues (e.g. Féry & Kügler, 2008). This is often exploited with a humorous effect¹⁹:

- (4) A: I heard you had to call a plumber over to your house this morning!
- B: Yeah, he’s still here and I’m ready to MURDER the ORANGUTAN!

¹⁹ Example adapted from Wennerstrom, 2011, p. 317.

The last word in (4), 'orangutan', is deaccented (as shown by the subscript text), although it would normally receive the nuclear pitch accent of the sentence. The humorous effect results from the listener's conclusion that this information should actually be understood as already introduced in the discourse, and thus, a co-referential relation between the plumber and the animal is created.

Focus

The given/new distinction is tightly related to that of (back)ground/focus. The term 'focus' was initially used by Halliday (1967b) and was introduced later in the generative literature by Jackendoff (1972) as part of the focus/presupposition distinction. In Jackendoff's work, 'presupposition' "denote[s] the information in the sentence that is assumed by the speaker not to be shared by him and the hearer" (Jackendoff, 1972, p. 230).

According to Ladd (2008, Chapter 6), it was this notion of 'focus' developed in the framework of the generative syntax (Jackendoff, 1972) which was adopted, together with Pierrehumbert's (1980) definition of 'pitch accent' based on Bolinger's work. In this sense, Vallduví and Engdahl (1996, p. 462) define 'focus' as "an informative, newsy, dominant, or contrary-to-expectation" part of the utterance as opposed to the expected one²⁰. Gussenhoven (1983a) coined the expression 'Focus-To-Accent' (FTA) approach to refer to the attempts to narrow the gap between the semantic/pragmatic notion of 'focus' and the phonetic/phonological notion of 'accent'. In FTA, the key issue turned around the relation between which parts of the utterance are focused and how a given pattern

²⁰ However, another tradition in the literature (Grosz & Sidner, 1986) employs the term 'focus' in a different way: a discourse entity is considered 'in focus' when it is the topic of the conversation, i.e. 'in focus' is equivalent of 'given' information rather than 'new' information and, thus, it is likely to be deaccented.

of focus is conveyed by the location of the accent. However, Ladd remarks that:

“The speaker’s decision about which word or constituent to focus is subject to all kinds of contextual influences which are at best poorly understood: these are the factors with which Bolinger, Chafe, Halliday, and others have always been concerned. However, once we specify the focused part of the utterance [...], the location of accent on a specific word within the focused constituent follows more or less automatically by language-specific rules or structural principles” (Ladd, 2008, p. 218).

Actually, there exists a strong correlation between metrical structure and the realisation of a nuclear pitch accent on a word introducing new information or a word with narrow focus. In Cole’s terms:

“A word located in a strong position in the phrase-level metrical structure is *prominent* relative to a word in a weak (or hierarchically lower) position, and the rightmost prominent word in the phrase is the head or nucleus of the phrase—unless a preceding word has a narrow or contrastive focus, in which case the focused word is assigned the nuclear prominence [...]. The nucleus is assigned an obligatory pitch accent. Additional pitch accents are optionally assigned to pre-nuclear words [...] based on their informativeness” (Cole, 2015, p. 9).

Thus, despite the role played by phonological rules in accent assignment, the notion ‘informativeness’ that Cole refers to makes that the location of a pitch accent is ultimately dependent on the context, as Ladd pointed out. For example:

- (5) We saw a tiger on the ROAD.
- (6) a. A: What did you see on the road?

- b. B: #We saw a tiger on the ROAD.
- c. B: We saw a TIGER on the road.

In sentence (5)²¹, the realisation of a pitch accent as a nuclear pitch accent indicates the focused constituent of the utterance, in this case as narrow focus. However, this leads to an infelicitous answer in (6b), where the focus shifts to ‘tiger’ according to the context provided in (6c).

An important aspect discussed in research is how focus interacts with syntax and phonology, in what is known as *focus projection* (Gussenhoven, 1983b; Selkirk, 1984). For example, it was observed that the nuclear accent occurring towards the end of a phrase may not only highlight the information conveyed by the accented word, but it can also work as a prosodic head that ‘projects’ to larger units, even to the whole verbal phrase (Birch & Clifton Jr., 1995). This is most evident in cases of broad focus.

- (7) a. A: What happened?
B: The bus driver [_F stopped at the STATION.]
- b. A: Where did the bus driver stop?
B: The bus driver stopped at [_F the STATION.]

For instance, in (7a)²², broad focus is realised through a nuclear pitch accent on the word ‘station’, which projects to the whole verbal phrase ‘stopped at the station’. In (7b) narrow focus is realized also through a nuclear pitch accent, but due to the context provided by the question, only the information about the place of the action is new to the hearer; consequently, there is no focus projection. Besides, both intonational contours are likely to differ, and in a natural conversation

²¹ Example adapted from Féry & Krifka, 2008, p. 124.

²² Example adapted from Falk, 2014, § 2.2.1.

the responding person in (7b) would probably just give the requested information: ‘at the station’.

In the case of contrastive focus, for example, it is assumed that a pitch accent usually falls on the single word being the contrasting information to a set of alternatives. However, according to Zimmermann (2007, p. 154), contrastive focus rather “express[es] a contrast between the information conveyed by the speaker in asserting [a certain information] and the assumed expectation state of the hearer.”

- (8) a. A: What did you drink at the dinner party?
 B: We drank WINE.
- b. A: Surely you drank WINE!
 B: No, we drank BRANDY!

For Zimmermann, contrastive focus is typically absent in answers to *wh*-questions (8a)²³, but is present in correcting statements (8b). He further explains:

“The most likely speech act following on a *wh*-question is an answer providing the required information. The speaker can also assume that the hearer will not be surprised by the choice of [*wine*] [...] and therefore will have no problems with updating the common ground accordingly. Hence, no need for contrastive marking. In [(8b)], in contrast, it follows from hearer A’s assertion that she does not expect to be contradicted” (Zimmermann, 2007, p. 155).

2.2.2.2 Accessibility: repetition, frequency, and probability

A different strategy to better understand notions of ‘givenness’ and ‘focus’ has been to assume that, as a discourse proceeds, information varies in its overall

²³ Example adapted from Zimmermann, 2007, pp. 154-155.

level of activation. In this sense, the relationship between prominence cues and information status has also been analysed according to the accessibility of the referents to the listener. Information that is highly activated tends to be highly accessible and is referred to by means of pronouns or shortened expressions, while less accessible material is referred to with full referring expressions (e.g. Grosz et al., 1995; see Arnold, J. E., 2010, for a review).

Accessibility is influenced by syntax, topicality, recency of mention, and other factors. It is possible that accessibility plays a role in accenting as well as in the choice between full noun phrases and pronouns, i.e. accessible information is deaccented, while non-accessible information is accented. This approach can capture given/new distinctions, since given information is likely to be accessible, while new information is not. Similarly, it also captures apparent exceptions, especially when accessibility to non-previously mentioned information can be inferred or derived from the discourse context. In addition, previously mentioned information is not always accessible, and thus, its reintroduction in the discourse would probably require an accent.

Different theories have tried to account for effects of the overall informativeness and predictability of words in speech production: the Probabilistic Reduction Hypothesis (PRH) (Jurafsky et al., 2001), the Smooth Signal Redundancy Hypothesis (SSRH) (Aylett & Turk, 2004), the Uniform Information Density (UID) (Levy & Jaeger, 2007). Basically, all share the fundamental insight that word forms are shortened and reduced when they have a higher probability of occurrence and recognition. For example, the Probabilistic Reduction Hypothesis (PRH) (Jurafsky et al., 2001) simply argues that word forms are reduced when they have a higher probability of occurrence. In this sense, lexical predictability can account for the reduction of word forms, as indicated by segment shortening. For example, using spontaneous speech, Fowler and Housum (1987) analysed the

acoustic correlates of prominence (f_0 , intensity, and duration) of the first two occurrences of words uttered in a broadcast radio monologue. They reported that duration was altered, but not f_0 or intensity, so that 71% of words uttered a second time were shorter. Secondly, in a subsequent perception experiment with extracts excised from the same monologue, they observed that listeners better identified speech items uttered for the first time, i.e. those introducing new information.

A similar study was conducted by Bard et al. (2000). This time the status of 'given' and the loss of intelligibility was tested in four different experiments according to whether speakers were aware or not of what listeners did and did not know. In a series of task-oriented dialogues, speakers and listeners reported to each other visual landmarks on their respective maps, which differed in some cases. Bard and her colleagues reported that repeated mentions of map landmarks by speakers became less clear, regardless whether listeners had heard (experiment 1) or seen the referent (experiment 2). In their turn, speakers reduced intelligibility when repeating a map landmark that had first been mentioned by the listener, including those cases in which speakers themselves could and could not see the referent (experiments 3 and 4).

Not only referent repetition, but also lexical frequency is related to the realisation of prominence. For example, shorter duration has been reported for high-frequency words when compared to low-frequency words. For example, Fosler-Lussier and Morgan (1999) carried out a corpus-based analysis of speaking rate and word frequency and observed that both factors were responsible for a different pronunciation of frequent words from the canonical, most likely pronunciation. In their turn, Pluymaekers et al. (2005) observed that lexical frequency also affected affix duration. Affixes that were attached to high-frequent words were shorter than those attached to low-frequent frequent words and, therefore,

less prominent.

Furthermore, next to repetition and lexical frequency, different forms of transitional probabilities can also affect the acoustic realisation of prominence. Gregory et al. (1999), for example, developed a probabilistic model including several measures such as lexical frequency, type of word collocation, and both discourse repetition and semantic relatedness. In the discussion of their results, they concluded that all these factors can be seen as part of the same process of prominence decrease, so that word duration tends to be reduced by all these factors. In a subsequent study, the same team analysed the contribution that each of them had separately to the observed shortening effect (Bell et al., 2009).

In their turn, Watson et al. (2008) observed an acoustic difference in the realisation of both less predictable words and important words in context. In the first case, words were uttered with longer duration and a greater pitch excursion, whereas in the second case, words were pronounced with greater intensity.

By the same token, Gahl and Garnsey (2004) reported an important difference in prominence realisation according to verb argument structure, i.e. verbs like 'believe' usually occurs more frequently with sentential complements than with direct objects, while verbs like 'confirm' are more often accompanied by direct objects rather than sentential complements. Thus, Gahl and Garnsey observed that verbs occurring with dispreferred arguments tended to be lengthened and preceded a longer pause when uttered.

2.2.3 Syntactic structures and prominence

The mentioned studies linking the realisation of prominence to a series of factors affecting intelligibility and word duration are also closely related to the syntactic structure of utterances. Initially, this was Chomsky and Halle's (1968) view, who put forward the idea that phrasal stress was independently determined

by a recursive algorithm that converted syntactic deep structures into a phonological transcription introducing prominence according to a series of phonological rules. Then, in order to relate a sentence to its context, i.e. signal the focus and presupposition structure, the interface system made use of the available stress prompted by syntax (Jackendoff, 1972). This view was explored further by Halle and Vergnaud (1987), who developed a metrical approach to this rule following the metrical tree notation proposed by Liberman (1979). However, Chomsky and Halle's view was contested by Bolinger, who insisted that prominence is realized through accents that reflect both the speaker's intentions and the information structure of the utterance:

“The distribution of sentence accents is not determined by syntactic structure [...]. Syntax is relevant indirectly in that some structures are more likely to be highlighted than others” (Bolinger, 1972, p. 644).

Along with Bolinger, several other authors discussed the relationship between prominence and syntactic structure. For example, Schmerling (1976) questioned the syntactic account of sentence stress assignment and proposed that prominence realisation was affected by a number of semantic constraints. Schmerling noticed that argument structure played an important role in phrasal stress. She also remarked that in languages such as German and English the direct object of a verbal phrase is always prominent, regardless of whether word order is OV or VO. Argument structure was also analysed by Gussenhoven (1984), who incorporated this semantic generalisation and proposed that a phrase carrying a single pitch accent may include a second argument if it is semantically justified and both are adjacent to each other. This was captured in his *Sentence Accent Assignment Rule* (SAAR) (1983a), which can be summarised as follows:

“If focused, every predicate, argument, and modifier must be accented, with the exception of a predicate that, discounting unfocused constituents, is adjacent to an argument” (Gussenhoven, 1992, p. 84).

Around the same time, the relationship between phonology and syntax in the *Standard Theory* outlined by Chomsky and Halle (1968) was revised by Selkirk (1984), who adopted the pitch accent view of intonation developed by Pierrehumbert (1980) a few years earlier. As a consequence, Selkirk’s work met with considerable success in accounting for patterns of phrasing in a variety of languages. Later on, the syntax-prosody mapping was addressed in some cases from the constraint-oriented framework of *Optimality Theory* proposed by Prince and Smolensky (1993) (e.g. Schwarzschild, 1999; Szendrői, 2001; Truckenbrodt, 1999).

The debate on the relationship between syntactic structure, semantics, and prosody mainly centered on the role of syntax (and semantics) in determining focus and accent. Gussenhoven’s (1983a; 1983c) views, differently from Selkirk’s (1984), were based on the surface position of the constituents and semantic roles they could take on. As a result, several studies investigated NP focus projection to other constituents (e.g. Birch & Clifton Jr., 1995; Bock & Mazzella, 1983; Nootboom & Kruyt, 1987). In this regard, as mentioned earlier (§ 2.2.2.1), it was observed that focused prosodic heads can project to larger units, even to a whole verbal phrase (Birch & Clifton Jr., 1995).

In the relationship between prosody and syntactic structure, it has been pointed out (Szaszák & Beke, 2017) that the syntax-phonology interface is, to some extent, exploitable to disambiguate ambiguous sentences that have interpretations that differ in their surface structure. Several grammatically correct prosodic phrasing and accentuation patterns may exist for a given syntactic representation. For example, the link between syntactic structure and prosodic breaks is formalized in the grammar as prosodic structure, which provides cues to syntactic structure

through knowledge of the grammar (Truckenbrodt, 1999). This is pretty much the opposite, however, when mapping prosody to syntax, since the ambiguity is by no means so easily resolved. Similar to studies reporting on the important role of prosodic breaks in resolving syntactic ambiguities, Schafer and her colleagues (2000) conducted a series of experiments to analyse the disambiguation of syntactic structure by means of prosodic prominence.

- (9) a. I asked the pretty girl who is COLD.
b. I asked the pretty girl WHO is cold.

In their first experiment, Schafer et al. compared the role of a pitch accent when signalling the word ‘cold’ in (9a) and when signalling the interrogative pronoun *who* belonging to an embedded question in (9b). Schafer and her colleagues ran the experiment using a short and long version of the same sentence (i.e. ‘I asked the pretty girl who is cold’ vs. ‘I asked the pretty girl who is very cold’) with different intonations, and they observed that syntactic analysis of such ambiguous sentences are disambiguated in favor of an embedded question by placing a pitch accent on the interrogative phrase.

A more complex view on prosody is also considered, so that together with syntax, other factors also have an influence on prosodic structure (e.g. Watson & Gibson, 2004). Besides, some languages need more intervention of prosody to mark phrasal stress, as English; while in other languages, such as Hungarian, syntax is more independent and flexible to convey information structure, thus allowing prosody to adopt functions not driven by syntax (e.g. Genzel et al., 2015). These cross-linguistic differences (see a more detailed discussion in § 2.2.5) can be grouped into a languages with rigid word order, which allow prosodic marking of focus at different locations in the sentence; and languages with a more flexible word order, which is exploited to mark focused constituents and exhibits less variation in the location of prosodic prominence (e.g. Donati & Nespør, 2003).

Non-canonical syntactic structures

Among the strategies used by intonational languages to mark information structure and focused constituents, there exists a three-way distinction. According to Skopeteas and Fanselow (2010), strategies follow a scale of structural complexity from less complex to more complex: (1) *in situ*, e.g. through prosody, (2a) *ex situ*, e.g. through word order, and (2b) *ex situ*, e.g. through cleft constructions²⁴. In the case of *ex situ* marking of prominence, non-canonical syntactic constructions result in different word order patterns, especially in languages with a more rigid word order such as English. For example, *argument reversal* can be realised as inversions and passive constructions, while cleft constructions include *fronting* (also ‘front-focus’ or ‘preposing’) and locate the highlighted constituent in sentence-initial position. In the subsequent examples, constituents in non-canonical positions are in boldface.

In the case of inversions (10), the prepositional phrase appears to the left of the verb, a position where the subject would canonically appear.

- (10) a. In one of the first-class carriages there is a baby [...]. There is a rusty old divine, in gilt-rimmed spectacles and a jowl, reading the *Guardian*. [...] And lastly, a little young lady, sitting facing the engine, with the dusty blast driving hot and full in her face, blinking, coughing, choking, with the utmost patience. **On her lap** lies a huge bunch of red and yellow roses and heavy-scented double-stocks, all limp and drooping and soiled. (Broughton, 1870, p. 73)
- b. He successively taught me by words, by diagrams, and on the ground [...]. The nights frequently passed in this occupation, **so much** did

²⁴ There also exist focus-related morphemes that are uncommon in intonational languages and will be omitted in this review with English examples.

this study absorb us, **so well** did my teacher know how to make it interesting. (Goodall, 2014, p. 45)

- c. **No evidence for evolution** pleased Darwin more than the presence in nearly all organisms of rudimentary or vestigial structures. (Gould, 1980, p. 29)

Passive constructions (11) promote the object to subject position, and a *by*-phrase introduces the logical subject, which is often omitted, as it conveys information considered irrelevant or given. In the following examples the passive verb is underlined.

- (11) a. Parisian director Paul Verhoeven is suing the Cannes film festival, claiming that his 2009 drama *Teenagers* was not accepted because of sexual prejudice. The lawsuit is the first ever filed against the festival. **The low-budget film**, which details the relationship between two teenage boys, was shown at other festivals around the world but it was rejected when submitted to Cannes. (*The Guardian*, March 2nd, 2015)
- b. Addressing the author as “Dear Orwell”, Eliot, then a director at publishing firm Faber & Faber, writes on 13 July 1944 that the publisher will not be acquiring *Animal Farm* for publication. [...]. **Animal Farm**, a beast fable that satirised Stalinism and depicted Stalin as a traitor, was rejected by at least four publishers, with many, like Eliot, feeling it was too controversial at a time when Britain was allied with the Soviet Union against Germany. (*The Guardian*, May 26th, 2016)

Cleft-sentences (12) split a given sentence into two different clauses. In the following examples it is possible to find cases in which the cleft-clause contains the focus, although this is not always the case. In cleft-sentences, prosody often

interacts with syntax to focus different constituents²⁵. In order to illustrate this interaction, focus in the next examples is indicated in upper case, other prosodically salient words are indicated in small caps, and other pitch accented words are indicated in regular lower case italics.

- (12) a. A: Oh, boss, I guess we made a mistake.
B: Stupid, **it was the CAR** we were *supposed* to STEAL!
- b. **What *these* warnings have *achieved*** is political COVERAGE for OFFICIALS. [03/11/01]
- b'. **What THESE warnings have *achieved*** is political COVERAGE for *officials*.
- c. It's got to be standardized, and **THAT'S what** the *President's proposal* will DO. [26/10/01]
- c'. It's got to be standardized, and **that's what** the President's PROPOSAL will *do*.

In the *it*-cleft sentence (12a), the focused information, 'the car', appears at the beginning of the sentence. The *wh*-cleft in (12b) and (12b') shows a difference in the prosodic realisation of the cleft clause, as showed by the highlighted narrow focus 'these' in (12b'), while in (12b) 'officials' is the focused constituent. In (12c), a pitch accent is located on 'that's' in the cleft clause, and the sentence concludes with a nuclear pitch accent on the verb 'do'. Conversely, in (12c') the narrow focus is realized through prosodic prominence on 'proposal'.

Similarly, *fronting* (13) also places the focal constituent in sentence-initial position, but it does not split the sentence.

²⁵ The examples showing a date in square brackets (e.g. [03/11/01]) have been adapted from Hedberg & Fadden, 2007 and appear as they cite them from the corpus they collected from the broadcast television program *The McLaughlin Group*.

- (13) a. A: Could I have a piece of pie?
B: Sorry, we're all out of pie. I can offer you [_F a MUFFIN.]
B: Sorry, we're all out of pie. **A MUFFIN** I can give you.
- b. A: Shall we start?
B: Nobody knows [_F WHEN] they will arrive.
B: **WHEN** they will arrive nobody knows.

In addition, fronting is superficially similar to clitic left-dislocation. However, in the case of left-dislocation (14), which is not part of the subsequent main clause, a coreferential pronoun appears in the canonical position of the constituent (underlined).

- (14) a. **That girl**, I saw her yesterday at Sally's party?
b. **One of the guys at the party**, he got so drunk that he couldn't even remember his name.

Both fronting and left-dislocation are considered as structures resulting in *topicalisation*, i.e. the promotion of a constituent to the topic of the utterance. For example, in the fronting case (13a), the focused constituent 'a muffin' is topicalised, as in the case of left-dislocation (14a) and (14b): the focus 'her' and the topic 'he' are highlighted. Leonetti and Escandell-Vidal note further that:

“Focalisation differs from clitic dislocation in a number of syntactic properties: there are no resumptive clitics, there is only one contrastive focus slot, and the construction shows all the typical features of operator-variable configurations, such as sensitivity to island contexts and weak cross-over effects” (Leonetti & Escandell-Vidal, 2009).

2.2.4 Text-to-speech synthesis and prominence

Depending on the syntactic differences among languages, top-down expectations corresponding to knowledge of syntactic structures by native speakers have been observed to influence prominence perception (Wagner, 2005). This insight has received support by research from text-to-speech synthesis, where development of algorithms usually include word class (part-of-speech), word length, and position in the sentence. For example, Lea (1980) developed an algorithm arranging word categories according to prominence reduction, but his algorithm tended to place a pitch accent on almost all content words when applied in Text-to-Speech synthesizers. This initial classification was later refined by Altenberg (1987), who established a less strict hierarchy of word class, in which, for example, some word classes, such as nouns, that are often accented may not receive prominence on some occasions. Along the same lines, Quené and Kager (1993) improved the set of function words that do not carry a pitch accent and extended it by including specific content words which were considered as having little semantic information, e.g. *maand* ('month') and *jaar* ('year'). In addition, they also included two rules: deaccent the middle one of three adjacent content words, and deaccent words conveying given information, as in adjectives before a proper name, e.g. *koningin Beatrix* ('queen Beatrix').

Later, Streefkerk (2002) also conducted a series of experiments to analyse the lexical and syntactic correlates of prominence in order to derive a set of rules later evaluated for prediction accuracy. Firstly, she labelled a set of training sentences according to word class, in which she included both lexical probability and contextual probability. Apart from confirming the difference previously remarked between content words and function words: she observed that verbs and adverbs were less prominent than the other analysed content words, while

nouns, adjectives, numerals, and negations showed the highest prominence values. Besides, in terms of word length, longer words tended to be more prominent, regardless of whether they were function words (usually monosyllabic) or content words (usually polysyllabic). Streefkerk also tested the position of words in the sentence and reported that nouns, adjectives, numerals, and negations tended to carry more prominence at the beginning of the sentence than at the middle or at the end. Finally, nouns were generally less prominent when preceded by an adjective than when standing alone, which implied that adjectives followed by a noun were to a great extent the most prominent item of the two.

The role played by aspects such as part-of-speech, word length, and position in the sentence were also complemented in other studies by modelling probabilities of word collocation, as previously mentioned (§ 2.2.2.2) (e.g. Bell et al., 2009; Gregory et al., 1999). In this regard, Gregory and Altun remark:

“Part of speech and the informativeness of a word do not capture all aspects of accentuation, as we see in this example taken from Switchboard, where a function word gets accented (accented words are in uppercase):

‘I, I have STRONG OBJECTIONS to THAT’.

[...] Additionally, whether the immediately surrounding words bear [a] pitch accent also affect the likelihood of accentuation. In other words, a word that might typically be accented may be unaccented because the surrounding words also bear pitch accent” (Gregory & Altun, 2004, p. 1).

Thus, for example, in order to account for word collocations, unigram and bigram probabilities were used by Nenkova et al. (2007) among a variety of other features in the prediction of prominence, which included: information status of a

word, i.e. new vs. old; part-of-speech; word length; position of word in the utterance; utterance length; animacy of nouns and pronouns, i.e. concrete, human, non-human, etc.; accent ratio, i.e. the likelihood of a word being accented; dialog act, i.e. statement, opinion, etc.; and some others. Similarly, different models for the automatic detection of phrasal stress have included a feature of different n -gram word probabilities (e.g. [Gregory & Altun, 2004](#); [Kakouros & Räsänen, 2015](#); [Ananthakrishnan & Narayana, 2008](#)) (see Table 5 for a summary).

Acoustic	Lexical	Syntactic	Semantic/Pragmatic
Fundamental frequency (f_0)	n -gram probability	Part-of-speech	Dialogue act
Intensity	Reverse n -gram probability	Position in the sentence	Semantic relatedness
Duration	Joint probability	Animacy	Information status
Spectral tilt	Accent ratio		Kontrast

Table 5: *Examples of acoustic, lexical, syntactic, and semantic/pragmatic features included in text-to-speech synthesis and automatic detection of prominence (e.g. [Gregory & Altun, 2004](#); [Ananthakrishnan & Narayana, 2008](#); [Nenkova et al., 2007](#)). Adapted from [Kakouros, 2017](#), p. 46.*

2.2.5 Cross-linguistic differences

The intonational system of each particular language determines how the combination of focus and prominence is achieved. It is the case, for example, that West-Germanic languages such as English, German, or Dutch usually mark contextually important information by means of prosody without resorting to syntactic operations (e.g. [Gussenhoven, 2005](#)). However, in certain Asian and African languages, pitch accents are hardly relevant. For example, in the case of Korean (e.g. [Jun, 2005](#)) or Japanese (e.g. [Venditti et al., 2008](#)) prosodic phrasing can completely replace the pitch accents that are characteristic of intonation languages, with one extreme case of prosodic marking, ellipsis, in which only the focused part of a sentence is pronounced, while the given part is just omitted.

Conversely, in most Romance languages such as Catalan, Spanish, or Italian, the same result is achieved mainly through altering word order, while intonation is less important (e.g. [Estebas-Vilaplana & Prieto, 2010](#)). More precisely, changes in word order in Romance languages allow particular items to be focused and to phonologically receive an accent. In this case, non-canonical word order patterns may also be accompanied by prosodic prominence.

It is widely accepted that most languages make use of some of the strategies observed cross-linguistically, and typically do not use them concurrently. Ladd ([2008](#), Chapter 6) enumerates three main cases in which cross-linguistic differences in prominence marking are most evident: both *yes-no* questions (YNQs) and *wh*-questions (WHQs); deaccenting and ‘semantic weight’; and predicates (verbs and predicate nouns or adjectives) vs. arguments (noun phrases syntactically linked to a predicate). In his detailed review, Ladd exposes differences that lie at the core of the debate around whether there exist cross-linguistic universals in highlighting information.

Nevertheless, in the case of intonational languages, as mentioned above, there are obvious differences in the use of prosodic and syntactic strategies to mark prominence. Swerts et al. ([2002](#)), for example, conducted a comparative study of accentuation strategies in adjective-noun phrases with Dutch and Italian speakers. The researchers elicited dialogues in which the information status within the noun phrase appeared in four different ways: all new information, single contrast in the adjective, single contrast in the noun, and double contrast in both adjective and noun.

The results were compared in three different analyses. Firstly, they observed that both new and contrastive information were always accented in Dutch, but given information was not. Conversely, in Italian, pitch accent marking was not observed to be a significant factor to distinguish information status, since in the

elicited noun phrases both adjective and noun were always accented, irrespective of the status of the discourse context. Secondly, Swerts et al. also observed that information status in Dutch was reflected in the prosodic prominence, and single contrastive accents were perceived as the most emphatic, while given words were perceived as the least emphatic. In Italian this was not so evident. Finally, a functional analysis of accent patterns was conducted to explore whether listeners were able to rely on acoustic prominence cues to reconstruct the context of a given utterance. The researchers reported that, consequently, Italian listeners did not manage to interpret acoustic cues in terms of the dialogue history.

Later on, Krahmer and Swerts (2004) conducted a follow-up study in which gestural cues were added to the acoustic cues of prominence that they had previously analysed (see § 2.3.8.2 for a description). In another study with Dutch and Spanish speakers, van Maastricht and her colleagues (van Maastricht et al., 2016) similarly observed that Spanish learners of Dutch neglected the role of pitch accents in the target language and always used the same intonation pattern throughout the experimental task irrespective of the focus location of the utterance.

Non-canonical word order in Spanish

The differences described by Swerts et al. (2002) can be explained in Vallduví's (1992) terms of *plastic* and *nonplastic* languages, i.e. whether or not the prosodic patterns of a given language can easily adapt and reflect information structure.

Romance languages are considered to be nonplastic languages. Spanish, for example, is considered to mark focal constituents both syntactically and prosodically, often combining both strategies. Contrastive focus in prenuclear position (15a) is achieved through a high pitch accent, in which a low accent occurs in pre-tonic syllable, the f_0 peak falls within the stressed syllable rather than in the

post-tonic syllable, and it is also possibly accompanied by longer duration and intensity (e.g. Face, 2001; de la Mota, C., 1997). This is followed by a low boundary tone, which results in a L+H* L% intonational pattern (Estebas-Vilaplana & Prieto, 2010). However, if contrastive focus occurs in nuclear position (15b), the pitch accent can also be high or low.

- (15) a. A: Me dijeron que Tania quiere comprar un violonchelo.
'They told me Tania wants to buy a cello.'
B: Yo entendí que quería [_F ALQUILAR] un violonchelo.
'I understood she wanted to [_F RENT] a cello.'
- b. A: Oí que Ana vino el martes.
'I heard Ana came on Tuesday.'
B: No, (Ana vino) [_F el LUNES.]
'No, (Ana came) [_F on MONDAY.]'

Contrastive focus can also be marked by means of cleft constructions. Nuclear stress on the focus can highlight either the subject (16a) or object (16b) in this non-canonical word order. Additionally, it can also appear in fronting constructions with a similar involvement of prosodic prominence (17).

- (16) a. A: Manuel se comió el chocolate, ¿no?
'Manuel ate the chocolate, didn't he?'
B: No, [_F fue ROSANA] quien se comió el chocolate.
'No, [_F it was ROSANA] who ate the chocolate.'
- b. A: Manuel se comió el chocolate, ¿no?
'Manuel ate the chocolate, didn't he?'
B: No, [_F fue la TARTA] lo que se comió (Manuel).
'No, [_F it was the CAKE] what (Manuel) ate.'

(17) A: ¿Han traído traído ya el paquete?

‘Have they brought the parcel yet?’

B: [F La CARTA] han traído.

‘[F The LETTER] they brought.’

It has traditionally been claimed that only contrastive focus can be fronted (e.g. Zubizarreta, 1998), but Vanrell and Fernández Soriano (2013) observed in a production experiment that also narrow focus is fronted in Spanish (18)²⁶.

(18) A: ¿Qué le dio el marinero al viejo?

‘What did the sailor give to the old man?’

B: [F La CARTA] le dio el marinero al viejo.

‘[F The LETTER] gave the sailor to the old man.’

However, Leonetti and Escandell-Vidal comment that:

“In the case of fronted lexical definites, there is also a strong requirement that the propositional content has been made accessible in the immediate context. Thus, a sentence like [(19)], uttered out of the blue, with no connection to any previous relevant information, is quite difficult to contextualise” (Leonetti & Escandell-Vidal, 2009, p. 171).

(19) A: ¿Y qué hiciste ayer?

‘And what did you do yesterday?’

B: # [F Un LIBRO] leí.

‘[F A BOOK] I read.’

²⁶ Example adapted from Vanrell & Fernández Soriano, 2013, p. 261.

Along the same lines, it has been claimed that the prosodic realisation of broad focus would be very similar to that of contrastive focus, even if the fronted phrase does not express any contrast.

- (20) A: ¿Y qué te han traído a ti?
'And what did they bring to you?'
[_F Unas ZAPATILLAS] me han traído.
[_F A pair of SLIPPERS] they brought me.'

Leonetti and Escandell-Vidal (2009, p. 163) report in (20) an example of fronted broad focus pointed out to them by an anonymous reviewer. However, in their interpretation, the researchers argue that “this [(20)] is still a case of contrastive focalisation and represents a marked way to convey the additional idea that the new piece of information is surprising or unexpected” (2009, p. 163)²⁷.

In their paper, Leonetti and Escandell-Vidal actually discussed a type of fronting in (21) that had hardly been discussed in the previous literature.

- (21) a. **ALGO** debe saber.
'S/he must know SOMETHING.'
b. **POCO** te puedo decir.
'LITTLE can I say to you.'
c. **Lo MISMO** digo (yo).
'I say the SAME.'

The previous examples, in which quantifiers are fronted (as shown in bold-face), are interpreted by them as cases of *verum focus fronting*, i.e. focus on the truth value of the sentence (also ‘sentential polarity focus’ or simply ‘polarity

²⁷ This view is very close to that expressed by Zimmermann (2007) (§ 2.2.2.1).

focus’). Thus, they interpret the sentences just seen in (21) as²⁸:

- (22) a. **ALGO** debe saber. → *Sí/seguro que sabe algo.*
‘S/he must know SOMETHING.’ → ‘Yes/sure that s/he knows something.’
- b. **POCO** te puedo decir. → *Sí/Es cierto que yo te puedo decir poco.*
‘LITTLE can I say to you.’ → ‘Yes/It’s true that I can say little to you.’
- c. **Lo MISMO** digo (yo). → *Sí/Es cierto que yo digo lo mismo.*
‘I say theSAME.’ → ‘Yes/It’s true that I say the same.’

2.2.6 Summary

Semantics, pragmatics and information structure

When syllables and words stand out from their environment, they do so due to semantic and pragmatic factors and/or syntactic and lexical elements in the utterance. Prosodic prominence signals the most salient element in the utterance, whose main role is generally accepted to be marking the status of a constituent within the discourse.

Different approaches to information structure take into account the potential of the semantics of sentences to change the previous context of the discourse. Three distinctions contribute to alter the common ground shared between speaker and hearer: given/new, topic/comment, and background/focus.

Information structure and the status of elements in the discourse interact in a process in which prosody plays a crucial role in many languages. Pitch accents can mark semantic contrasts, and words carrying a pitch accent are traditionally seen as referring to new information, while deaccented words were perceived as information already given in the discourse (e.g. [Birch & Clifton Jr., 1995](#); [Bolinger,](#)

²⁸ Verum focus fronting is also considered to be one of the strategies used in Spanish to express irony ([Escandell-Vidal & Leonetti, 2014](#)).

1958, 1961; Cutler, 1976; Dahan et al., 2002). However, previous mention of a word in the discourse is not sufficient for already conveyed information to be deaccented (Brown, 1983; Terken & Hirschberg, 1994).

The distinction of background/focus is closely related to that of given/new. ‘Focus-To-Accent’ (FTA) approaches, a term coined by Gussenhoven (1983a), attempted to narrow the gap between the semantic/pragmatic notion of ‘focus’ and the phonetic/phonological notion of ‘accent’. The way that focus—signalled by pitch accents—interacts with syntax and phonology is at the core of this relationship. It was observed that nuclear accents may work as prosodic heads that ‘project’ to larger units, even to the whole verbal phrase they occur in, in what is known as ‘focus projection’ (Birch & Clifton Jr., 1995).

Additionally, a pitch accent, traditionally considered to fall on the single word that provides contrasting information to a set of alternatives (i.e. contrastive focus) has been claimed to actually express a contrast between the information conveyed by the speaker and the informational expectation of the hearer (Zimmermann, 2007; Leonetti & Escandell-Vidal, 2009).

The concepts of givenness and focus have been also addressed according to the accessibility by the listener to the referents uttered in the discourse: information that is highly activated is also highly accessible and is referred to through pronouns or shortened expressions (e.g. Arnold, J. E., 2010; Grosz et al., 1995). This approach can capture given/new distinctions—given information is likely to be accessible, while new information is not—as well as apparent exceptions. The Probabilistic Reduction Hypothesis (PRH) (Jurafsky et al., 2001), the Smooth Signal Redundancy Hypothesis (SSRH) (Aylett & Turk, 2004), and the Uniform Information Density (UID) (Levy & Jaeger, 2007) are the best-known frameworks aiming at explaining the effects of overall informativeness and predictability of words in speech production.

Lexical predictability can also account for the reduction of word forms, i.e. temporal shortening of segment (e.g. Bard et al., 2000; Fosler-Lussier & Morgan, 1999; Fowler & Housum, 1987; Pluymaekers et al., 2005). However, next to predictability and lexical frequency, different forms of transitional probabilities can also affect the acoustic realisation of prominence (e.g. Bell et al., 2009; Gregory et al., 1999; Watson et al., 2008).

Syntactic structures and prominence

In order to signal the focus and the presupposition structure of an utterance, prosody makes use of the stress made available by syntactic means. This was the idea put forward by Jackendoff (1972) based on Chomsky and Halle's work (1968), although it was soon contested by Bolinger (1972) and Schmerling (1976). In the mid-80's Selkirk (1984) reviewed the *Standard Theory* based on Pierrehumbert's (1980) notion of pitch accent. Later, the syntax-prosody mapping was also addressed from the constraint-oriented framework of *Optimality Theory* (Prince & Smolensky, 1993).

The syntax-phonology interface can be exploited to disambiguate ambiguous sentences whose interpretations differ in their surface structure (e.g. Truckenbrodt, 1999; Schafer et al., 2000). However, the interaction between prosody and syntax is also seen as a more complex one, so that other factors have also an influence on prosodic structure apart from syntax (Watson & Gibson, 2004).

Additionally, non-canonical syntactic structures are used by intonational languages to mark information structure and focused constituents. If *in situ* marking of prominence involves prosody, *ex situ* involves either a different word order or cleft constructions. A different word order can be realised for example as inversions or passive sentences, while cleft constructions—including fronting—locate the highlighted constituent in sentence-initial position.

Syntax was also initially exploited in text-to-speech to develop algorithms to automatically detect prominence (Lea, 1980; Altenberg, 1987; Quené & Kager, 1993; Streefkerk, 2002). These algorithms usually included word class (part-of-speech), word length, and position in the sentence and transitional probabilities (e.g. Ananthakrishnan & Narayana, 2008; Bell et al., 2009; Gregory et al., 1999; Gregory & Altun, 2004; Kakouros & Räsänen, 2015; Nenkova et al., 2007).

Cross-linguistic differences

The intonational system of each particular language determines how the combination of focus and prominence is achieved. Thus, in some languages prosody plays a more important role, while in other languages word order and non-canonical syntactic constructions are preferred (e.g. Gussenhoven, 2005; Ladd, 2008; Venditti et al., 2008). In the case of intonational languages, there are obvious differences in the use of prosodic and syntactic strategies to mark prominence, as the differences observed between Dutch and Italian (Swerts et al., 2002).

In Spanish, a language whose prosody cannot easily adapt to reflect information structure, focal constituents can be marked both syntactically and prosodically. Both strategies are often combined, as in the case of cleft constructions, including fronting. Thus, contrastive and broad focus can be realised in non-canonical syntactic constructions through a pitch accent (e.g. Leonetti & Escandell-Vidal, 2009; Vanrell & Fernández Soriano, 2013).

2.3 Gestural correlates of prominence

In this section, an overview of gesticulation, located at one of the extremes of the so-called Kendon's continuum, is offered. This is followed by a brief review of some of the most conspicuous studies on gestures prior to the 20th century and

by a detailed account of those carried out in the 20th century. Different criteria to categorise gestures are later discussed, especially the one developed by McNeill (1992). Subsequently, head nods and eyebrow raises are also dealt with in more detail, which leads to a discussion on the temporal coordination of gesture and speech. Finally, a large subsection is devoted to the phenomenon of *audiovisual prosody*, i.e. the interaction of gestures and verbal prosody, with especial attention to studies on the interaction between gesture and speech in signalling prominence.

2.3.1 Introduction

Apart from speech, humans also perform gestures in their communication, which are usually defined as spontaneous, and often unwitting, body movements accompanying speech. These gestures, performed with hands, fingers, arms, head, face, eyes, eyebrows, or trunk, are also known in the literature as *gesticulation* (Kendon, 1982) and differ in fundamental ways from the gestures performed in pantomime, from those gestures known as emblems, or from the gestures of sign languages. In this case, and in keeping with most of the literature, the word *gesture* is used to refer to *gesticulation*.

2.3.2 The Kendon's continuum

Following McNeill (1992; 2000), different gestures can be classified according to the relationship that body movements have to speech, and they can be organised depending on their increasing degree of conventionalisation along the so-called Kendon's continuum (as mentioned in the introduction § 1, Figure 1). In their relation to speech, body movements at the left end of the continuum are produced unawares together with speech. Conversely, at the right end, the

movements performed in sign languages are produced consciously to communicate, which typically occurs in the absence of speech (Figure 4).

<u>Gesticulation</u>	→	<u>Emblems</u>	→	<u>Pantomime</u>	→	<u>Sign Language</u>
Obligatory		Optional		Obligatory		Obligatory
presence of		presence of		absence of		absence of
speech		speech		speech		speech

Figure 4: *Gestures in their relationship to speech in the Kendon’s continuum. This one and the following continua are adapted from McNeill, 2000, pp. 2-5 and Loehr, 2004.*

However, McNeill considers several other dimensions of body movements in this continuum. In their relation to linguistic properties, he points out that at the left end, movements have neither a lexicon of agreed-upon symbols nor a phonological, morphological, and syntactic system to combine its constituents. In this sense, the gestures at the left end are not conventionalised, i.e. there is not a conventional code for their correct realisation, while sign languages at the right end share a conventionally structured code that is necessary for communication (Figure 5).

<u>Gesticulation</u>	→	<u>Pantomime</u>	→	<u>Emblems</u>	→	<u>Sign Language</u>
Absence		Absence		Presence		Presence
of linguistic		of linguistic		of some linguistic		of linguistic
properties		properties		properties		properties
(Not conventionalised)		(Not conventionalised)		(Partly conventionalised)		(Fully conventionalised)

Figure 5: *Linguistic properties and degree of conventionalisation of gestures in the Kendon’s continuum.*

Finally, according to the semiotic characteristics of body movements along the continuum, it is observed that when a meaning is conveyed, it can be done in either a segmented or in a global way. In the former case, the meaning of the whole is determined by the combination of its smaller parts, as morphemes do when they are combined into larger meanings in spoken language (bottom-up).

In the latter case, conversely, the meaning of the parts is determined by the whole (top-down). McNeill illustrates this last case by means of an example in which an individual utters the sentence 'he grabs a big oak and he bends it way back'. Upon the words 'bends it way back', the speaker's hand appears to grip something in front of him and to pull it back and down towards his shoulder (Figure 6). The speaker's hand, its movement and the direction can be understood as the parts making up the gesture, although they cannot be combined independently as morphemes in speech. The movements back and down performed by the speaker's hand acquire their meaning by the whole gesture's enactment of bending back a tree. It is in this sense that McNeill refers to global semiosis as a case in which the meaning of the parts depend on the meaning of the whole.

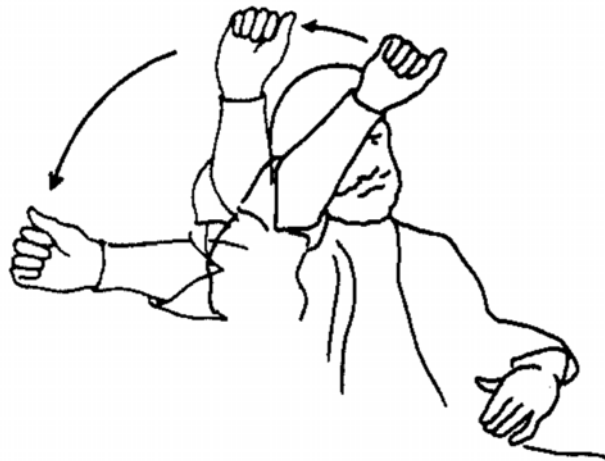


Figure 6: *Depiction of the gesture performed by a speaker while uttering the sentence 'he grabs an oak and he bends it way back'. Adapted from McNeill, 1992, p. 12.*

Furthermore, a gesture is considered synthetic when it combines different meanings into one symbol stretching across several words. In the previous example offered by McNeill, the single gesture performed by the speaker captured the aspects of actor, action, and manner corresponding to the uttered words 'he',

'bends', and 'back'. By the same token, a sentence is analytic in that each successive word is a symbol.

At the far right of the spectrum, for example, movements performed in sign language are conventionalised, i.e. they are agreed-upon linguistic symbols. Besides, signs are analytic because their meaning is encoded in a series of symbols, not in a single one. In sign languages, elements such as hand form and direction typically contribute to the meaning of the sign, as morphemes do in speech, and so signs in sign languages are segmented (bottom-up meaning).

In the case of pantomime, however, there is not a series of conventionalised symbols. The movements performed by a mime occur in the absence of speech, yet their meaning is understandable. Pantomime is also analytic because each movement has meaning of its own. However, and differently from sign languages, each of the parts that make up the movement do not contribute separately to the global meaning; instead, they are understood according to the movement as a whole (top-down meaning).

Emblems, in their turn, are conventionalised symbols and, although they can be understood in the absence of speech, they often accompany speech and present some linguistic properties. For example, an emblem that is widespread across many cultures is the 'OK' movement corresponding to a clenched fist with the thumb extended upwards²⁹. This movement modifies the noun or situation being communicated, and it is synthetic because it encodes the meaning of its own. Nevertheless, emblems are considered by McNeill as a case of segmented semiosis, presumably because larger meanings can result from the combination

²⁹ As symbols with an agreed-upon meaning, emblems usually differ according to each culture. For example, the 'OK' sign is usually rendered in North America by means of the thumb and the index in contact forming a circle, which in turn means 'money' in Japan, 'sex' in Mexico, and 'zero' in France (Aronson et al., 2010, p. 90).

of several of them.

Finally, at the left end of the spectrum, gestures are holistic because they depend on speech for their meaning to be adequately understood. They also stretch throughout space and time and lack linguistic properties (Figure 7).

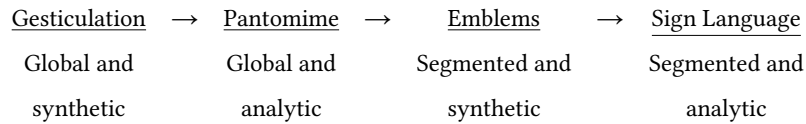


Figure 7: *Semiotic characteristics of gestures in the Kendon's continuum.*

2.3.3 The study of gestures until the 20th century³⁰

2.3.3.1 Antiquity

In the Antiquity, gestures were recognised as a fundamental feature of human expression. For example, in the Indo valley manual gestures accompanied the ritualistic pronunciation of the Sanskrit Vedas, so that the height of tonal accents were matched by the height of the performed gestures (Shukla, 1996). In the Greek and Roman tradition, gestures were performed with a more purposeful goal, and public speakers considered them as part of the theatrical techniques applied in the oratory to sway the feelings of the crowds. Among other commentators on the role of gestures, such as Cicero, it was the Roman rhetorician Quintilian (ca. 35-100), in the first century AD, who discussed in depth the role of gesture in his *Instutio oratoria* ('Education of the orator'). By the use of the Latin word 'gestus', Quintilian not only referred to the movements performed with arms and hands, but also to those aspectos of the nonverbal communica-

³⁰ In this review of studies on gesture before the 20th century, I basically follow Kendon, 2004, pp. 17-61.

tion that can have an influence on the audience, such as the carriage of the body and its posture, the movements of the head and the face, and the glance.

2.3.3.2 The Renaissance and the Enlightenment

Beyond the rhetoric efforts of public speaking, gestures became more relevant in relation to how to conduct and express oneself during the Renaissance. Social position given by birth was not anymore the only way for courtiers to acquire social status, but it was something to achieve now by meeting the new civil standards set by works such as *Il libro del cortegiano* ('The book of the courtier', 1528) or *Il Galateo, overo de' costumi* ('Galateo: the rules of polite behaviour', 1558), which circulated among the European courts. Around the same time, the first European contacts with America arose the interest in gestures as a universal means of communication in the absence of a common language. However, it was *L'arte de' Cenni* ('The Art of Signs', 1616) by Giovanni Bonifaccio, which was the first work to address the topic of gesture systematically.

Previously, and also later, gestures were considered to be a 'natural', a more 'primitive' language, and as a consequence they were seen to form a universal way of communication in the face of the numerous languages that the fall of Babel gave rise to. These ideas prompted a series of philosophical reflections on the nature and origin of language. Even if Giambattista Vico is sometimes considered as one of the first thinkers to have written on a possible gestural origin of language (1725/1744, see Danesi, 1993, for a review), such an idea became more well-known through the work *Essai sur l'origine des connaissances humaines* ('Essay on the origin of human knowledge', 1746), by the French philosopher Étienne Bonnot de Condillac. One of Condillac's main arguments focused on how accidental signs may have become conventionalised. Thus, after a first stage in which the natural gestures done by another individual may be interpreted as having

communicative content, gestures are thought to have been produced voluntarily and transformed into instituted signs. This first form of linguistic expression would have been previous to any vocal sounds, which may have arisen at a later stage (see [Burling, 2000](#), for a modern account of sign conventionalisation and its possible role in the origin of language).

Around the same time, in his *Lettre sur les sourds et les muets* ('Letter on the Deaf and Dumb', 1751) the famous philosopher Diderot also showed interest in the relation that gestures had to thought and speech. For him, thought was a reflection of the linear nature of spoken language, and so he wondered about the structure of thought that gestures might impose on people deprived of the ability to hear. The constraints imposed on thought by speech seemed to him more artificial when compared to those of gesture, which he considered a medium better suited for the expression of thought.

Interest in the gestures that deaf people make to communicate were started to be seen as *signs* in their own right, i.e. as elements conveying meaning. Consequently, the deaf could be regarded as having normal intelligence and even able to be educated, once they learnt a proper system to communicate. This was the task that the Abbé Charles-Michel de l'Épée committed himself to as the educator of deaf children. Helped by a group of deaf children who had already created their own signs, de l'Épée developed a method to teach a system of manual signs to French deaf children so that they could later develop their own system of signs (1776). Nevertheless, the idea of sign languages like those developed by de l'Épée and his successor, the Abbé Roch-Ambroise Cucurron de Sicard, was not new, since John Bulwer had earlier developed a finger-spelling alphabet and dedicated a series of works to communication and the body (1644).

However, the idea that gestures, as those in sign languages, could be the basis for a universal language was criticised later by Joseph Marie de Gérando in his

work *Des signes et de l'art de penser, considérés dans leurs rapports mutuels* ('On signs and on the art of thinking considered in their relations to each other'), which was published in 1800.

2.3.3.3 The 19th century

In the 19th century, the fundamental interests that had previously been manifested in relation to gestures continued: the possibility of considering a gestural stage previous to speech in the origin of language; the relationship between the medium of communication and thought; and whether gestures can be seen as a more fundamental form of expression. However, in the 19th century different disciplines opened up new perspectives, and archaeology, geology, and biology made important contributions to the study of gestures.

For example, the archaeological discoveries that had been made at Herculaneum and Pompeii unburied artifacts and mosaics depicting old Romans. This had an important impact on how modern common people living in the same area saw themselves, which was made evident in the work *La mimica degli antichi investigata nel gestire napoletano* ('Gestural expression of the ancients in the light of Neapolitan gesturing'), published in 1832 by Andrea de Jorio. The book offered an explanation of the gestural expressions of the modern inhabitants of Naples, which, for de Jorio, were a continuation of the cultural and gestural practices that had been maintained since old Roman times. The abundant descriptions of the gestures used in de Jorio's time in Naples are still an important sample of ethnographic research.

The relationship between the observed cultural differences and the basic unity of the human mind was discussed by Edward Tylor in his *Researches into the Early History of Mankind* (1865). Tylor, who was later regarded as one of the founders of what later came to be known as cultural anthropology, began his book with a

reflection on language. Tylor's underlying argument was that the expression of thoughts and ideas by human beings through speech can also be achieved mainly through gestures, pictures, and writing. Tylor proposed that gesture-language and picture-writing share an ability to represent meaning. In the development of this argument, he reviewed different examples in which the gestures of sign languages were used to communicate, as seen in the communication of the deaf, in the sign language used by Native American Indians in specific situations, and in the gestural behaviour of everyday life. By highlighting the striking similarities among such groups, Tylor discussed in detail the conveyance of meaning.

Similarly, Garrick Mallery, a US colonel involved in the military campaigns against the Native American Indians in the late 19th century he was assigned to the US Bureau of Ethnology. In such a position he dealt with the sign language that Native Americans used on certain occasions. The progress report on his studies on this matter was published in 1881 as *Sign Language Among North American Indians Compared with that Among other Peoples and Deaf Mutes*. In this work, Mallery discussed the nature of the primordial language, as argued in the biblical tradition, and defended that abstract thought and reason were possible for humans even before they were capable of speech. For him, concepts can be formed and expressed by other means different to sound, as it is the case with gestures. Mallery concluded that, although both voice and gesture might have arisen in parallel in the origin of language, gesture could have been more important at an early stage.

Also in the 19th century, Charles Darwin's book *The Descent of Man* (1871, p. 56) included a comment on the origin of language, in which gestures were referred to as the aid of "man's own distinctive cries", some of which are made in "imitation of natural sounds". Darwin understood human gesture in a purely biologist manner as part of the nonverbal communication that is genetically de-

terminated in our species. In his *The Expression of the Emotions in Man and Animals*, published one year later (1872), he developed the idea that body movements, and especially facial expressions, were, therefore, universal and tightly linked to the expression of emotion. Darwin proposed an evolutionary account of these facial expressions and argued that they were useful to the individual for the performance of some basic actions. He went on to explain that raising the eyebrows would increase the range of vision, but such an action would be also of great adaptive advantage to human beings, for example, for communicating. For Darwin, facial expressions were linked to our emotions and cannot be language-specific, but universal.

Finally, at the end of the century, language started to be studied as a psychological phenomenon. Although Wilhelm Wundt was not the first researcher to carry out psychological experiments, he was the first to open a laboratory officially designated for experimental psychology. His initial studies on sensation and attention gave way to the study of social phenomena, and a full range of topics were covered in his ambitious treatise *Völkerpsychologie. Eine Untersuchung der Entwicklungsgesetze von Sprache, Mythos und Sitte* ('Social psychology. An investigation of the laws of evolution of language, myth, and custom'), published between 1904-1920.

In his *Völkerpsychologie*, Wundt devoted the first volume to human language, and whose second chapter dealt with gestures. He presented a large review on the topic and included examples of the sign language of the deaf that he had personally encountered in several places; the sign language of Native America as Mallery had described it; the gestural tradition of Naples—based on de Jorio's work—; and the sign language used by Cistercian monks. Wundt offered a classification of the different types of gesture and argued that, contrary to words, the origin of gestures can be easily identified, once the principles of expressive move-

ment is understood. For him, these basic expressive movements correspond to the affective states that the individual experiences, which induce similar feelings when observed by others. Communication is therefore the way these feelings are conveyed. Beyond this, humans also share their conceptualised experiences, which are shared by means of body movements in the form of voluntary gestures. The conceptual reference so expressed is, for Wundt, what gestures add to the emotional content of expressive movements. Thus, Wundt argued that gesture is a primitive form language prior to speech because it combines both concept and emotional content so that it relates naturally to its conveyed meaning.

2.3.4 Gesture studies in the 20th century

2.3.4.1 The importance of spontaneous gestures

Around the turn of the century, interest in gesture and language decreased. Until Critchley's work *The Language of Gesture*, which was published in the late 30's (1939), there were few relevant publications on the topic. Nevertheless, little after, Efron's *Gesture, Race, and Culture* (1941/1972) was published, a work that is considered the first modern study devoted to spontaneous gestures.

At the height of the racist theories purported in Nazi Germany, Efron studied the hand movements performed by the Jewish and Italian immigrants that had recently settled down in New York. One of his main conclusions was that gesticulation differed among cultures, although the gesticulation of the immigrants' children and grandchildren did not show any differences, as they had already been assimilated by the Anglo-American culture.

One of Efron's main contributions was the development of a careful methodology intended to account for his object of study. Efron classified gestures into two main categories: gestures whose meaning was independent of speech, which

he named *objective* gestures; and gestures tightly linked to the meaning conveyed in speech, which he called *logical-discursive* gestures. Previously, gestures had been classified in relation to some particular aspect: for example, Wundt, similar to Efron, had also based his classification on semiotic principles and on the relation between gestures and meaning.

Efron never presented his classification in a systematic way, although his work was the basis for later classifications, especially that of Ekman and Friesen (§ 2.3.4.3). Nevertheless, a comprehensive summary of Efron's typology can be found in Ekman's introduction to the reissue of Efron's work in the 70's (1941/1972). In his classification, Efron mainly focused on the different perspectives from which gestures can be considered as well as on the different ways they can be used.

Efron studied gesture, as performed with arms and hands, from three main perspectives: a *spatio-temporal* perspective, dealing with the kinesic features of movements; an *inter-locutional* perspective, focused on the interactional functions of gestures; and finally, a *linguistic* perspective, based on how gestures convey meaning. He divided his linguistic perspective into gestures that have an 'objective' meaning and those that have a 'logical' or 'discursive' meaning. In addition, his category of objective gestures included *deictic* gestures, those that point to an object; *physiographic* gestures, those that depict the form or the spatial relationship of an object (*iconographic* gesture) or of a bodily action (*kinetographic* gesture). Finally, next to the categories of logical-discursive and objective, Efron also focused on what he called *symbolic* or *emblematic* gestures, which correspond to the conventionalised emblems mentioned above, which possess agreed-upon meanings (§ 2.3.1).

The second category proposed by Efron, logical-discursive gestures, become relevant together with the speech they co-occur with. For him, they actually

reflect the cognitive process involved in speaking. Thus, rather than referring to an object, the type of logical-discursive gestures, dubbed by him as *batons* in a comparison to a conductor's baton, reveal the linguistic mental activity of the speaker. Finally, within this second category, Efron further considered what he called *ideographic* gestures, i.e. gestures that depict in the space a sketch of the abstract movements carried out by the speaker in their thinking process.

2.3.4.2 Kinesics and linguistics

Around the same time as Efron published his work, an interest beyond gesture arose and extended into nonverbal communication (also 'nonverbal behaviour'), which became a topic of research at the crossroads of linguistics, psychology, and psychiatry. The use of simultaneous audio- and video-recordings soon made obvious that far more than words were involved in communication: tones of voice and vocal quality, patterns of intonation, style of talking, modes of hesitation, body posture and body movements, facial expressions, etc.

A multidisciplinary project called the *Natural History of an Interview* (Bateson, 1958; McQuown, 1971) brought together the linguists McQuown and Hockett (and, at a later stage, Trager and Smith) with a team of psychotherapists and anthropologists (Birdwhistell among them) to analyse in detail the body behaviour and the speech used in social interactions. Using an interview previously recorded by one of the researchers, a detailed description of everything that could be observed in an interaction with a psychiatric patient was analysed. Among the numerous observations made by the group, McQuown not only noted that behaviour was reflected in prosody (e.g. a narrowed pitch register could be suggestive of depression, apathy, or boredom), but also that speech and body movements were tightly linked.

The interest in the relationship between language and motion had firstly led

Birdwhistell to conduct research into *kinesics* (1952; 1970), and he proposed a classification of movements that included *kines*, *allokines*, *kinemes*, and *kinemorphemes*, in an analogy to the linguistic categories of phones, allophones, phonemes, and morphemes. His contention was that “linguists [had to] turn to body motion for data to make sense out of a number of areas now hidden in the parts of speech” (1970, p. 127). This undoubtedly called for a shift in the traditional framework that was applied in linguistics to account for unresolved theoretical issues. In his attempt to extend his analogy between kinesics and linguistics, Birdwhistell also pointed out the close relationship between body movements and intonation, although he did not pursue that line of research.

In his collaboration with Birdwhistell, the psychiatrist Albert Scheflen, defined a series of body movements according to the role they played in social interactions (Scheflen, 1964). For example, Scheflen observed that head shifts used to accompany different arguments made by the speaker and torso shifts were associated with larger units of conversation. Similarly, Scheflen and Birdwhistell also described how eyeblinks, head nods, and hand movements coincided with the end of clauses. One of their main insights was that the movements made by speakers are highly patterned and show structural features that are analogous to those of speech.

2.3.4.3 Redefinition of gesture categories

At a moment when the concept of ‘nonverbal behaviour’ had become known, and psychiatrists had started to get interested in the topic, Freedman and Hoffman (1967) developed a classification of hand movements as observed in psychiatric interviews. They divided hand movements into two broad categories: *body-focused* movements, in which the hands touch the body and are not speech-related (e.g. grooming); and those related to speech, dubbed by them as *object-*

focused movements. The latter category included five gesture types along a continuum of increasing information content and decreasing integration with speech: *punctuating movements*, which accentuate the accompanying speech, but do not add any information to it; *minor qualifying movements*, which are simply a “turning of the hand from the wrists” and add some information to the speech content; *literal-reproductive movements*, which often portray a concrete object or an event; *literal-concretization*, which also describe an event, but have no physical referent, e.g. an emotion; and finally, *major-qualifying movements*, which are poorly integrated with speech and are often disruptive to it.

In their turn, Ekman and Friesen (1969) offered a well structured view of gestures. Their work described a series of studies on body and face movements observed in social interactions as well as in cross-cultural situations. They re-defined and extended the previous classification of gestures developed by Efron and developed a classification that included five categories of body movement. Their first category corresponded to the previously mentioned *emblem*, a term introduced by them, and which refers, as previously noted, to a sign whose meaning is conventionalised and culture-specific. They also included, under the category of *illustrators*, six types of movement, most of which were borrowed from Efron’s classification: *batons*, used to mark emphasis; *pictographs*, movements imitating some of the characteristics possessed by their referents; *kinetographs*, portraying a bodily action; *ideographs*, depicting the course of thought; *deictics*, gestures physically pointing towards their referent; and *spatials*, illustrating spatial relationships. Next to these, a seventh type of gesture was later added by Ekman, *rhythmic movements*, which depict the rhythm or pacing of an event. The third category, *regulators*, include the movements that regulate the interaction and determine turn-taking.

The final two categories developed by Ekman and Friesen fell out the cat-

egory of gesture as defined by Efron. The first of these two categories was *affect displays*, movements revealing the affective or emotional state of the speaker, which are mostly performed by the face, as in the case of an expression of enthusiasm. The initial idea was that these may be universal, and Ekman later renamed this category as *emotional expressions* (Ekman, 1999). The second one, dubbed by them as *adaptors*, are movements performed to adapt to the environment; they include movements such as yawning or adjusting one own's glasses, for example, and were later renamed as *manipulators* by Ekman (1999) (see Table 6, § 2.3.5, for an overview of different gesture classifications).

2.3.4.4 Micro-analyses of gestures in relation to speech

So far, studies on gesture and classifications by different authors encompassed a wide range of body movements and did not explicitly focused on gesticulation (e.g. emblems were also analysed), even if in the second half of the 20th century important insights were gained into how the body movements of gesticulation were related to speech.

Condon's initial studies started in the early 1960's and were intended, similarly as in McQuown's work, to contribute to psychiatric studies. Condon worked together with William Ogston, and both relied on filmed interactions, as previous studies had done. However, differently from them, Condon and Ogston used a time-aligned oscilloscope to thoroughly analyse how the body parts moved in combination and formed hierarchical units in relation to speech. They observed that large movements such as those performed by the head or the arms encompassed smaller ones, such as the tiny movements of eyebrows and hands, which were recognisable in subtle changes of direction and speed within the continuous movement (Condon & Ogston, 1966, 1967).

Condon and Ogston also noticed that some of the rhythmic patterns of body

movements were performed at the smallest level of the linguistic hierarchy, i.e. the phone, while some others corresponded to syllables and words. They described two more levels of bodily rhythmic behaviour, which coincided with verbal stress, and occurred at a half-second cycle and at one-second cycle (Condon, 1976), respectively. This tight relationship between movement and speech was seen by them as a case of *self-synchrony*.

Nevertheless, Condon (1976) observed that not only were the speaker's movements synchronised with their speech, but also that "the listener moves in synchrony with the speaker's speech almost as the speaker does" (Condon, 1976, p. 305)³¹, even in the case of infants as young as 20 minutes old. This effect was named *interactional synchrony* by Condon, who presumed that it might result from a basic and common neurological processing of movement in both speaking and listening, thus allowing a listener to entrain their own rhythmic patterns to those of the speaker.

2.3.4.5 Related hierarchy of gesture and speech

Kendon drew on much of Condon and Ogston's work and focused on the spontaneous gestures that accompany speech. Using the same methodology, which also combined filmed material and the use of an oscilloscope, he took a step forward in the classification of the hierarchical units of body movements, especially manual gestures (Kendon, 1972). In his study, Kendon confirmed that a nested hierarchy of speech phrases was matched by a similar nested hierarchy of gesture units.

³¹ For a more recent account of how listeners' gestures adapt to those of speakers in face-to-face conversation see Mol et al., 2012. This adaptation to another individual's gestures seems to be mediated by a motor copy of the observed body movement through the mirror neuron system (e.g. Montgomery et al., 2007).

For Kendon, the smallest gesture unit and nucleus of the gesture was the *stroke*, which is the distinct effort of the gesture that spans over a small interval of time. Strokes may be preceded by a preparatory phase and are typically followed by a retraction phase. In the optional *preparation*, the articulator, for example the hand, is brought to the point where the stroke is initiated; and in the *retraction* (also called ‘recovery’), the articulator is brought back to a resting position, either to its starting point or to any other point from where the next gesture can eventually be initiated. The typical combination of a preparation and a stroke was defined by Kendon as a *gesture phrase*, which forms a *gesture unit* when the retraction is included and the articulator comes back to a resting position (Figure 8). Furthermore, several gestural units, defined as successive rests of the articulator, can be grouped by a common feature, typically a repeated head movement. Finally, a consistent body posture and a consistent use of the articulator involved, for example the arm, were proposed by Kendon as the highest level of the nested gestural hierarchy.

As observed in Figure 8, the entire gesture unit begins when the speaker starts moving his hand away from its resting position on the table and ends when he returns his hand to the table. During the course of this gesture unit, the speaker carries out the stroke. The preparation of the gesture is initiated when the speaker lifts his hand away from the table to a position in front of him where the stroke is to be performed; for this, the orientation of the hand and the position of the fingers also change. After the stroke, the recovery starts, and the hand returns to its resting position, which is included in the gesture unit but not in the gesture phrase.

The speaker performs a manual action for ‘scattering’ that co-occurs with the verb ‘throw’ and which is semantically coherent (‘co-expressive’ in McNeill’s terms) with the meaning of the verb. Thus, the verbal expression of the utter-

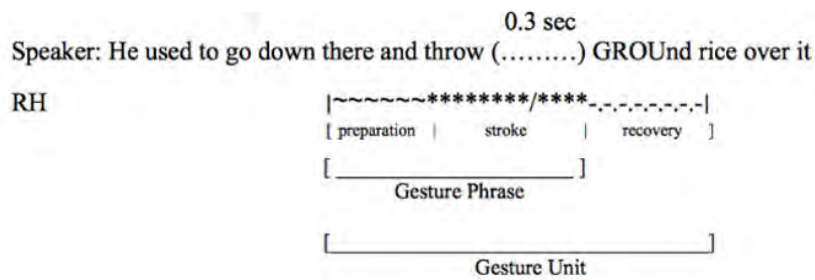
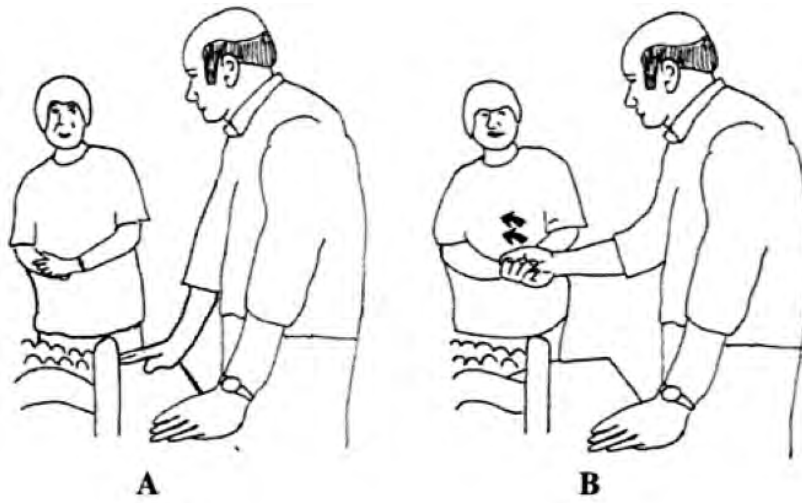


Figure 8: On the left, picture A shows the speaker's rest position before and after the stroke. On the right, picture B shows the stroke of the speaker's gesture phrase. RH stands for right hand; in the displayed sentence, (...) stands for a pause in speech together with its length in tenths of a second; the stressed syllable appears in capital letters. Adapted from Kendon, 2004, p. 114.

ance is exemplified by the movement of the hand. However, the speaker does not necessarily express the same meaning through both modalities. In this case, the verb 'throw' has a general and abstract meaning, and the way in which the action of throwing is carried out ultimately depends on the shape, size, weight and material of what is being thrown. The manual gesture specifies the action of throwing and makes it more precise. The preparation for the gesture is undertaken before the verb is uttered in a combined expression, which shows how the

speaker gets ready to perform the gesture prior to the verbal component of the utterance.

Also, Kendon described in detail the alignment of this gestural hierarchy with intonation. He observed that the stroke coincided with the onset of the *stressed syllable*, or was just previous to it; while a *tone group* corresponded to a gestural phrase. In the case of gestural units, which included a recovery phase next to the preparation and the stroke, the intonation equivalence was a complete sentence (or *locution*, in Kendon's terminology). Finally, consistent head movement, and consistent arm use and body posture corresponded to a *locution group* (i.e. a group of sentences with a common intonational pattern), and to a *locution cluster* (i.e. a paragraph), respectively. As observed in Figure 9, the larger the intonation unit, the greater the change in the gestural hierarchy, perhaps as a form of working memory control.

The gestural hierarchy proposed by Kendon was later extended. Kita (1993) observed the occasional occurrence of both a *pre-stroke hold* and a *post-stroke hold*. Strokes may optionally be preceded by a brief hold before its beginning or followed also by a hold in which the articulator is maintained in the position at which it arrived and allows the stroke to be prolonged. Similarly, subsequent studies called for a more precise analysis of the nucleus of the gesture, and the category of *apex* was introduced (Levelt et al., 1985) as a subphase within strokes corresponding to the peak of effort that occurs at an instant in time, i.e. the "kinetic 'goal' of the stroke" (Loehr, 2004, p. 89). Finally, Kipp (2003) pointed out that in certain occasions, after a forceful stroke, the hand lashes back from the end position of the stroke, and the resulting movement can be interpreted as a *recoil phase*.

The existing synchrony of certain gestural structures with certain phonological structures led Kendon to conclude that there exists a highly specialised

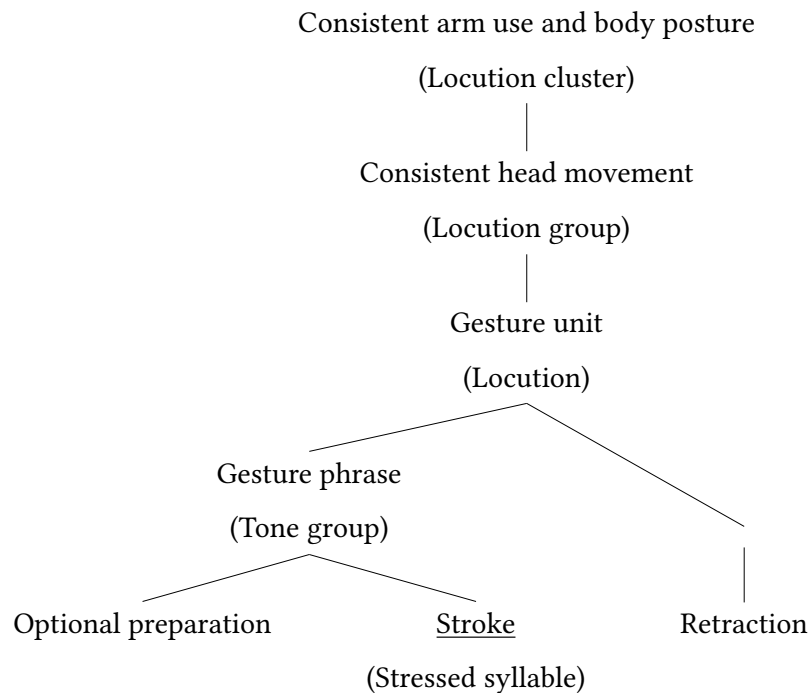


Figure 9: *Gestural hierarchy, matched by the intonational hierarchy (in brackets), as developed by Kendon (1972). Gestural phrases cannot exist without a stroke, the nucleus of the gesture, which appears underlined in the diagram. Adapted from McNeill, 1992, p. 82.*

coordination of body movement and speech. Kendon developed this idea further and went on to suggest that both speech output and kinesic output are two aspects of the same process, i.e. both are the visible forms of a single underlying utterance (Kendon, 1972, 1980) and form an *idea unit*.

2.3.4.6 Gesture and the cognitive foundations of language

The synchrony of spontaneous gestures and speech was started to be seen as a relationship that revealed important aspects of the cognitive foundations of language. Based on Kendon's work, this relationship was explored by McNeill (e.g. 1985; 1992), who also posited that gesticulation is so tightly linked with speech that both must be seen as two aspects of an utterance.

For McNeill, this dialectic relationship is seen in cognitive processes involved in speech, imagistic thinking, and linguistic categorial thinking. He proposed the term *growth point* to refer to the starting point from which the utterance unfolds and develops into a combination of visible gesture and speech:

“The growth point is the speaker’s minimal idea unit that can develop into a full utterance together with a gesture [...]. The content of the growth point tends to be the novel departure of thought from the presupposed background. It is the element of thought that stands out in the context and may be the point of greatest relevance.

The concept of the growth point unites image, word and pragmatic content into a single unit. In this way it is a unit with the properties of the whole and could initiate the microgenetic process of the utterance formation. It is also a unit that encompasses properties of an opposite character—both imagistic *and* linguistic, idiosyncratic *and* social, global *and* segmented, holistic *and* analytic—and this is a source of instability [which is the foundation of its dialectic nature]” (McNeill, 1992, p.220).

This view is against the idea that an utterance is built in a linear way, piece by piece, as proposed by Levelt (1989). Rather it defends that an utterance contains from its very beginning what is to be expressed and combines different modes of representation at an early computational stage. Both imagistic thinking and thinking involving linguistic categories engage in a cognitive process that grows towards a shared expressive end.

In McNeill’s view, several facts support the notion of a common initial stage of gesture and speech. On the one hand there exist structural and functional parallels between both modes of expression, and language acquisition shows

the synchronised development of speech and gesture. It has also been observed that different types of aphasia may manifest themselves similarly in gesture and speech. On the other hand a synchrony between both modes of expression can be observed at different levels. McNeill formulated several rules that account for this synchrony. His *phonological synchrony rule* is based on Kendon's (1972; 1980) observation that the stroke of the gesture is prior to, or coincides with, the stressed syllable, but it never occurs after it. Also, McNeill's *semantic synchrony rule* predicts that gestures and speech relate to the same idea unit and convey the same meaning simultaneously. According to McNeill, although spontaneous gestures can theoretically convey an unrelated meaning to that of speech, it demands an effort to produce a gesture contradicting the conveyed idea unit. Nevertheless, pauses, multiple gestures, and gestures corresponding to more than one clause can blur the limits of the semantic synchrony rule. Finally, the *pragmatic synchrony rule* reveals, in a similar way, the underlying cognitive connection between gesture and speech. For McNeill, gestures and speech perform the same pragmatic function when they co-occur. This is evident mainly in the case of metaphoric gestures, which may not necessarily relate to the semantic content, but rather have the same pragmatic function. For example, in the spoken utterance 'the lesson went on and on', a circular movement performed with the hand co-expresses the same idea on a pragmatic level.

Nevertheless, McNeill's views on the shared computational stage of gesture and speech and their interdependence were later called into question. For example, it was argued that in order to understand the cognitive relation between both modes of expression it was necessary to analyse in more detail the precise nature of their temporal coordination (Butterworth, 1989; Feyereisen, 1987). Along similar lines, it has been claimed that:

“Simultaneity of two events is not evidence that the two events

are more tightly coupled (i.e. with less variance) than two events that occur at a fixed lag. Evidence that two events exhibit a functional linkage or coupling must come from examining variability” (Leonard & Cummins, 2010, p. 69).

In this sense, according to McNeill (2005), synchrony between gesture and speech lies in the process of conceptualization, which is not necessarily reflected in a strict synchrony in production. Yet, Leonard and Cummins (2010) found evidence of the phonological synchrony rule. In their study, they observed that listeners systematically detected mismatches in co-speech beat gestures if they occurred only 200 ms later than naturally produced ones³². Similar experiments analysed and confirmed synchrony of gesture and speech at semantic, functional, and prosodic levels (e.g. Esteve-Gibert & Prieto, 2013; Ishi et al., 2014; Kim et al., 2014; Loehr, 2012; Özyürek et al., 2007; Rochet-Capellan et al., 2008).

Following McNeill’s insights, others have also put forward a different theoretical model to account for the cognitive integration of gesture and speech. For example, Kita (2000) suggested that both modalities do not belong to the same cognitive process, but are separate processes whose close coordination results from their interaction at an early stage, when information is packaged, organised, and distributed across both modalities. Different from McNeill’s *growth point*, in Kita’s *Information Packaging* hypothesis, gestures are produced by what he calls ‘spatio-motoric thinking’, as opposed to the ‘analytic thinking’ of speech suggested by McNeill. Kita’s view assumes that gesture is involved in the conceptualisation of information for speaking and both forms of thinking are combined into a single utterance. Thus, an increased use of gesture is predicted by a difficulty in conceptualizing information (Kita & Davies, 2009). This increased use of gesture is also predicted in the case speakers have strong visuo-spatial skills but

³² See note 38.

weak verbal skills (Hostetter & Alibali, 2007), or even when new information is introduced into the dialogue (Bergmann & Kopp, 2006).

In De Ruiter's view (2000), and contrary to McNeill, gesture and speech do not relate to each other in a dialectic way at an early computational stage. In his view, which is an extension of Levelt's (1989) linear information processing model, gestures arise in a process that is parallel to the different stages that produce an utterance. Each stage is realised by individual modules, which are connected to one another in a linear sequence. According to De Ruiter, the parallel generation of gestures expresses the parts of the idea unit that cannot be expressed linguistically, in a process that he calls *sketching*.

Differently from the previous views, the production of gestures have also been associated to verbal disfluencies and to hesitation pauses or pauses before words indicating problems with lexical retrieval (Butterworth & Beattie, 1978; Dittmann & Llewelyn, 1969). Based on this evidence, the *Lexical Retrieval* hypothesis (Krauss & Hadar, 1999) postulates that both gesture production and speech production are rooted in separate memory representations, i.e. visuo-spatial and propositional, respectively. In this case, the semantic features of both gesture and speech are thought to be activated and processed by each computational system without any coordination between them. Cross-modal interaction of gesture and speech is believed to result when features happen to be processed by both systems at a later stage, when a selected gesture can prime words during lexical retrieval. Nevertheless, different studies have yielded results that contradict some of the basic principles of the Lexical Retrieval hypothesis. For example, it was observed that speakers performed fewer gestures, not more, when there were filled pauses—pauses used by the speaker to search the next word, i.e. interruptions in the flow of speech such as 'uh', 'ah', 'er', and 'um' (Christenfeld et al., 1991). Similarly, in an experiment where speakers were forced not to gesture,

speech was not found to be altered (Hoetjes et al., 2014). At any rate, the formulation of the Lexical Retrieval hypothesis came to stir up the debate around the primary role of gestures, i.e. whether they are a fundamental part of the speech production process or whether rather they serve an expressive communicative purpose.

2.3.5 Categorising gestures

Different schemes aiming at ordering and classifying gestures may be traced back to the Antiquity, as in the case of Quintilian's categorisation based on the body part involved in the execution of the gesture and its functional character. More recently, the renewed interest that the subject experienced since the middle of the 20th century made this classificatory effort more explicitly categorical. The criteria used for classifying the phenomenon of gesture are extensive; however, most authors include a basic division between gestures referring to an object by pointing at it (deictic gestures) and gestures characterizing the object in some way (representational gestures). Several authors also include a category for expressive gestures accounting for a state of mind or a mental attitude, and in many cases, they discuss gestures that punctuate or make reference to aspects of the structure of the discourse, either to its phrasal organization or to its logical structure. Thus, the terminology used in the literature is changing, and authors differ in the emphasis they place on different possible criteria, which may include:

“[W]hether [gestures] are voluntary or involuntary; natural or conventional; whether their meanings are established indexically, iconically, or symbolically; whether they have literal or metaphorical significance; how they are linked to speech; their semantic domain – for example, gestures have been divided into those that are ‘objective’, serving to refer to something in the external world, and those

that are ‘subjective’, serving to express the gesturer’s state of mind. Gestures have also been classified according to whether they contribute to the propositional content of discourse, whether they serve in some way to punctuate, structure or organise the discourse, or indicate the type of discourse that is being engaged in; and whether they play a primary role in the interactional process, as in salutation, as a regulator in the process of turn-taking in conversation, and the like” (Kendon, 2004, p. 84).

Kendon (2004) himself, for example, analysed spontaneous gestures along several categories; for example, in relation to their referential meaning, and in relation to their function. In the latter case, he distinguished between *substantial* gestures, which contribute to the semantic content of the utterance, and *pragmatic* gestures, i.e. those that convey aspects of the situational embedding.

Kendon also described four gestures families based on their kinesic characteristics and their formation patterns: firstly, two precision grip actions, i.e. *grappolo* (or *g-family*), in which all the fingers are brought together until they are in contact; and the *ring* (or *r-family*), in which the thumb and index finger are put into contact at their tips (Figure 10).

And secondly, two families of the open hand, in which the hand shape is ‘open’ but fingers are not spread, i.e. the *palm-down family*, in which the forearm is always in a prone position, so that the palm of the hand faces either toward the ground or away from the speaker (used when something is denied, negated, interrupted or stopped, whether explicitly or by implication); and finally, the *palm-up family*, in which the forearm is always supine, so that the palm of the hand faces upwards (used when the speaker is offering, giving or showing something or requesting the reception of something).

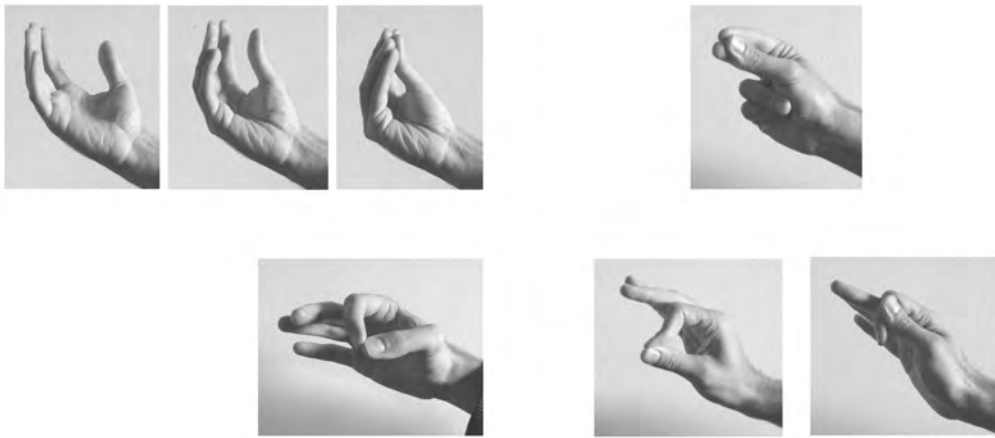


Figure 10: Examples of the precision grip actions described by Kendon (2004): in the upper row, the kinesics of the grappolo family (g-family), with a variant on the right; in the lower row, three examples of the ring family (r-family).

McNeill's classification

In previous sections, the classifications developed by some authors in the 20th century have been briefly discussed, e.g. Efron (§ 2.3.4.1), Ekman and Friesen (§ 2.3.4.3), and Freedman and Hoffman (§ 2.3.4.3). Efron's work, for example, although did not categorise gestures systematically, was the starting point for Ekman and Friesen's own classification. Both have important similarities, such as the various ways in which gestures relate to speech and how they convey meaning. On the contrary, the classification of Freedman and Hoffman has a different orientation, mainly because it was based on gestures that were observed during psychotherapy sessions. For example, Freedman and Hoffman's *object-focused* movements were categorised according to their degree of information content and their integration with speech. Differently, McNeill worked on gestures performed by participants engaged in storytelling. His classification, which has been widely adopted, followed the tradition of both Efron, and Ekman and Freedman, and is based on the symbolical expressiveness of spontaneous gestures in relation

to the context created by the speech they co-occur with.

As previously mentioned, McNeill made a series of distinctions based on Kendon's work, which are splendidly summarised in the Kendon's continuum. They include: (1) how gestures relate to speech (Figure 4); (2) the extent to which they have linguistic properties (Figure 5); (3) the extent to which they are conventionalised (Figure 5); and (4) how they contrast in terms of their semiotic properties (Figure 7).

At the left end of the continuum, gesticulation always appears temporally coordinated with speech, lacks linguistic properties, and its kinesic features are not conventionalised. Consequently, they have special semiotic characteristics, i.e. they are both global and synthetic: the whole meaning of a gesture determines the meanings of each of its parts, and each of its parts are synthesised into a single gesture. These gestures, as studied by McNeill (1992), do not combine with each other to form larger, hierarchically structured gestures; rather, when they occur successively within a clause, each corresponds to an idea unit. Additionally, the fact that gestures at the left end of the continuum are both created at the time of speaking and are not determined by a code makes them incorporate only the most salient aspects of the context in which they occur. Thus, a speaker's gesture for a precise referent may show several forms to highlight a variety of contextual aspects at different moments. As a result, this lack of standards in the production makes gestures become idiosyncratic, i.e., they largely vary from speaker to speaker.

In his classification, McNeill established a fundamental division between *imagistic* gestures and *non-imagistic* gestures. The former are those that depict the shape of an object, display an action or represent some pattern of movement. Depending on whether they are concrete or abstract, they can be grouped into: *iconic*, which display, in the form and manner of their execution, aspects of the

same concrete scene that is presented in speech; and *metaphoric*, which display an image, either as a shape or a movement, that represents or stands for some abstract concept.

Non-imagistic gestures include pointing gestures, rhythmic gestures that highlight either segments of the discourse or its rhythmic structure, and gestures associated to speech failures. Firstly, pointing movements, i.e. *deictic gestures*, are prototypically performed with the pointing finger, although any extensible object or body part can be used, including the head, nose, chin, as well as manipulated artifacts. Deictic gestures relate utterances to the circumstances of space and time in which they occur. Secondly, within rhythmic gestures, McNeill includes *beats*, which are movements that do not represent a discernible meaning and are simple up-and-down or back-and-forth movements of the hand; and *cohesives*, which are used by the speaker to join together thematically related, but temporally separated, parts of the discourse. For McNeill, cohesives can be imagistic or can be performed as pointing gestures. A final category is what McNeill calls *butterworths*, which are typically made when a speaker is trying to recall a word or another verbal expression (Table 6).

Although McNeill's classification has been widely used by scholars, several points concerning his categorisation have been raised. For example, Kita considered that iconics and abstract deictics should be grouped into the category of *representational* gestures:

“Representational gestures are defined here as iconic gestures and abstract deictic gestures (McNeill, 1992). In an iconic gesture there is a certain degree of isomorphism between the shape of the gesture and the entity that is expressed by the gesture. An abstract deictic gesture points to a seemingly empty space in front of the body, as if establishing a virtual object in the gesture space or point-

Literature review: prominence and language

McNeill	Efron	Ekman and Friesen	Freedman and Hoffman	Brief description
Non-imagistic	Logical-discursive	Illustrators	Object-focused	
Deictics	Deictics	Deictics	Literal-reproductive ³³	Pointing gestures
Beats	Batons	Batons	Punctuating	Bi-phasic movements without meaning
		Rhythmics	Minor and major qualifying	Rhythmic movement
Cohesives	–	–	–	Movements to join separate parts of the discourse
Butterworths	–	–	Speech failures	Movements made in a speech failure
<hr/>				
Imagistic				
Metaphorics	Ideographics	Ideographs Spatial	Literal-concretization –	Movements portraying an abstract idea or emotion Movements depicting spatial relationships
<hr/>				
Objective				
Physiographics:				
Iconics	– Iconographics – Kinetographics	Pictographs Kinetographs	Literal-reproductive	Movements portraying a referent Movements depicting a bodily action

Table 6: Summary of gesture classifications that correspond to McNeill’s categories for spontaneous co-speech gestures, i.e. manual gesticulation. Adapted from *McNeill, 1992, p. 76* and *Loehr, 2004, p. 31*.

ing at such a virtual object. Because these gestures have a relatively transparent form-function relationship, they play an important role in communication” (Kita, 2000, p. 162).

Different views on what iconic gestures should include are abundant. In this sense, it has been pointed out that ‘iconicity’ is a concept from the field of semiotics, which has its “own specific, and at times complex, meaning in the history

³³ Ekman and Friesen themselves comment on Freedmann and Hoffman’s category: “They included [...] literal-reproductive movements (which would include our sub-types of deictic, iconographic, kinetographic, pictographic), and literal-concretization (which would be our ideographic)” (Ekman & Friesen, 1969, p. 68). However, in this extract, Ekman and Friesen’s surprisingly mentioned an inexistent ‘iconographic’ category within their *illustrators*. It is possible that they might have meant *spatials* or that they referred to one of the possible aspects of coding meaning in gestures.

of ideas” (Mittelberg & Evola, 2013, p. 1740). Thus, it is possible to find different iconic relations between gesture and speech. As Mittelberg and Evola themselves (2013, p. 1741) sum up, these iconic relations can occur:

1. Between an individual gestural sign carrier and what it evokes or represents, e.g. *iconic gestures, representational gestures*.
2. Between gestures and the concurrent speech content as well as prosodic contours.
3. Gestural behavior of interlocutors, e.g. mimicry (Kimbara, 2006).
4. As iconic patterns emerging from gestural forms recurring within the same discourse, e.g. locution clusters (Kendon, 1972).
5. Across discourses and speakers, e.g. recurrent gestures (e.g. Bressem, 2013) and geometric and image-schematic patterns in gesture space (e.g. Cienki, 2005).
6. Across different languages (e.g. Özyürek et al., 2005).
7. Across different age groups, clinical groups, social groups, or cultures (e.g. Cocks et al., 2013).

2.3.6 Head movements and facial expressions

Most of the work discussed above concentrated on the arm and hand movements typically found in gesticulation, and researchers consequently based their classifications on them. There have also been some efforts to classify gestures performed with the head, especially as part of the nonverbal behaviour studied

in psychiatry (e.g. [Freedman & Hoffman, 1967](#)). Head gestures performed in face-to-face conversation, similar to manual gestures, are non-conventionalised and lie at the left end of the Kendon's continuum.

Analyses of head gestures have reported that consistent head movements are associated to the production of several sentences sharing a common intonation pattern (e.g. 'locution group', a group of sentences with a common intonational pattern) ([Kendon, 2004](#)). The interrelation between head movements and speech was explored in detail by [McClave \(2000\)](#), who identified semantic, discourse, and communicative functions associated to head gesturing, following the well-established tradition of micro-analysing movements using video-recordings:

“Speakers predictably change the position of their heads at the beginning of direct quotes, and for alternatives or items in a list. In narration head orientation functions to locate a referent in abstract space. Because such deictic head movements precede verbalization itself, they cannot be nonverbal translations of speech. Thus, like manual gestures such movements are manifestations of core concepts that are expressed both nonverbally and verbally ([McNeill, 1992](#)). Some speaker head nods have been shown to have an interactive function in triggering backchannels. Listeners are extraordinarily sensitive to such signals and respond within a second” ([McClave, 2000](#), p. 876).

For [McClave](#), both head nodding and head shaking, associated to positive and negative responses, respectively, are the most common semantic function together with a sweeping lateral movement of the head to refer to 'everyone' or 'anything'. In her study, [McClave](#) finally concluded that there might exist cross-linguistic differences in some of the head movements she studied, especially in

the nods functioning as backchannel requests (i.e. phatic communication from the listener), while some other movements might be universal, such as deictic movements performed with the head that serve to locate referents in abstract space.

Nonetheless, differently from manual gestures, and partly due to the bio-mechanical constraints of head gesturing (Figure 11), movements performed with the head are often regarded as serving pragmatic rather than semantic functions.

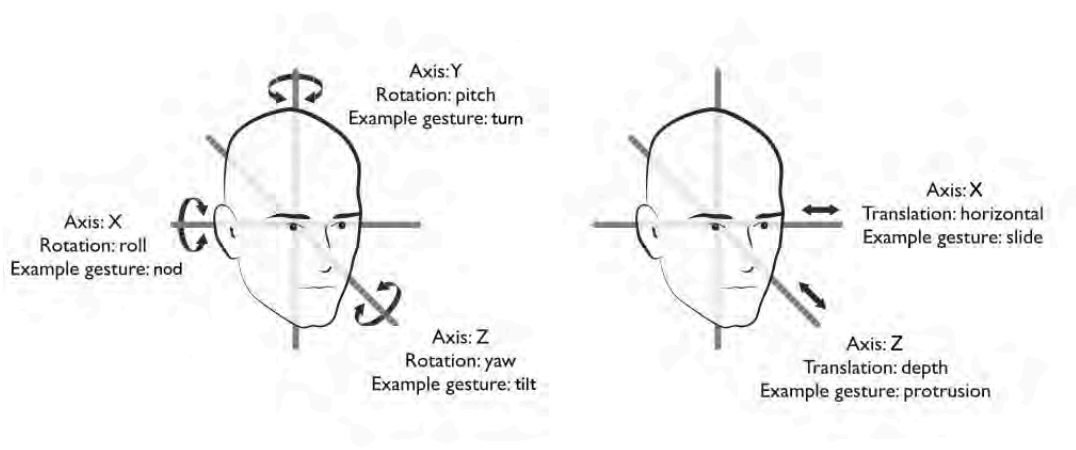


Figure 11: *Schematic overview of the biomechanics of the head. Movements along the three axes and their respective rotations are accompanied by examples of gestures performed in head gesturing. Adapted from Wagner et al., 2014, p. 212.*

Head gestures are performed around the three axes X, Y, Z corresponding to rotations known as ‘roll’, ‘pitch’, and ‘yaw’, respectively. Rotation around the X axis in up-down movements, i.e. nods, and down-up movements, i.e. jerks, are responsible for a head gesturing predominantly performed by the listener as a backchannel signal. Despite the fact that head nods and jerks are more numerous in the role of listener, they can also be performed by speakers, probably with important cross-linguistic differences, as in the case of Japanese head nods in conversations (Ishi et al., 2014). Similarly, lateral head shakes are performed around

the Y axis, and angular rotations around the Z axis result in a tilt movement, ie. the lateral bending of the head. Apart from the direction, the ‘cyclicity’ of some head movements also play an important role in distinguishing among apparently similar head gestures. For example, repeated lateral head shakes contrast with single turns of the head to either side (Heylen, 2008). In this sense, differences between linear and cyclic movements are associated to different pragmatic functions, e.g. single and repeated nods relate to turn-taking and responses to questions, respectively, which again show cross-linguistic differences (e.g. for Swedish, Cerrato, 2007; for English, Hadar et al., 1985; for Japanese, Ishi et al., 2014).

Head movements also convey attitudinal and emotional information. For example, a single and rapid head nod might signal impatience (Hadar et al., 1985). In this sense, head gestures closely interact with facial expressions, especially eyebrow movements.

The expressivity of the face in conversation was initially addressed by Ekman and Friesen (1969), who created the category ‘affect displays’. In a line that can be traced back to Darwin (1872) (§ 2.3.3.3), Ekman and Friesen proposed the existence of several universal primary affects that are expressed through the movements of the facial muscles: happiness, surprise, fear, sadness, anger, disgust, and interest. Being aware of the criticisms that their categorisation might raise, Ekman and Friesen stated that the question of these universals “is not as yet a proven fact, and many would disagree”³⁴ (1969, p. 73). For Ekman and Friesen, these primary affects play several roles in communication:

“Facial behavior in general, and affective displays in particular,

³⁴ Certainly, the controversy was considerable at the time, although the debate around which emotions constitute universal facial expressions is still open. See, for example, Crivelli et al., 2016 or Jack et al., 2012.

receive great attention and external feedback from the other interactant" [...]. Affect displays can be related to verbal behavior in a number of ways. They can repeat, qualify or contradict a verbally stated affect, or be a separate, unrelated channel of communication" (Ekman & Friesen, 1969, p. 77).

However, a different line of research from the study of the communicative effect of emotions as expressed through facial expressions came from the observations that gesture and speech are temporally coordinated. Initial insights gained in the 60's and 70's pointing to this synchrony were previously discussed (§ 2.3.4.4 and § 2.3.4.5). In the next section, a more detailed account of the time-alignment between manual gestures and speech is offered. The interaction of both head and eyebrow movements with speech will be dealt with in § 2.3.8.1.

2.3.7 Temporal coordination of gesture and speech

As mentioned earlier, the claimed synchrony between gesture and speech was put to the test for a better understanding of how both modes of expression relate to each other. Both gesture production and speech production had been shown to have a certain degree of synchronisation (e.g. Kendon, 1972), which led to hypothesise about their cognitive interdependence (e.g. Kita, 2000; McNeill, 1992). Subsequently, the task of precisely measuring the alignment of gesture and speech was based on the estimation of gesture effort maxima³⁵ that co-occurred

³⁵ In this regard, several criteria appeared in the literature. For example, Kita et al. argued that: "Different types of phases can be identified by different foci of 'effort' [...]. In a preparation and a retraction, the effort is focused on reaching their respective end points (the beginning of the following stroke and the resting position). In contrast, in a stroke, the effort is focused on 'the form of the movement itself – its trajectory, shape, posture.' (McNeill, 1992, p. 376)" (Kita et al., 1998, p. 27).

with speech: either with stressed syllables (e.g. McClave, 1991; Tuite, 1993), with the lexical affiliates of the gesture (e.g. Levelt et al., 1985; Morrel-Samuels & Krauss, 1992; Schegloff, 1984), with focused words (e.g. Butterworth & Beattie, 1978), with intonation peaks (e.g. De Ruiter, 1998, experiment 2), or both simply co-occurred without being temporally aligned (McClave, 1994). Methodological differences yielded inconsistent results on the precise temporal coordination of gesture and speech, and as a result several views on the matter were put forward.

2.3.7.1 Precedence of gesture

Initially, Butterworth and Beattie (1978) observed that iconic gestures tended to start in a pause just before the corresponding focused word, whether noun, verb, adverb, or adjective. In a similar way, words in a contrastive focus condition were found by Roustan and Dohen (2010) to coincide with the apex of pointing gestures. They concluded that apexes were aligned with articulatory vocalic targets (in their case, peak of amplitude of either lip opening or lip protrusion) and not with acoustic correlates (peaks of f_0 and intensity), so that “focus and manual gestures are coordinated in the sense that focus ‘attracts’ the manual gesture” (Roustan & Dohen, 2010, p. 4).

Other studies also reported that the onset of gestures generally precedes that of speech. For example, Scheloff (1984) analysed the temporal relationship of gestures with their lexical affiliates (either the accompanying word or a phrase sharing a common meaning with the gesture) and observed that the stroke phase of deictic gestures were generally produced in anticipation (similar results were obtained by Ferré, 2010). Interestingly enough, Morrel-Samuels and Kraus (1992) reported that the precedence of gesture to speech was reduced in proportion to the speaker’s familiarity with the uttered word. Besides, this familiarity was also related to the duration of the gesture, which was observed to be shorter as

the word was more familiar to the speaker. The role played in the temporal coordination of gesture and speech by the semantic content was also explored by Bergmann and her colleagues (2011). They found that the asynchrony of gesture and speech decreased when both modalities were redundant and expressed the same content, while it increased when gesture complemented the semantic content of speech.

2.3.7.2 More accurate measurements

In an attempt to gain precision, Loehr (2004; 2012) measured the temporal coordination between accented stressed syllables occurring in spontaneous dialogues and the apex of several types of gestures: deictics, iconics, metaphoric, beats, head movements, and emblems. In his results, Loehr observed that apexes of all gestures (a more exact temporal measure within strokes corresponding to an effort maximum) significantly tended to align with accented stressed syllables. Similarly, in a study conducted by Jannedy and Mendoza-Denton (2005) using filmed material of a political speech, there were almost no apexes that did not coincide with a pitch accent. Additionally, Flecha-García (2006; 2007), using spontaneous speech material, found statistical significance for the time-alignment of eyebrow raises and surrounding pitch accents, especially following accents. In a similar way, alignment between pitch and intensity maxima of accented stressed syllables was also reported by Nobe (1996, experiment 4), who analysed deictic, iconic, and metaphoric gestures, following McNeill's methodology of recording the narration of cartoon films by participants.

Furthermore, Rochet-Capellan and her colleagues (2008), in a study to establish more accurately the synchronization between gesture and speech, analysed the alignment of deictic gestures with jaw movement. They measured the jaw-opening maximum in the stressed syllable of oxytone and paroxytone words and

what they called the ‘pointing plateau’, i.e. the amount of time the finger remained pointing at the target. Their results showed that the alignment varied according to the different stress patterns: in paroxytone words, the maximum opening of the jaw aligned with the beginning of the pointing plateau, while in the case of oxytone words, it aligned with the end.

Conversely, De Ruiter (1998) obtained contradictory results for the influence of different stress patterns in the temporal coordination of gesture and speech. In a first experiment analysing the alignment of deictic gestures with different stress patterns in disyllabic and trisyllabic words, he observed that the temporal alignment of the apexes was not affected by stress position. However, in a second experiment, De Ruiter found that the stress pattern did affect the temporal coordination of gesture and speech in a context of contrastive focus in adjective-noun phrases. He actually observed that the duration between the beginning of the preparation and the apex was longer in the case of oxytone words that were accented in phrase-final position. Additionally, this was the only context in which the apex preceded the stresses syllable without overlap (Figure 12).

A similar study in a context of contrastive pitch accents and different syllable position was conducted some time later by Rusiewicz (2010) in her doctoral thesis. She analysed the alignment of the apex of deictic gestures with the utterance of disyllabic compound words (e.g. words sharing the first syllable: bathrobe, bathtub; and sharing second syllable: suitcase, briefcase) whose stress patterns were manipulated in production by participants with a contrasting pitch accent. Differently from De Ruiter (1998), Rusiewicz observed that the timing of the gesture was not influenced by the actual stress pattern because the apex of gestures was time-aligned with the onset of the target word rather than with the prosodically prominent syllables.

³⁶ As De Ruiter explains: “The reason that the stressed syllable in adjectives with final stress

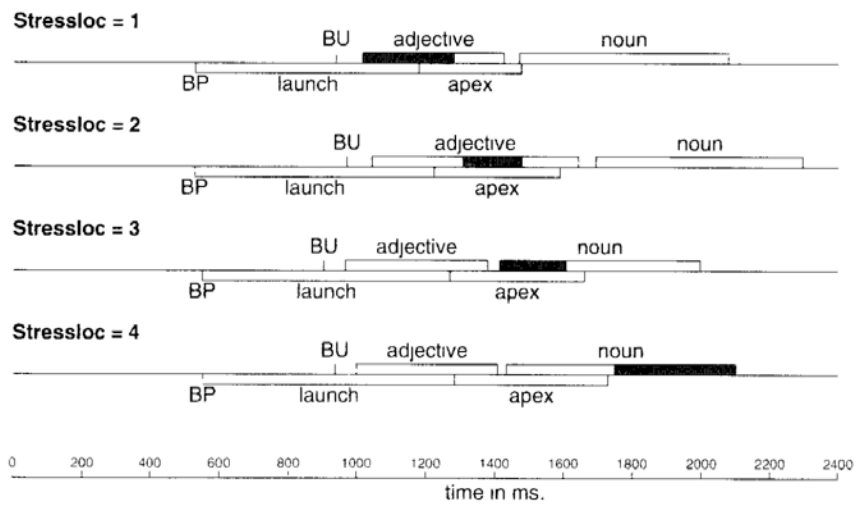


Figure 12: Results of De Ruiter's (1998, experiment 2). Apex becomes longer towards stress location 4, where there is not overlapping. Abbreviations: Stressloc (stress location; 1 and 2 correspond to contrastive focus in adjectives, and 2 and 3, in nouns), BP (beginning of preparation), BU (beginning of utterance). The black bar corresponds to the temporal location of the stressed syllable³⁶. Adapted from De Ruiter, 1998, p. 64.

2.3.7.3 On the applicability of the Phonological Synchrony Rule (PSR)

Some of the obtained results in different studies called into question the strict application of McNeill's (1992) 'phonological synchrony rule' (PSR). For example, around the time McNeill's formulated the rule, McClave (1991; 1994) conducted a series of experiments and did not find any evidence of temporal coordination; actually she observed that beat gestures occurred with stressed syllables as frequently as with unstressed syllables.

Consequently, Karpiński et al. (2009) set out to explore the applicability of McNeill's rule in an analysis of the temporal coordination of gesture phrases and are not at the end of the word is that in Dutch, adjectives are inflected when they are preceded by definite articles. For example, the root form /vi-o-lEt/ then becomes /vi-o-lE-tə/" (De Ruiter, 1998, p. 64).

intonational phrases in Polish task-oriented dialogues. The temporal coordination between the stroke of gestures and what they labelled as ‘strong prosodic prominence’³⁷ was observed in only 5% of cases for a short difference of 100 ms³⁸ between the beginning of the stroke and the beginning of the prominent syllable, and in 40% of cases for a 200-millisecond difference. They reported several reasons for the failed occurrence of the phonological synchrony rule, e.g. stroke repetitions, inertial echoes, ambiguous hand excursions. For gestural phrases and intonational phrases, Karpiński and his colleagues found these did not synchronise very precisely, although more coordination was observed between them, especially because gestures partly overlapped with co-occurring major intonational phrases.

2.3.7.4 Redefining alignment landmarks

Such inconsistent results called for a closer examination of the landmarks that could serve as targets of coordination between gesture and speech. Four studies offered stronger evidence of the exact nature of this temporal coordination.

Firstly, as mentioned above, Leonard and Cummins (2010), in a production-perception study, analysed the variability of several kinematic landmarks of beat gestures together with three potential anchor points in the co-occurring speech. Their choice of beats (also known as ‘batons’) was motivated by the fact that beat gestures integrate more precisely into the continuous stream of speech due both to their lack of meaningful content compared to iconic and metaphoric gestures

³⁷ In the absence of a more precise definition of their categories ‘strong’ and ‘weak’ prosodic prominence, it is possible that ‘strong prosodic prominence’ might refer to the accented stressed syllables of the intonational phrases they analysed.

³⁸ For the sake of comparison, and as Rusiewicz conveniently puts it: “100 ms is the benchmark for software developers so that processing will seem instantaneous and is also the human reaction time for advanced athletes” (Rusiewicz, 2010, p. 203).

and to its distinct kinematic strength and time of occurrence. Authors selected five kinematic landmarks for each gesture, i.e. the onset of the gesture, the peak velocity of the extension phase, the point of maximum extension of the hand before recovery, the peak velocity of the recovery phase, and the end of the gesture. Each of these was compared to three landmarks of speech, i.e. the vowel onset of the stressed syllables, the estimated P(erceptual)-centre (also known as ‘stress beat’)³⁹, and the maximum pitch of the stressed syllable. Leonard and Cummins, by analysing the variability of each landmark, observed that not all gesture landmarks related equally to the measured speech landmarks. For example, the apex of beat gestures showed the least variability in its relative timing regardless of the examined anchor points of speech. Besides, the pitch maximum of accented stressed syllables was found to be the closest speech landmark to the apex of beat gestures, in line with Loehr’s (2004) previous observations.

Secondly, a study by Esteve-Gibert and Prieto (2013) reviewed previous research on the topic and pointed out that the differences found in the literature stem from a heterogeneity of temporal measures. In their turn, they conducted an analysis of temporal coordination of deictic gestures and speech in Catalan, similar to previous studies involving contrastive focus in three different stress patterns, i.e. oxytone, paroxytone, and monosyllabic words (e.g. De Ruiter, 1998, experiment 2; Rusiewicz, 2010). Then, differently from Rusiewicz, for example, who took the vowel midpoint as speech landmark in her comparison with the different gesture landmarks, Esteve-Gibert and Prieto chose the f_0 peak of the accented stressed syllable, i.e. the intonation peak. Additionally, they measured the onset and the duration of the preparation, stroke, and retraction phases of

³⁹ “P-centres are obtained by estimating the energy over the frequency range of the first two formants to identify the sonority rise at the onset of the nuclear vowel. A [stress] beat [i.e. a P-centre] is defined as occurring halfway through this rise” (Cummins & Port, 1998).

gestures together with the point in time when the apex occurred. In their results they observed that, on the one hand, the position of the apex and that of the intonation peak were the two measurements with the strongest correlation, i.e. gesture apexes were anchored in intonation peaks. On the other hand, Esteve-Gibert and Prieto also reported that the timing of gesture apexes and intonation peaks was similar and was determined by the prosodic structure of the utterance, in line with previous results obtained by [Shattuck-Hufnagel et al., 2007](#); [Yasinnik et al., 2004](#). In their case, they found that both apexes and intonation peaks anticipated respect to the end of the accented syllable in phrase-final position (monosyllables and oxytone words), while both were retracted when the accented syllable was in non-phrase-final position (paroxytone words) (Figure 13).

Thirdly, Krivokapić et al. ([2015](#); [2016](#)) conducted a series of studies to also analyse the temporal alignment of gesture and speech including different prosodic structures, i.e. contrastive focus, narrow focus, broad focus, and also a deaccented context (a detailed account of all experiments and results is in [Krivokapić et al., 2017](#)). Krivokapić and her colleagues contributed to the debate on gesture-speech alignment with the recording of the simultaneous kinematics of both arm movement during the production of a deictic gesture and vocal articulatory gestures by means of motion tracking and electromagnetic articulography (EMA), respectively. Target words differing in their stress patterns (oxytone vs. paroxytone) were elicited in the four prosodic structures mentioned. Similar to other experimental tasks (e.g. [Esteve-Gibert & Prieto, 2013](#); [Rochet-Capellan et al., 2008](#)), participants (N = 2 for the data collection used in all three experiments; see [Krivokapić et al., 2017](#) for details) had to read sentences on a screen and point with their index finger at the pictures displayed as they read the associated target words.

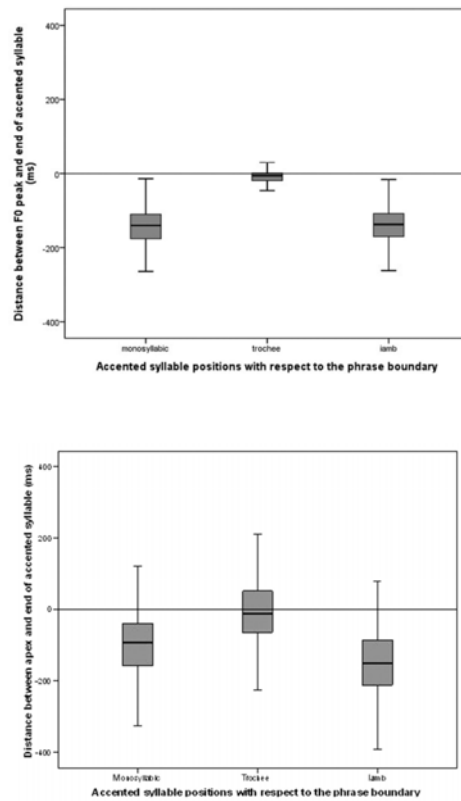


Figure 13: Results of Esteve-Gibert and Prieto (2013, pp. 857-858) showing the time difference (in milliseconds) of both the intonation peak (upper panel) and the gesture apex (lower panel) respect to the end of the accented syllable for each of the analysed stress patterns (monosyllabic, paroxytone, and oxytone words). Error bars indicate 95% confidence intervals.

In their first experiment, the authors investigated the effects of phrase-level stress; in their second experiment, they analysed the effects of phrase-initial prosodic boundaries; and in their third experiment, they focused on the effects of phrase-final prosodic boundaries. Overall, their main finding, similar to that of Esteve-Gibert and Prieto (2013), was that the realisation of both manual and vocal gestures was influenced by prosodic structure. In their case, Krivokapić and her colleagues reported lengthening in both modalities at prosodic boundaries, with a local effect that increased with boundary strength.

Finally, and differently from the previous studies just mentioned, an experiment conducted by Parrell et al. (2014) reinforced the idea of the temporal coordination of gesture and speech. In a motor task that involved finger tapping synchronously with speech, speakers were unable to desynchronize tapping and emphatic stress. They observed that emphasis in either speech or manual behavior automatically resulted in lengthening the production in the other domain.

2.3.8 Audiovisual prosody

There has also been a large number of studies on the interaction between gesture and speech that have corroborated the strong connection between the production of movements and the production of speech. For example, eyebrow movements have been found to correlate with f_0 (Cavé et al., 1996), although results for manual beat gestures were more ambiguous (McClave, 1998), which was related to the fact that manual gestures and articulatory gestures are controlled by the same brain areas (e.g. Bernardis & Gentilucci, 2006).

As for speech perception, one line of research has focused on the perceptual effects of the visual component of speech. For example, the so-called McGurk effect (McGurk & MacDonald, 1976) initially established that visual information, more precisely, lip movements, affect speech perception. Later, not only lips, but also the rest of the face (e.g. Pelachaud et al., 1996) was reported to affect speech perception. As a consequence, it was observed that both facial expressions and body movements play an important role in conveying functions traditionally associated to prosody, such as phrasing and emphasis. In parallel, an interest in the visual aspects of speech was developed by researchers working on audiovisual speech synthesis. On the other hand, the use of talking heads resulted in the development of adequate animations not only of lip movement, but also of facial movements to help identify, for example, the most prominent words of the

utterance and its information structure (e.g. Beskow, 1995).

These two lines of research have converged in the study of how the information conveyed in the visual modality interacts with prosody. Partly due to the advances in technical equipment and computer power, this kind of research has become easier and has also gained more attention. Swerts and Krahmer (2005), for example, explored the verbal and visual cues that reveal the listeners' knowledge of precise information after being asked a question. In their paper, they defined the term *audiovisual prosody* in the following terms:

“So far, research has focussed primarily on analyses of verbal features, such as particular intonation patterns, that are encoded in the speech signal itself. *Visual cues*, such as a wide range of gaze patterns, gestures, and facial expressions, are a natural and important ingredient of communication as well. It seems a reasonable hypothesis that these might be informative just as verbal features are, though this issue is largely unexplored [...]. [For t]he basis of combinations of verbal and visual features, [...] we will use the term *audiovisual prosody*” (Swerts & Krahmer, 2005, pp. 81-82).

Audiovisual information has been shown to play a crucial role in a wide range of communicative functions. Generally, the interplay between visual cues and verbal cues has been found to influence both speech intelligibility and comprehension. It is also involved in signalling higher-level pragmatic aspects such as emotion, attitude, and engagement (e.g. Cafaro et al., 2012; de Gelder & Vroomen, 2000; Ekman, 1999). Besides, it has been observed that visual cues can also express communicative elements such as uncertainty and frustration (e.g. Barkhuysen et al., 2005; Swerts & Krahmer, 2005). Such pragmatic aspects are also expressed prosodically in dimensions that include valency and arousal rather than in discrete categories such as anger, fear or joy (Schröder et al., 2001).

Both modes of expression share the potential for nuancing these non-discrete elements of communication. For example, the pitch accent on an utterance can be produced with more or less excursion along an intonation phrase, thus expressing the novelty or importance of the corresponding item to various degrees.

One of the main functions associated to communicative visual cues is that of enhancing traditional prosodic functions, e.g. phrasing, face-to-face grounding, question intonation, and prominence signalling (e.g. [Barkhuysen et al., 2008](#); [Hadar et al., 1983](#); [Krahmer & Swerts, 2007](#); [Nakano et al., 2003](#); [Srinivasan & Massaro, 2003](#)). In this sense, not only manual gestures, but also head movements and eyebrow movements contribute to the prominence lending properties of prosody in order to highlight certain parts of the utterance (e.g. [Granström et al., 1999](#)). For example, eyebrow raises, head nods, and beat gestures increase the perceived prominence of the words they co-occur with, and reduce that of the surrounding words, although this is not restricted to hand or head gestures but any body part can assume this highlighting function ([Bull & Connelly, 1985](#)).

2.3.8.1 Multimodal prominence signalling

Most methods applied to study the interaction of audiovisual and verbal prominence have so far made use of both lip-synchronised animated agents and experimental settings in which gestures are elicited with controlled speech stimuli. In the former case, one of the main advantages of animated agents is the possibility of manipulating visual cues while preserving acoustic information. Conversely, research in prominence by means of elicited gestures in experimental settings also presents some difficulties. Only a minor number of studies have applied spontaneous speech to the study of perception of multimodal prominence. In any case, all these approaches have restricted themselves to analyses of head movements and facial expression, especially eyebrow movements, while just in

some cases the role of hand gestures have been analysed.

2.3.8.2 Studies using an animated agent

Head nod vs. eyebrow raise

Granström and his colleagues (1999), for example, conducted a perception experiment in Swedish, in which an eyebrow raise performed by an animated agent was created and presented with different content words of a given sentence. The eyebrow raising was subtle, so that its perception was distinctive but not too obvious (Figure 14). In total, it lasted 500 ms and included a dynamic raising part (100 ms), a static raised period (200 ms), and a dynamic lowering part (200 ms).

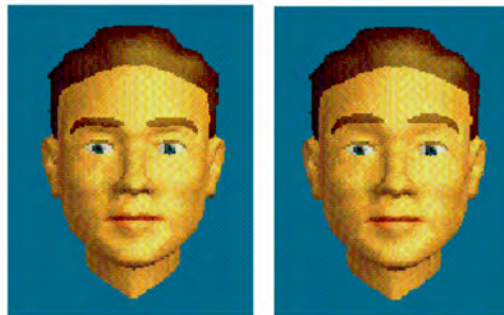


Figure 14: *Animated agent with neutral expression (left) and showing an eyebrow raise (right).*
Adapted from *Granström et al., 1999.*

In their results, Granström et al. reported that those words co-occurring with an eyebrow raise were generally perceived as more prominent, even if no strong acoustic cues of prominence existed. Besides, differences were found for a subgroup of non-native Swedish speakers, who tended to give more prominence marks for those words with a concomitant eyebrow raise.

The same team of researchers carried out a second experiment, which this time also included head nods (House et al., 2001). The objective of the study was to test the strength of head nods vs. eyebrow raises as visual cues of prominence

as well as the perceptual sensitivity in their time-alignment with the syllable they co-occurred with. This time, eyebrow movements were shorter than in the previous experiment, and they lasted a total of 300 ms, divided in three time spans of 100 ms (Figure 15).



Figure 15: 3D animated agent used in House et al.'s experiment. Adapted from House et al., 2001.

As the researchers observed, the combination of head nods and eyebrow raises was a powerful cue of prominence when they co-occurred with the stressed vowel of a potentially prominent word, although head nods were perceptually more relevant than eyebrow raises. Additionally, Granström and his colleagues reported that the integration window of both visual cues with the co-occurring syllable lay around 100 ms, much in line with previous studies on the gesture-speech synchrony (e.g. Karpiński et al., 2009)⁴⁰.

Eyebrow raise as a cue of contrastive focus

In their turn, Kraemer and his colleagues (2002a; 2002b) tested further the role of eyebrow movements in their interaction with pitch accents when cueing contrastive focus. In the creation of their stimuli, they established the duration of eyebrow movements as House et al. (2001) did, in a three-fold sequence of 100 ms per movement phase (Figure 16).

⁴⁰ See note 38.



Figure 16: *Krahmer et al.'s talking head while uttering a stimulus sentence. It shows an eyebrow raise (left) and a neutral expression (right). Adapted from Krahmer et al., 2002a.*

In addition, they used a disyllabic adjective-noun phrase, in which either one of the two words or both words had a pitch accent. Eyebrow raises co-occurred with the first or the second word, so that both the pitch accent and the eyebrow raise sometimes coincided on the same word, sometimes they were placed on different words; sometimes both words carried a pitch accent but only one eyebrow raise on either the first or the second word.

In their first analysis (Krahmer et al., 2002a), participants were asked to report what the preceding (contrasting) utterance might have been, given the information structure of the presented stimuli, e.g. to the stimulus *blaue vierkant* ('blue square'), the possible contrasting sentences were '**red** square' (focus on the first word), 'blue **triangle**' (focus on the second word), and '**red triangle**' (focus on both words). The researchers reported that both auditory and visual cues had an important influence on focus, although this effect was different in magnitude: pitch accents had a larger effect than that of eyebrow raises. Subsequently, in a second analysis (Krahmer et al., 2002b), participants were asked to choose their preferred animation from a pair of similar stimuli as those described above. In this case, the results showed that listeners preferred the occurrence of the eyebrow raise coinciding with a pitch accent, although there was a slight preference for both acoustic and visual cues co-occurring on the first word when both words were accented. In addition, Krahmer et al. (2002b) also observed that visual cues

had an additive effect in the perception of prominence when they coincide with auditory cues, although a clashing effect happened when visual cues appeared immediately before or after an accented word.

Interestingly, the same researchers, using a similar procedure to the one just mentioned, conducted a third experiment and found cross-linguistic differences between Dutch and Italian (Krahmer & Swerts, 2004). In the case of Dutch, speakers tended to prefer audiovisual stimuli when the eyebrow raise co-occurred with the word carrying a pitch accent rather than with the unaccented word (as they had already observed, Krahmer et al., 2002a,b). Conversely, Italians consistently preferred the eyebrow movement co-occurring on the first word, regardless of which of the two words in the phrase was accented. By the same token, it was observed that Dutch listeners were able to detect the focus of the utterance, but Italians were not, as previously observed for auditory cues alone (Swerts et al., 2002) (see § 2.2.5 for a brief description), and similar results were later obtained with Dutch and Spanish speakers (van Maastricht et al., 2016). Finally, the role of eyebrow raises in enhancing acoustic prominence was clearer in Dutch than in Italian. All these differences were explained by prosodic differences between both languages. Thus, Krahmer and Swerts concluded that the location of eyebrow raises was found to be language-dependent and their functional contribution can vary according to each specific language.

Word recognition mediated by realistic 3D facial expressions

A subsequent study on the perception of audiovisual prosody, also using an animated agent, was conducted by Munhall et al. (2004). The researchers prepared a realistic 3D talking-head animation and analysed how head movements interacted with speech in word recognition. More precisely, although they did not test the prominence effects of head movements, Munhall and his colleagues

analysed whether head movements could improve the intelligibility of Japanese sentences in a speech-in-noise task.

Differently from the studies just mentioned (Granström et al., 1999; House et al., 2001; Krahmer et al., 2002a,b), the realistic animated agent was modelled this time from the face and head motions of a Japanese speaker, which had been obtained in a previous study. Using this realistic talking head, which was depicted wearing sunglasses due to the difficulty of modelling eye movement (Figure 17), four conditions were created: (i) normal head motion, (ii) a head motion with twice the amplitude in all directions of normal head motion, (iii) no head motion, and (iv) a blackened screen with audio-only as available prominence cues. Munhall and his colleagues reported that identification of individual words in a set of statement sentences was improved in all the first three conditions compared to the audio-only condition. In their analysis, they remarked that both f_0 and intensity were highly correlated with the kinematics of the head in their animation, which had been preserved from the original recording used to create the 3D avatar.



Figure 17: *Realistic 3D animated agent while pronouncing different speech sounds corresponding to a different position and orientation. Adapted from Munhall et al., 2004.*

Both eyebrow movements and head movements and their interaction with speech prosody were also studied by Al Moubayed and his team in two different experiments (Al Moubayed & Beskow, 2009; Al Moubayed et al., 2010). In their case, they used animations of a talking head that had been previously developed

by his colleagues (House et al., 2001) (Figure 18).

In their first experiment, Al Moubayed and Beskow selected 40 sentences from a corpus—of which they do not provide further details—ranging between 6 and 10 words. These auditory stimuli were manipulated with a noise-excited vocoder and then lip-synchronised with an accompanying animated agent. The sentences had been previously annotated by an expert listener on a 4-point scale. Participants were administered 30 sentences in 5 blocks of 6 sentences (after a first 10-sentence trial), with each block possessing different visual cues of prominence: (i) no gesture, i.e. only visible lips and jaw movements, (ii) a 350-millisecond-long head nod located on the most prominent syllable of the sentence, (iii) an equally-long eyebrow raise also located on the most prominent syllable of the sentence, (iv) an eyebrow raise placed on steep pitch movements, (v) and eyebrow raise placed randomly on syllables in order to test whether these random gestures improved or hindered the perception of prominence⁴¹. In their results, they observed that participants better identified those words in vocoded sentences that were accompanied by a head nod or an eyebrow raise (Figure 19).

⁴¹ Al Moubayed and his team later published a book chapter (2011), where they included a sixth condition in the description of their experiment: (vi) ‘automatically detected prominence with eyebrows’, in which an eyebrow raise was coupled with prominent syllables that were automatically detected using an automatic prominence detection model explained in their book chapter.

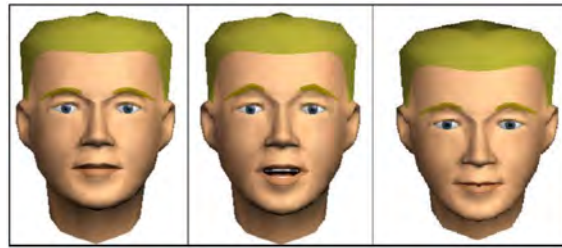


Figure 18: Three different views of the animated agent used in Al Moubayed and Beskow's experiment. The animated character shows no gesture (left), a head nod (centre), and an eyebrow raise (right). Adapted from Al Moubayed & Beskow, 2009.

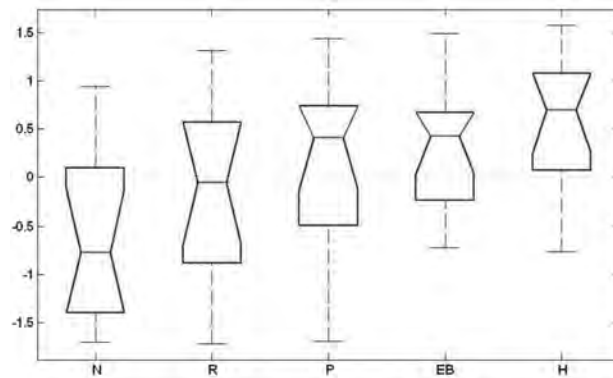


Figure 19: Results of Al Moubayed and Beskow (2009, p. 45) showing normalised percentage score of correct word recognition for the five tested conditions: (N)o gesture; (R)andom eyebrow raise; eyebrow raise on steep (P)itch movement; (E)ye(B)row raise on prominent syllable; (H)ead nod on prominent syllable.

Focus detection using auditory and visual cues

Finally, two more perception experiments analysing the detection of focus by means of both auditory and visual cues were conducted by Prieto and her colleagues (2011)⁴². The perception of narrow focus and contrastive focus were explored using a mismatch paradigm between visual cues and auditory cues. Firstly, the intonation peak of the stressed syllable was manipulated so that it

⁴² A more detailed version of both experiments was published later as a journal paper (Prieto et al., 2015).

ranged, in four successive 1.5-semitone steps, from a narrow focus statement to a contrastive focus statement. Secondly, the visual cues conveyed by an animated agent were changed from a neutral state to a more expressive one by progressively increasing the movement of an eyebrow raise in combination with a head nod. In a forced-choice task, participants were asked to identify narrow focus (as a 'statement' sentence) and contrastive focus (as a 'correction' sentence). In their results, Prieto et al. found that the combination of clear visual cues, i.e. those with the largest movement excursion or with the highest intonation peaks, allowed participants to best detect contrastive focus. Interestingly, visual cues proved stronger cues than auditory ones. Besides, the analysis of reaction times showed that none of the presented combinations of visual cues and auditory cues were perceived by participants as incongruent.

In their second experiment, Prieto and her colleagues studied the independent contribution of both eyebrow raise and head nods to the perception of contrastive focus. For this, they created visual-only stimuli with the combination of both types of gesture involving four degrees of activation, from less pronounced movement to more pronounced movement (Figure 20). The same participants tended to detect contrastive focus more easily this time using the cues conveyed by the movement of the head rather than that of the eyebrows.

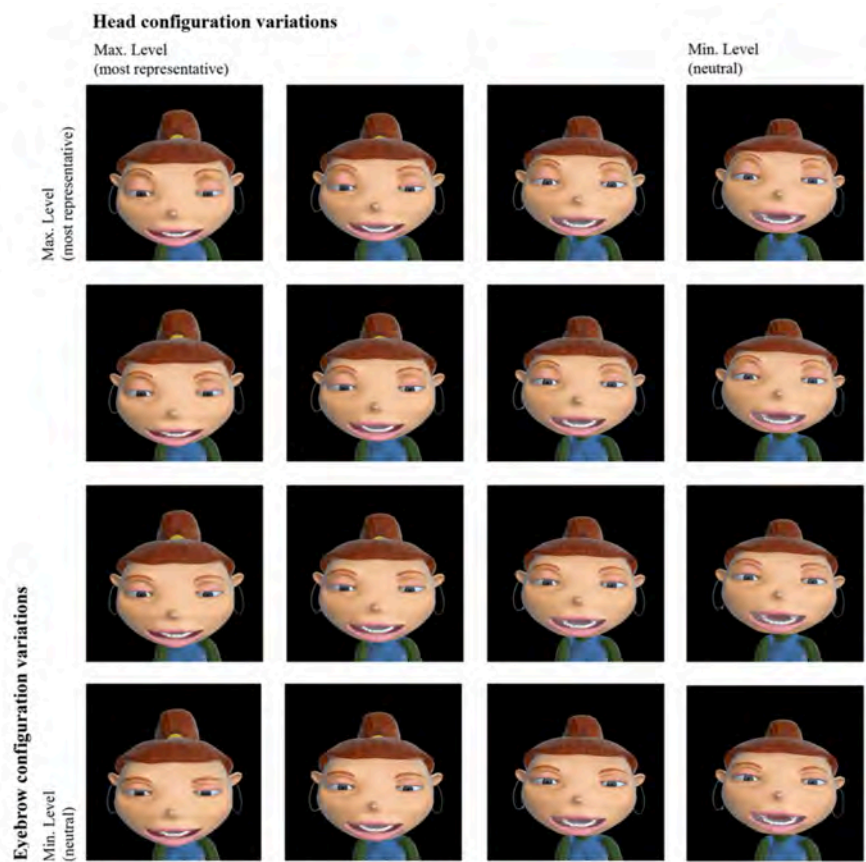


Figure 20: Sixteen stills showing combinations of an eyebrow raise and a head nod varying in their degree of intensity. The four head configurations are displayed in the X axis and the four eyebrow configurations are in the Y axis. Adapted from Prieto *et al.*, 2015, p. 51.

2.3.8.3 Studies in experimental settings

The study of audiovisual prosody in experimental settings involves the elicitation of gestures either by means of dialog-oriented tasks or by overtly requesting participants—sometimes even professional actors—to produce gestures while uttering a series of given sentences.

‘Producing beats, hearing beats, seeing beats’

Swerts and Krahmer, for example, conducted a series of studies in an experimental setting to explore different effects of audiovisual prosody in speech production and perception (Krahmer & Swerts, 2007; Swerts & Krahmer, 2004, 2008, 2010). Differently, from their previous research using an animated agent, Swerts and Krahmer (2004) initially recorded participants uttering a sequence of three nonsense syllables (Figure 21).

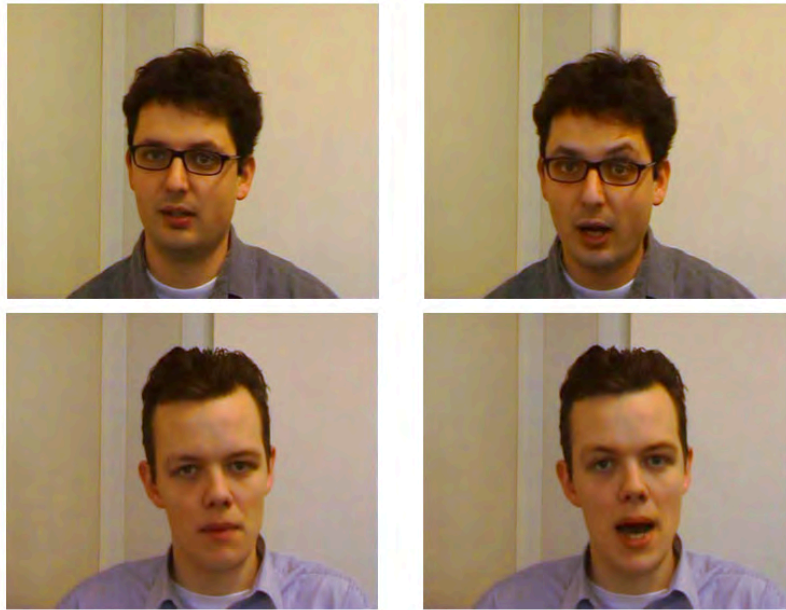


Figure 21: *Stills showing two of the recorded speakers uttering a three-syllable word. On the left, both speakers appear pronouncing an unaccented syllable; on the right, an accented one. Adapted from Swerts & Krahmer, 2004, p. 70.*

Using these stimuli, they conducted two experiments on stress perception. In the first case, the stimuli were administered to listeners in three different conditions, auditory, visual, and audiovisual. In the second experiment, the researchers created a mismatch between both modalities in the audiovisual recordings, so that auditory cues and visual cues never occurred on the same syllables. In

their results, they observed that listeners had no problem to detect the stressed syllable in the first experiment, and that auditory cues were stronger cues of prominence when compared to visual cues in the second experiment.

In a subsequent, larger study, Krahmer and Swerts (2007) centered on beat gestures in order to analyse: the realisation, the auditory perception, and the audiovisual perception of beat gestures. In the first experiment of their study, ‘producing beats’, they asked participants to utter a four-word sentence with one visual beat performed with hands, head, or eyebrows, so that the beat gesture coincided with a pitch accent on either of the two target words (underlined) or on none of them: *Amanda gaat naar Malta* (‘Amanda goes to Malta’). For this, participants were allowed to train until they felt they could not improve the realisation of the gesture accompanying the uttered sentence. In some cases, they were also asked to produce a mismatch between the gesture and the pitch accent, so that each fell on a different target word.

Firstly, Krahmer and Swerts confirmed that the realisation of a pitch accent resulted in significant higher f_0 and intensity, and longer segment duration. Additionally, they observed that prominent vowels accompanied by a beat gesture were produced with more prominent acoustic features than those without a gesture, regardless of the position of the pitch accent. Additionally, words accompanied by a gesture had longer segmental duration and lower second formant (F2), suggesting that pitch accents and visual beat have similar emphasizing functions⁴³.

In order to analyse whether these results were perceptually relevant, the researchers conducted two more experiments. In the first one, ‘hearing beats’, the audiovisual recordings obtained in the previous experiment were administered

⁴³ Alteration of F2 has been explained as a transfer of the social intention from the gesturing effort to a more marked tongue protrusion (Barbieri et al., 2009; Bernardis & Gentilucci, 2006).

to three trained listeners only in the auditory modality for prominence marking. Marking for prominence was conducted on a 3-point scale: 0 for no prominence, 1 for a minor pitch accent, and 2 for a clearly perceived pitch accent. Total perceived prominence was thus computed adding the given prominence score per word, so that a maximum score of 6 points was possible (corresponding to a clear pitch accent marked by all three trained listeners). The statistical analysis was conducted on the 'perceived prominence difference score', which resulted from subtracting the prominence score of the first target word to that of the second one. Additionally, agreement—computed as Pearson's correlation of the three listeners—ranged between 0.58 to 0.65 for the first word and 0.66 to 0.70 for the second word. In their results, Krahmer and Swerts observed that the production of a visual beat on a certain word had a clear impact on the prominence perceived in the auditory modality, so that the relative spoken prominence of that target word increased, while it decreased in the other one.

Finally, a third experiment, 'seeing beats', was conducted to assess the perceptual effect of visual beat gestures—in this third experiment, head nods were excluded from the experiment and only those gestures performed with hands and eyebrows were included. Three of the audiovisual recordings obtained in the first experiment were administered to participants in both audio-only and audiovisual modality. The experimental task consisted in marking prominence, first on one, then on the other target word, using a 10-point scale (1 for 'no prominence' and 10 for 'strong prominence'). Similar to the experiment just mentioned, researchers computed a 'visual difference score' for perceived prominence by subtracting the prominence score obtained in the audio-only modality from the score in the audiovisual modality: a positive result would suggest that the visual cues were responsible for an increase in perceived prominence.

Krahmer and Swerts concluded in this third experiment that seeing visual

beats actually had an effect on the perception of prominence. Besides, when participants saw a beat gesture on one word, the perceived prominence of the other word automatically decreased, and this effect was stronger in the case of the first word. As for gestures, manual beat gestures were observed to play a more important role than rapid eyebrow movements, and the perceived prominence also depended on the way each of the three speakers gestured (Figure 22).

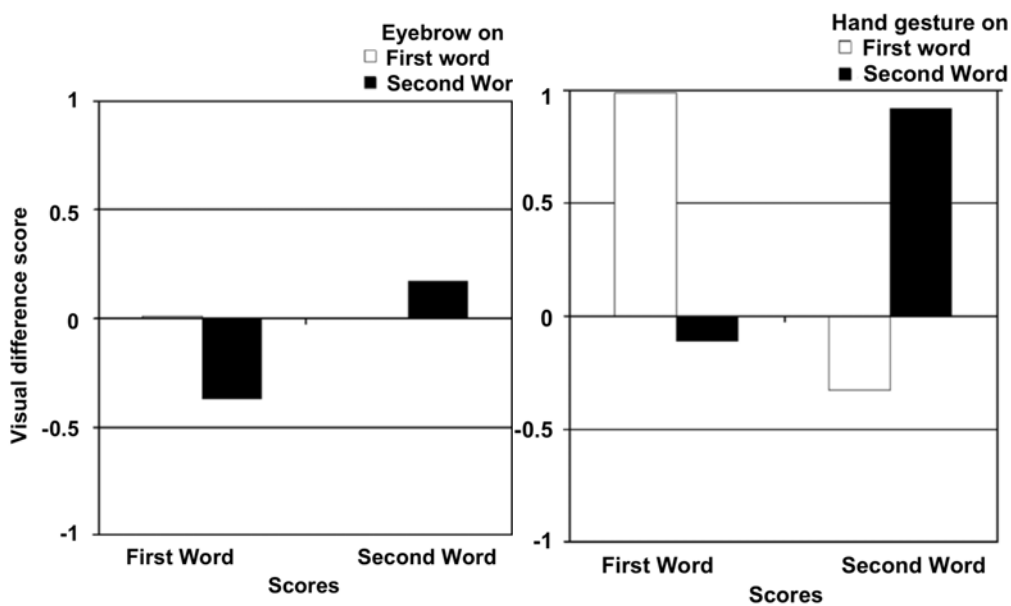


Figure 22: Average prominence scores obtained by Krahmer and Swerts (2007, experiment 3) for both target words after subtracting the score of the audio-only modality from that of the audiovisual modality. A positive score reflects an increase in prominence perception due to the occurrence of visual cues. Left chart shows hand beats; right chart shows eyebrow beats. Adapted from Krahmer & Swerts, 2007, p. 409.

A follow-up study conducted by the same researchers registered participants' reaction times to analyse both the effects of modality and face area in the audiovisual perception of prominence (Swerts & Krahmer, 2008). Six native speakers of Dutch (two of them the researchers themselves) were recorded uttering variants of a sentence containing three target words (underlined) *Maarten gaat maandag*

naar Mali ('Maarten goes Monday to Mali'), which corresponded to the different narrow focus versions: 'Who will go on Monday to Mali?'; 'When will Maarten go to Mali?'; 'Where will Maarten go on Monday?'; plus one prosodic neutral version corresponding to broad focus.

The stimuli were then manipulated to create audiovisual mismatches. Participants' reaction times were recorded for an experimental task consisting in recognising the most prominent target word in the sentence. In their results, the researchers observed that congruent auditory and visual cues of prominence, i.e. those occurring on the same words, were identified more rapidly than incongruent cues. Additionally, a second experiment in which the audiovisual information was partially blackened, revealed that the upper face had a stronger prominence cueing effect than the bottom part of the face, and so was the left part respect to the right part (Figure 23).

No mismatch, but degraded speech

Dohen and Lœvenbruck (2009), using a careful controlled experimental setting, tested the role of visual cues in the audiovisual perception of contrastive focus⁴⁴.

In their rationale, they explained that the auditory perception of contrastive focus in French usually shows a ceiling effect, i.e. close to 100% of successful identification. For this reason either audiovisual mismatches or acoustically degraded stimuli were necessary to assess the potential enhancement of prominence perception associated to the visual modality. Since audiovisual mismatches had been previously used by Swerts and Kraemer (2004; 2008), Dohen and Lœvenbruck used whispered speech stimuli. In this experimental paradigm, there is no inton-

⁴⁴ Their study extended the results of a previous experiment (Dohen & Lœvenbruck, 2005) with a slightly different methodology

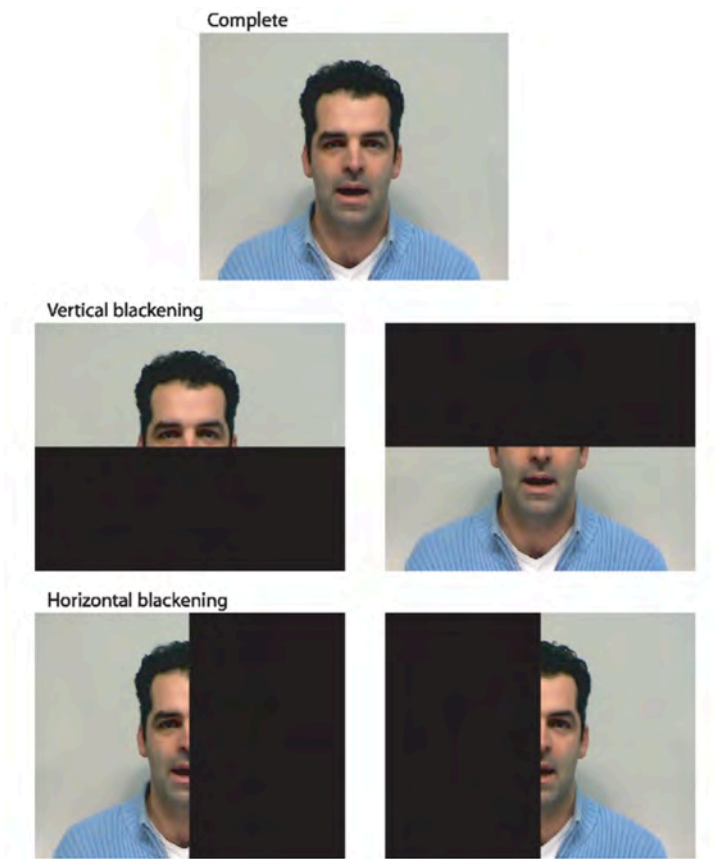


Figure 23: *Different versions of the stimuli presented in Swerts and Krahmer's study (2008, experiment 2) to analyse the effect of prominence cued by several face areas. Adapted from Swerts & Krahmer, 2008, p. 229.*

ational (f_0) information because there is no vibration of the vocal cords, although duration and intensity remain intact. In this way, it is more difficult to perceive prosodic focus auditorily.

Two French native speakers were recorded in a correction task, in which narrow focus sentences were elicited in whispered speech using four question-answer pairs. The resulting sentences were used as stimuli, which were administered to 13 listeners in all three conditions, i.e. audiovisual, audio-only, and visual-only; and two different angles, i.e. front and profile views of the speaker.

Listeners were instructed to identify the focused word.

In the results of this first experiment, Dohen and Lœvenbruck reported that detection of focus was significantly better in the audiovisual condition than in both audio-only and visual-only conditions, with no differences between the latter two. In their second experiment, similar stimuli were elicited as those in experiment one, with the exception that the narrow focus sentences were uttered in both normal speech and whispered speech and the speaker was recorded only from a frontal angle. Their stimuli, consisting this time in two sentences uttered in two speech modes in three focus conditions, were administered to 31 listeners in the same way as in their first experiment.

Dohen and Lœvenbruck observed that focus was more easily perceived in the normal speech condition as well as in the audiovisual modality, as it was in their first experiment. In addition, no differences were found in normal speech between modalities (audio-only vs. audiovisual), while in whispered speech correct answers were significantly higher in the audiovisual modality. By the same token, perception of focus was more difficult in whispered speech than in normal speech in the audio-only modality, but it was easier in the visual-only modality. Analysis of reaction times also showed that listeners processed speech faster in the audiovisual modality in whispered speech when compared to when they just heard speech (audio-only) or saw the speaker (visual-only). Such a difference in reaction times disappeared, however, in normal speech when the audiovisual modality was compared to the audio-only modality—although it persisted when the audiovisual modality was compared to the visual-only modality.

Do visual cues help detect f_0 changes or intensity changes?

Another study on the perception of audiovisual prosody conducted with controlled speech stimuli in an experimental setting was conducted by Foxtan et al.

(2010). In their case, the researchers investigated the role of visual information on the detection of threshold-level differences in f_0 and intensity. For the creation of the experimental stimuli, a female French actress was instructed to utter the two-word phrase *si chic* ('so smart') three times: emphasizing the first word, the second word, or none of them, paying attention to pronouncing each of them with the same duration. As for the visual cues, the joint movement of eyebrows and head were filmed as naturally co-occurring on the emphasised word.

The researchers first homogenised the acoustic parameters of f_0 and intensity in the 'no-emphasis' condition, and they also matched the duration of the vocalic segments. Subsequently, stimuli with f_0 and intensity differences were created, ranging stepwise from 0.1 to 24 semitones and from 0.2 to 25 dB, respectively. In a set of congruent stimuli, f_0 and intensity had the same values as in the original recording, while in a different set of incongruent stimuli they did not match (Figure 24).

The recordings were presented to participants in a forced-choice task both in the audiovisual modality and in the audio-only modality. Participants had to identify which of the stimuli contained an emphasised word (either by increased f_0 or increased intensity), although the audiovisual information was congruent only in half of the stimuli, a factor that was not present in the audio-only condition.

Foxton and her colleagues observed that participants were able to detect both f_0 changes and intensity changes in both modalities, although they reported that congruent visual cues of prominence helped participants to better detect the thresholds of auditory cues than when they could only hear the stimuli, and this especially so for intensity thresholds. This is in line with previous results suggesting that amplitude changes are more affected by visual prosodic information than pitch changes (Scarborough et al., 2009), since intensity is thought to

be more perceptible in articulatory gestures than f_0 changes.

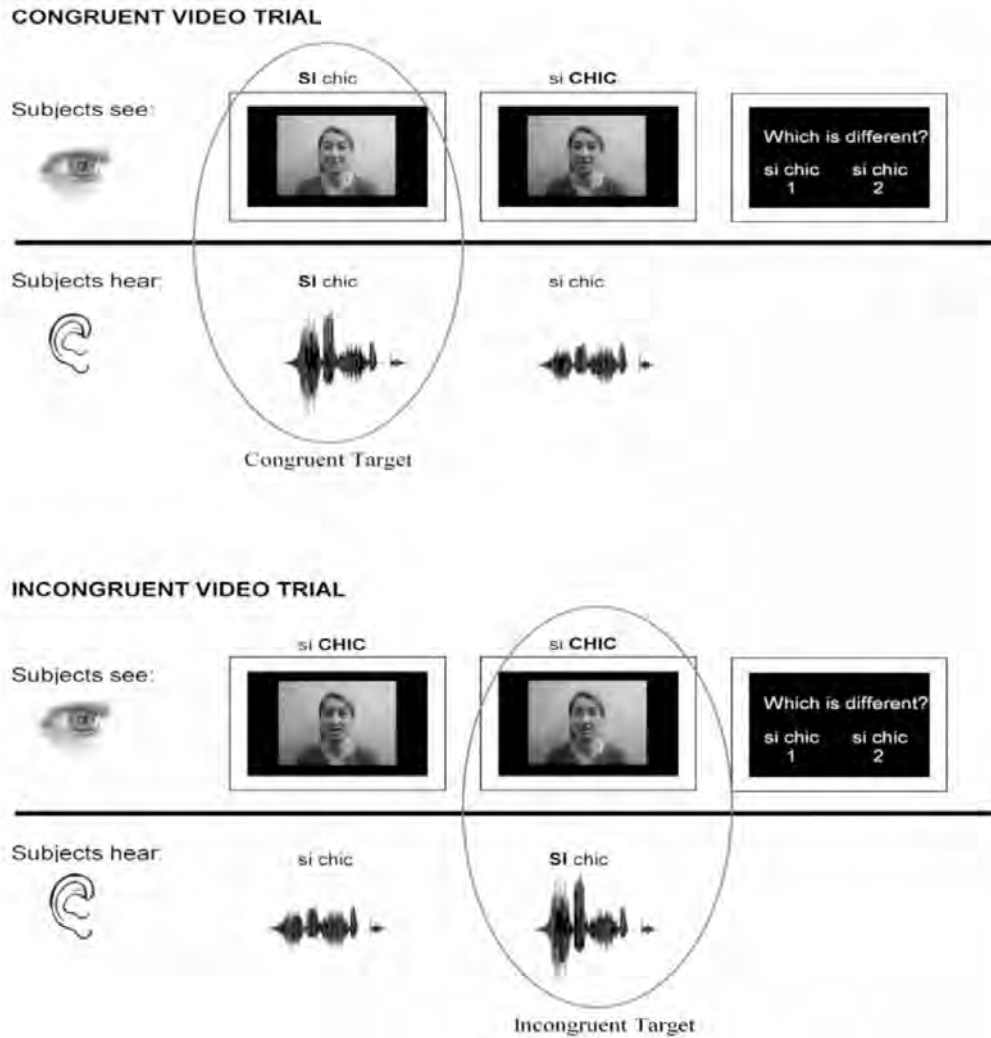


Figure 24: Example of two audiovisual stimuli belonging to the congruent and the incongruent set of stimuli, respectively, used for the task of prominence perception cued by intensity. The capitalised word on top of each video image corresponds to the realisation of the visual cues of prominence that participants saw in the video, while the oscillogram extract below the video image shows the acoustic cues that participants heard. Adapted from [Foxtan et al., 2010](#), p. 74.

2.3.8.4 Studies using spontaneous speech

Prominence perception using spontaneous speech stimuli

Swerts and Krahmer (Swerts & Krahmer, 2010) envisaged to analyse the effects of facial expressions on speech production using spontaneous stimuli obtained from TV newsreaders, after their previous studies on audiovisual prominence (see § 2.3.8.2 and § 2.3.8.3 for details). They had previously noticed the need to “supplement the current findings [Krahmer & Swerts, 2007] with data about gestures (both manual and facial) in spontaneous speech, although it is difficult to see how incongruent utterances could be triggered naturally” (2007, p. 411). Their call for spontaneous speech makes evident that, in most studies, stimuli often contain only one gesture, which is very far from the natural interweaving of gestures performed with hands, head, and eyebrows typically found in everyday spoken language.

For their analyses, the researchers selected 60 fragments from the recordings of two male and two female Dutch TV newsreaders in which utterances ranged between 4s and 12s. The stimuli were then presented to a group of 35 participants in the auditory modality for binary prominence marking (prominent vs. non-prominent). The responses were later used to create a prominence scale, based on the idea that prominence can be related to the proportion of subjects that agrees on a word being prominent, as done in previous studies (Mo et al., 2008; Streefkerk, 2002; Swerts, 1997). Thus, if a word did not received any mark of prominence, it was categorised as having no accent, while if it received at least half of the possible marks, it was considered as having a strong accent. Word falling in between these two extremes were classified as having a weak accent. Subsequently, two independent researchers visually annotated the stimuli for eyebrow and head movements.

Swerts and Krahmer's analysis of the distribution of gestures and pitch accents showed that the joint movements of eyebrows and head were used together with auditory cues to mark certain words as prominent. More precisely, words categorised as 'strong accent' mostly occurred with an accompanying eyebrow movement; however, the mere presence of an eyebrow movement did not imply the presence of a strong accent: only 47 out of 303 eyebrow movements corresponded to a strong accent. The distribution of head movements followed a similar pattern, so that 60 out of 228 head nods were associated to a strong accent. Interestingly, strong accents were especially marked by combinations of eyebrow and head movements (in 67.2% of the cases); conversely, single eyebrow or head movements hardly coincided with strong accents, while the joint movement of eyebrow and head occurred on 19 out of 138 words having a strong accent.

Distribution of gestures in spontaneous speech samples

Similarly, Ambrazaitis and House (2017) also used spontaneous speech from several TV newsreaders to explore the audiovisual realisation of multimodal prominence. Differently from Swerts and Krahmer (2010), the researchers did not conduct a task to perceptually detect prominence. In their study on the interaction between beats performed with head and/or eyebrows and pitch accents signalling focal constituents, the TV recordings were annotated for gestural and acoustic cues of prominence by two independent annotators plus the main author of the study (Table 7).

Inter-rater agreement, measured as Fleiss's kappa (1971), ranged between 0.69 and 0.77. Interestingly, occurrences of head movements not always coincided with a focal accent, differently from observations from previous studies (e.g. Swerts & Krahmer, 2010; Yasinnik et al., 2004), which the researchers relate

Prominence marker	Occurrences	% of words
FA only	128	12.98
FA+HB	126	12.78
FA+EB	3	0.30
FA+HB+EB	39	3.96
HB only	58	5.88
HB+EB	10	1.01
EB only	15	1.52
None	607	61.56
Total	986	100.00
Total		
FA (focal accent)	296	30.02
HB (head beat)	233	23.63
EB (eyebrow beat)	67	6.80

Table 7: *Distribution of the prominence markers FA (focal accent), HB (head beats), and EB (eyebrow beats). The upper part of the table shows the total distribution of markers including individual and combined occurrences. Adapted from Ambrazaitis & House, 2017, p. 105.*

to the purpose of their annotating only focal or ‘big’ [sic] accents. In the case of words annotated as containing only a head beat, a post-hoc annotation of pitch accents was conducted by two raters to determine whether the word in question carried a minor pitch accent or was completely deaccented. In this second round, inter-rater agreement ranged between 0.46 and 0.52, and many annotations previously considered as head beats unrelated to pitch accents were now linked to minor, non-focal pitch accents (labelled by the authors as ‘small’ accents).

Subsequently, Ambrazaitis and House carried out a functional analysis including a comparative assessment of the combined prominence markers. The researchers found patterns for the usage prominence markers, which focused on how they were used across different news topics, on repetitions of topics by the same speaker, and on their use by different speakers. Text and information structure were the two main analysis criteria together with speaker expressivity.

In their results, Ambrazaitis and House reported a different distribution of focal accents and head beats (but not eyebrow beats) within a news story uttered by the newsreader. In this case, focal accents were preferably used in the first half of the text, while head beats and the combination of focal accents and head beats occurred during the second half of the text. This was explained as a consequence of information structure, since the initial part of a text often presents the theme or define a common ground, while the second part usually corresponds to the rheme. Thus, head beats seemed to highlight the most important information in a piece of news once it had already been presented in the first half of the text.

2.3.8.5 Gender differences in the audiovisual perception of speech

Finally, behavioural and neuroanatomical differences between men and women have been reported for the audiovisual perception of speech (Dancer et al., 1994; Öhrström & Traunmüller, 2004; Ruytjens et al., 2006, 2007; Watson et al., 1996; see Alm & Behne, 2015, for a summary). More precisely, women have been observed to perform better at speech-reading than males (e.g. Dancer et al., 1994; Strelnikov et al., 2009), which has been related to the fact that women could be more active gazers than men (e.g. Johnson et al., 1988). Additionally, women have been reported to be more sensitive to visual cues than men in audiovisual speech perception (Aloufy et al., 1996; Irwin et al., 2006; Öhrström & Traunmüller, 2004). For example, Öhrström and Traunmüller (2004) showed that women were significantly more influenced by the visual modality than men in perceiving incongruent Swedish vowels embedded in a syllable. Similar results were reported by Irwin et al. (2006), who studied the influence of visual speech for the syllable /ba/. Irwin et al. suggested that such a gender difference might be due to a different pattern in language processing, with a stronger activation in bilateral brain areas causing a more efficient audiovisual language processing in women (e.g. Baynes

et al., 1994; Coney, 2002).

Furthermore, neuroanatomical differences point to a stronger activation of brain areas associated with speech perception in women (Ruytjens et al., 2006, 2007). In addition, it has been claimed that gender differences in audiovisual speech perception may emerge in the context of challenging stimuli (Jaeger et al., 1998), which can be related to the results observed in the third experimental condition Exp3, where duration was the only acoustic cue available to participants.

2.3.9 Summary

The Kendon's continuum classifies gestures according to the relationship that body movements have to speech and their increasing degree of conventionalisation. In their relation to speech, body movements at the left end of the continuum are produced unwarily together with speech and are known as 'gesticulation'. Conversely, at the right end, the movements performed in sign languages are produced consciously to communicate typically in the absence of speech. Gesticulation does not have agreed-upon movements conveying meaning or any phonological, morphological, and syntactic system to combine its constituents. Therefore, these type of gestures are not conventionalised, while sign languages at the right end share a conventionally structured code necessary for communication. Similarly, gestures at the left end convey meaning in a global way, i.e. they can not be segmented; while at the right end, sign languages convey meaning by combining smaller parts, as morphemes do when they are combined into larger meanings in spoken language.

Gesture studies

Research on gestures experienced a renewed interest since the mid-20th century, partly due to the convergence of linguistic, psychological, and psychiatric

studies. Nevertheless, different aspects of gestures can be traced back to the Antiquity. One of the main questions arisen at different moments in history is the role played by gestures in the cognitive foundations of language. For example, some authors, such as Giambattista Vico (1725/1744) and Étienne Bonnot de Condillac (1746), already considered gestures as a fundamental aspect of human communication underlying the evolution of the language faculty. Such an idea has been later reformulated (e.g. Armstrong et al., 1995; Corballis, 2002; Stokoe, 2001), partly buttressed by neurophysiological evidence confirming the sharing of neural circuits between speech and gesture (e.g. Meister et al., 2003).

Also, a theoretical framework to account for their cognitive integration has been put forward by McNeill (1992), although similar accounts have also been offered by other authors, e.g. Kita (2000) and De Ruiter (2000). The particularities of each of them turn around the stage at which the interaction of gesture and speech takes place, although their interaction is not questioned for language production and language perception (Biau & Soto-Faraco, 2013; Biau et al., 2015). Subsequently, the temporal coordination of both phenomena has been argued to reflect their close connection. Quite a large number of studies have tried to pin down the precise time-alignment of gesture and speech and its implications for a wider perspective on language (e.g. Esteve-Gibert & Prieto, 2013; Karpiński et al., 2009; Krivokapić et al., 2015, 2016; Loehr, 2004; Morrel-Samuels & Krauss, 1992; Rochet-Capellan et al., 2008).

Furthermore, different criteria have been used in an attempt to order and classify gestures. Important researchers on the phenomenon of gesture have contributed in the 20th century with several classifications (Ekman & Friesen, 1969; Ekman, 1999; Efron, 1941/1972; Freedman & Hoffman, 1967; Kendon, 1972; McNeill, 1992). One of the most widespread distinction among authors is based on semiotic principles, and gestures are divided into those which refer to an

object by pointing at it and those which characterise the object in some way. Nowadays, the most widely accepted classification of gestures is that developed by McNeill (1992).

In his classification, McNeill established a fundamental division between *imagistic* gestures and *non-imagistic* gestures. The former are those that depict the shape of an object, display an action or represent some pattern of movement. Depending on whether they are concrete or abstract, they can be grouped into: *iconic* and *metaphoric*. *Non-imagistic* gestures include pointing gestures, rhythmic gestures that highlight either segments of the discourse (*cohesives*) or its rhythmic structure (*beats*), and gestures associated to speech failures (*butterworths*).

Some have also explored the interaction between gestures and verbal prosody. It has been observed that gestures have much in common with prosody in their potential, for example, for adding non-discrete nuances that serve interactive functions and facilitate comprehension (Foxton et al., 2010; Munhall et al., 2004). Furthermore, gestures have been also found to possess similar prominence-increasing effects to those of speech, with prominence marking as one of the many possible interactions between both modalities (Krahmer & Swerts, 2007).

Temporal coordination of gesture and speech

The temporal coordination of gesture and speech has also been studied, and knowledge on their precise synchronisation has come a long way in the last decades. The ‘phonological synchrony rule’, as stated by McNeill (1992), has been contested and put to the test (e.g. McClave, 1991, 1994; Karpiński et al., 2009), although several studies have proved the tight time-alignment between gesture and speech (Esteve-Gibert & Prieto, 2013; Krivokapić et al., 2015, 2016; Leonard & Cummins, 2010; Loehr, 2004).

Audiovisual prosody

On the other hand, research on the visual component of communication has begun to give an account of how the visual correlates of prominence (e.g. gestures performed with hands, eyebrows, or head) interact with verbal prosody (e.g. Al Moubayed et al., 2010; Granström et al., 1999; Kim et al., 2014; Krahmer & Swerts, 2007; Prieto et al., 2011; Scarborough et al., 2009). In the case of prominence perception, visual cues result in stronger production and perception of verbal prominence (Krahmer & Swerts, 2007); in this way, facial gesturing, for example, has been found to systematically influence the perception of verbal prominence (Dohen & Løevenbruck, 2009; House et al., 2001; Swerts & Krahmer, 2008).

Most studies having addressed the interaction of visual and verbal prominence have so far made use of both lip-synchronised animated agents (e.g. Al Moubayed & Beskow, 2009; Granström et al., 1999; House et al., 2001; Krahmer et al., 2002a,b; Prieto et al., 2011) and experimental settings in which gestures are elicited with controlled speech stimuli (e.g. Dohen & Løevenbruck, 2009; Foxton et al., 2010; Krahmer & Swerts, 2007). In both cases, the visual cues of prominence—limited to beats produced by eyebrow raises and head nods, and occasionally also by hand gestures (Krahmer & Swerts, 2007)—have been observed to enhance verbal prominence perception.

Finally, a difference exists between men and women in the audiovisual perception of speech, which has been supported by neuroanatomical differences, with a stronger activation in bilateral brain areas causing a more efficient audiovisual language processing in women (Dancer et al., 1994; Öhrström & Traunmüller, 2004; Ruytjens et al., 2006, 2007; Watson et al., 1996).

Methodology

This chapter firstly presents a review of the methods found in previous studies on prominence perception. Then, the statistical background applied in the experimental analysis conducted in this study is discussed. After these two sections, the details of the methodology applied in the experimental part of this study is discussed and lays out the theoretical grounds on which it is based (§ 3.3.1). An account is given of the spontaneous speech materials used in both experiments as well as of the process by which the stimuli were created. Following this, some aspects of the methodology that are common to both experiments are referred, e.g. participants, gesture annotation of stimuli, and data analysis. Logically, the specific details differing between experiments, such as experiment design, procedure, etc., will be discussed particularly in their respective sections.

3.1 Previous methodological approaches

The review of the methodology is summarised here below in Table 8. This table shows the details reported in studies on prominence perception published between 1992 and 2017. For example, the number of participants in the over twenty experiments reviewed ranged between 3 and 74, and they mostly included naïve listeners, but also, to a lesser extent, trained listeners.

Author	N – Listeners	Task	Trials	Scale	N – Stimuli	Presentation	Modality	Speech material
Streefkerk et al., 1997 (experiment 1)	8 – naïve	Mark every prominent syllable in the sentence	3	0–1	81 sentences	Audio and printed sentence	A	Polyphone Corpus (telephone speech)
Streefkerk et al., 1997 (experiment 2)	8 – naïve	Mark every prominent word in the sentence	3	0–1	81 sentences	Audio and printed sentence	A	Polyphone Corpus (telephone speech)
Streefkerk et al., 1998	10 – naïve	Mark every prominent word in the sentence	1	0–1	500 sentences	Audio and printed sentence	A	Polyphone Corpus (telephone speech)
Granström et al., 1999	21 – naïve	Two-alternative forced choice	3	0–1	20 sentences 2 target words in a six-word sentence	Animated agent	AV	Synthesised speech
House et al., 2001	33 – naïve	Two-alternative forced choice	<i>ad lib.</i>	0–1	12 sentences 2 target words in a eight-word sentence	Animated agent	AV	Synthesised speech
Eriksson et al., 2001	18 – naïve	Mark every prominent syllable in the sentence	<i>ad lib.</i>	0–100	13 syllables in a nine-word sentence	Audio and printed sentence	A	Elicited speech
Krahmer et al., 2002a,b	25 – naïve	Three-alternative forced choice	<i>ad lib.</i> (2002a) 2 (2002b)	0–1	36 sentences 1 two-word sentence	Animated agent	AV	Elicited speech
Swerts & Krahmer, 2004	55 – naïve	Mark the most prominent syllable in the word	3	0–1	20 words 1 three-syllable pseudo-word	Audio and printed sentence Speaker on screen	A-V-AV	Elicited speech
Kochanski et al., 2005	2 – trained (marking from corpus) 1 naïve – 1 trained	Mark every prominent syllable in the sentence	<i>n.a.</i>	0–1	13437 syllables (sentence number is not given)	Sentence waveform Audio and printed sentence	A	IVIE Corpus (spontaneous speech)
Krahmer & Swerts, 2007 (experiment 2)	3 – trained	Mark every prominent word in the sentence	<i>ad lib.</i>	0–2	360 sentences 2 target words in a four-word sentence	Sentence waveform Audio and printed sentence	A	Elicited speech
Krahmer & Swerts, 2007 (experiment 3)	20 – naïve	Mark the target word in the sentence	1	0–10	72 sentences 2 target words in a four-word sentence	Audio Speaker on screen	A-AV	Elicited speech
Swerts & Krahmer, 2008 (experiments 1 and 2)	42 – 66 – naïve	Three-alternative forced choice	1	0–1	54 sentences 3 target words in a five-word sentence	Speaker on screen	AV	Elicited speech
Mo, 2008a	74 – naïve	Mark every prominent word in the sentence	2	0–1	18 excerpts of about 20 seconds long	Audio and printed sentence	A	Buckeye Corpus (spontaneous speech)
Cole et al., 2014	15 – naïve	Mark every prominent word in the sentence	<i>n.a.</i>	0–1	16 excerpts of about 18 seconds long	Audio and printed sentence	A	Buckeye Corpus (spontaneous speech)
Al Moubayed et al., 2010 (labelling)	1 – trained	Mark every prominent word in the sentence	<i>n.a.</i>	0–3	30 sentences ranging between 6 and 10 words	Printed sentence Animated agent	AV	Corpus (no further details provided)
Dohren & Levenbruck, 2009 (experiment 1)	13 – naïve	Four-alternative forced choice	1	0–1	192 sentences 4 sentences ranging between 4 and 5 words	Audio and printed sentence Speaker on screen	A-V-AV	Elicited (whispered) speech
Dohren & Levenbruck, 2009 (experiment 2)	31 – naïve	Four-alternative forced choice	1	0–1	192 sentences 2 sentences ranging between 4 and 5 words	Audio and printed sentence Speaker on screen	A-V-AV	Elicited (whispered) speech
Swerts & Krahmer, 2010 (experiment 1: auditory mark-up)	35 – naïve	Mark every prominent word in the sentence	<i>ad lib.</i>	0–1	60 sentences ranging between 4 and 12 seconds	Printed sentence	A	Audiovisual recordings from TV newsreader (spontaneous speech)
Ambrazaitis & House, 2017	3 – <i>n.a.</i>	Mark focal accents and head and eyebrow beats	<i>n.a.</i>	0–1	31 news items ranging between 1 and 3 sentences	Sentence waveform Speaker on screen	A-AV	Audiovisual recordings from TV newsreader (spontaneous speech)

Table 8: Methodological details of previous studies. Swerts and Krahmer’s study (2004) is the only one exploring lexical stress. Abbreviations: *n.a.* (not available) *ad lib.* (ad libitum, ‘no limit’).

The type of experimental task the participants were expected to conduct and the number of trials they were allowed to receive each stimulus are also detailed. Finally, the table also offers the scale to mark prominence used by participants, the presentation and modality through which stimuli were administered, as well as the speech material from which the stimuli were created.

3.1.1 Procedures

Previous studies on prominence perception have conducted prominence marking in different ways. As seen in the literature, perceptual studies mostly involve experimental tasks done by either naïve listeners (Mo et al., 2008) or by trained listeners (usually fewer than naïve listeners) (e.g. Al Moubayed et al., 2010; Krahmer & Swerts, 2007, experiment 2). In order to conduct the experimental task participants are generally given minimal instructions to mark:

1. Every syllable in the sentence (e.g. Eriksson et al., 2001; Kochanski et al., 2005; Streefkerk et al., 1997).
2. Only the most prominent syllable in the target word (e.g. Swerts & Krahmer, 2004).
3. The target word in the sentence (Krahmer & Swerts, 2007, experiment 3).
4. Every prominent word (e.g. Cole et al., 2014; Krahmer & Swerts, 2007, experiment 2; Mo, 2008a).
5. Only the most prominent word (e.g. Dohen & Løevenbruck, 2009; Swerts & Krahmer, 2008).

Thus, whether the perceptual target is the syllable or the word, the experimental task makes possible to mark either all prominent elements in their respective environments or just limit the marking to just one target element. For

this, it is possible to use a binary scale, e.g. prominent vs. non-prominent (e.g. Cole et al., 2014), or a gradient scale, e.g. a Likert scale, a 4-point scale, a 10-point scale, etc. (Krahmer & Swerts, 2007, experiment 3); however binary prominence marking has also been conducted using 2-, 3- and 4-alternative forced-choice tasks (e.g. Granström et al., 1999; Dohen & Løevenbruck, 2009; Krahmer et al., 2002a,b) (Table 8).

The choice between target syllables and words, on the one hand, as well as between a binary scale and a gradient scale for marking, on the other hand, depends on several factors: the nature and length of the stimuli, the cognitive effort required from the listeners, and the fine-grained detail of the expected results. Marking words for binary prominence in short sentences, for example, requires less cognitive effort from listeners than marking every syllable on a gradient scale in longer sentences, and it also results in a more consistent agreement on the prominent elements. In this sense, Streefkert et al. (1997), for example, compared their results for both syllables and words using the same speech material. She and her colleagues observed remarkable differences in agreement among listeners, with words being more meaningful units than syllables in an experimental task consisting in marking phrasal stress by naïve listeners. Additionally, Rosenberg and Hirschberg (2009) reached similar conclusions for the automatic identification of pitch accents.

It is also possible either to choose certain target syllables or words within a sentence or to make all elements available for marking in the sentence. For example, using short sentences, House et al. (2001) and Krahmer and Swerts (2007) presented two target words to listeners for marking. However, Streefkert et al. (1997) used read-aloud sentences from a corpus and, in two different experiments made firstly all syllables and then all words available for marking. Conversely, Dohen and Løevenbruck (2009) applied a four-alternative forced-choice task in a

series of five-word sentences for listeners to detect contrastive focus. Other studies employed a gradient scale to determine prominence, mostly on small samples of speech material (Eriksson et al., 2001). In their turn, Al Moubayed et al. (2010), for example used a 4-point scale in an experiment on the multimodal perception of prominence using an animated agent (Table 8).

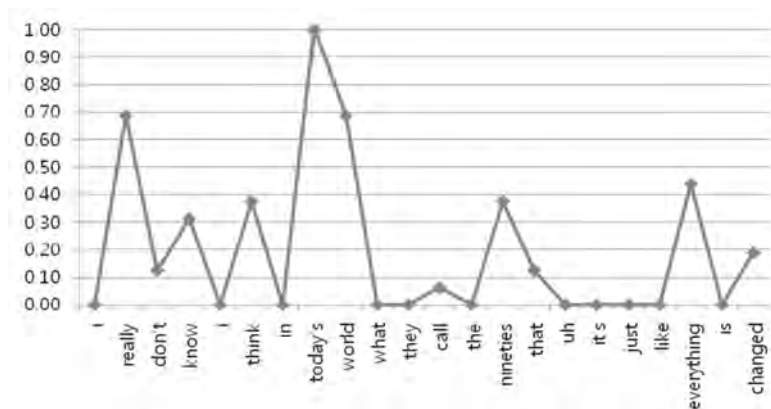


Figure 25: Example of P-scores resulting from pooling together the prominence marks given by a large number of listeners, which is then transformed into a probabilistic score. Adapted from Mo, 2008a, p. 260.

From this analysis of the literature, it can be observed that, despite the higher accuracy achieved by nuanced gradient scales, the use of a binary scale is more straight-forward to naïve listeners, thus reducing greatly the complexity of the experimental task. Additionally, binary prominence marking presents the advantage that it can be easily made into a more fine-grained scale by adding up the individual scores obtained from listeners. Cumulative marks of prominence are an elegant way to analyse prominence in more detail (Cole et al., 2010; Krahmer & Swerts, 2007; Mo, 2008a; Swerts, 1997). Thus, the proportion of listeners agreeing on prominent elements can be computed through a probabilistic P(prominence)-score that reflects the probability of a given word to be prosodically perceived as prominent (Figure 25).

3.1.2 Speech materials and type of stimuli

Not only the setup of an optimal experimental task, but also the experimental paradigm to be used represents a great challenge for the study of prominence perception, especially in its relation to the visual cues of prominence. Undoubtedly, methodological difficulties lie at the heart of all attempts to account for the interaction of visual cues and acoustic cues in prominence perception. So far, most methods applied to study the interaction of auditory and visual cues of prominence have made use of both animated agents (e.g. Al Moubayed et al., 2010; Granström et al., 1999; House et al., 2001; Krahmer et al., 2002a,b; Prieto et al., 2015) and elicited gestures with controlled speech stimuli in experimental settings (e.g. Dohen & Lœvenbruck, 2009; Krahmer & Swerts, 2007), as previously explained (§ 2.3.8.2 and § 2.3.8.3).

Despite their proved usefulness, both animated agents and elicited gestures and speech in the laboratory present some shortcomings for the study of the multimodal perception of prominence. Certainly, the use of animated agents has been shown to provide intelligibility (e.g. Agelfors et al., 1998), and they make possible to manipulate visual cues while preserving acoustic information. However, they have been limited mainly to the study of eyebrow and head movements (Krahmer et al., 2002a,b; Prieto et al., 2011), so that no gestures performed with hands, for instance, have been studied yet using animated agents, let alone a combination of manual, facial, and head gestures.

Additionally, stimuli created in experimental settings usually involve audio-visual recordings, in which it is difficult to elicit spontaneous gestures (e.g. Krahmer & Swerts, 2007). Although face expressions and head nods have been consistently analysed, manual gestures have been more systematically neglected. Speakers do not always behave as naturally when asked to in front of a camera

as when they gesticulate in natural conversations. Hence, the ecological validity of methods used in experimental settings is somewhat reduced, and the generalisation of results obtained with animated agents is to some degree limited.

It is also worth noting that the speech materials used are sometimes characterised by being carefully produced (e.g. [Dohen & Loevenbruck, 2009](#); [Krahmer & Swerts, 2007](#); [Krahmer et al., 2002a,b](#)), thus lacking some of the acoustic phenomena of spontaneous speech ([Face, 2003](#); [Laan, 1997](#)). For example, Krahmer and Swerts (2007) created their audiovisual stimuli by instructing participants to utter a four-word sentence and produce subsequently a quick visual beat with either hand, head, or eyebrows on one of the two target words of the sentence. Participants were also allowed to rehearse until they were not able to improve the combination of gesture and speech. Then the researchers administered these stimuli to a group of naïve listeners for prominence marking. Similarly, Foxton and her colleagues (2010) instructed a French actress to pronounce a two-word utterance to study the perception of co-occurring facial gestures.

Differently, several studies have employed spontaneous speech in the study of audiovisual prosody, and stimuli have been created either from speech corpora applied to animated agents ([Al Moubayed et al., 2010](#)) or from audiovisual recordings ([Ambrazaitis & House, 2017](#); [Flecha-García, 2006, 2007](#); [Loehr, 2004](#); [Swerts & Krahmer, 2010](#); [Yasinnik et al., 2004](#)). For example, Loehr (2004) filmed four subjects in natural conversation with friends for the study of the temporal coordination of visual and auditory prosody; Yasinnik (2004) filmed three academic lecturers in short monologues also with a similar purpose; and Flecha-García (2006) elicited task-oriented dialogues with which she analysed the participants' eyebrow movements in relation to discourse structure and utterance function. In their turn, Ambrazaitis and House (2017) and Swerts and Krahmer (2010) resorted to TV recordings of newsreaders, and both analysed eyebrow movements and

head movements, although only Swerts and Krahmer (2010, experiment 1: ‘auditory mark-up’) explored prominence perception, and this, only in the auditory modality.

From this it can be gleaned that previous methods present some important shortcomings. Similarly, it is also apparent that spontaneous speech has not been fully exploited. Therefore, the creation of stimuli from spontaneous speech samples can both extend the research questions to be addressed and increase the ecological validity of results obtained in the multimodal perception of prominence.

3.1.3 Summary

Studies on prominence perception have set up experimental procedures in several ways, differing in the perceptual target unit: whether syllable or word (e.g. Streefkerk et al., 1997); the nature of the marking: binary scale (e.g. Swerts & Krahmer, 2004), binary scale in a forced-choice task (e.g. Granström et al., 1999), a 4-point scale (e.g. Al Moubayed et al., 2010), or a 10-point scale (e.g. Krahmer & Swerts, 2007, experiment 3); and the presentation of stimuli: audio-only (e.g. Streefkerk, 2002), visual-only, (e.g. Swerts & Krahmer, 2004), and audiovisual (e.g. Dohen & Løevenbruck, 2009).

Next to this, some important differences among studies are seen in the participants conducting the task: naïve (e.g. Mo, 2008a) or trained (e.g. Krahmer & Swerts, 2007, experiment 2), and the speech material stimuli are created from: elicited speech (e.g. Dohen & Løevenbruck, 2009), synthesised speech (e.g. House et al., 2001), or corpus speech (e.g. Mo, 2008a). Only a small fraction of studies have used spontaneous speech, but none to study the multimodal perception of prominence.

From the literature, it seems that, despite the higher accuracy achieved by

nuanced gradient scales, the use of a binary scale is more straight-forward to naïve listeners. Binary prominence marking also presents the advantage that it can be easily made into a more fine-grained scale by adding up the individual scores obtained from listeners (e.g. Cole et al., 2010; Mo, 2008a).

Furthermore, shortcomings are associated to the methodological paradigms used in the study of the multimodal perception of prominence. For example, studies conducted using animated agents have been limited mainly to the study of eyebrow and head movements (e.g. Krahmer et al., 2002a,b; Prieto et al., 2011). Therefore, gestures performed with hands, for instance, have not been studied using animated agents, let alone a combination of manual, facial, and head gestures. On the other hand, it is usually difficult to elicit spontaneous gestures in experimental settings (e.g. Krahmer & Swerts, 2007), and manual gestures have also often been omitted from analysis. As a result, the ecological validity and the generalisation of results are to some degree limited.

Although several corpora of spontaneous speech have been used in the study of prominence perception—only in the audio-only modality (e.g. Mo, 2008a)—, only few studies have used spontaneous speech in the audiovisual modality (e.g. Flecha-García, 2006, 2007; Loehr, 2004; Yasinnik et al., 2004) and still fewer have done so in the study of the multimodal perception of prominence (Ambrazaitis & House, 2017; Swerts & Krahmer, 2010).

3.2 Review of statistical methods

In this section, first, the details of linear mixed models (LMMs) are discussed. LMMs are then extended and generalised linear mixed models (GLMMs) handling non-normal distributions are presented, with special emphasis on the binomial distribution. The different estimation methods for GLMMs are explained

together with the model selection procedure that permits to establish the model, from a set of estimated models, that best accounts for the data. Finally, an introduction to the *Akaike information criterion* (*AIC*), as a method to compare models, is offered. *AIC*, differently from traditional *p*-values, simply provides an ordinal value assessing the quality of models respect to their complexity, thus lacking any meaning of its own except as a way to rank models form a set (Akaike, 1973).

3.2.1 From ANOVAs towards Linear Mixed Models (LMMs)

Statistical analyses in the field of psycholinguistics have traditionally applied analyses of variance (ANOVAs) often involving full factorial designs with repeated measures. Such experimental designs aimed at increasing statistical power and precision. Typically a particular sample of participants gave responses to a fixed set of stimuli. By rotating stimuli over experimental conditions undesired repetition effects were avoided.

In such designs, the collection of more than one response per participant necessarily breaks the assumption of independence of observations—more precisely, independence of residuals. This principle is inherent to any linear regression model, since more than one response from a given participant is usually correlated and, therefore, responses are not independent. Logically, several observations from the same participant are tend to be more similar to each other than two observations obtained from different participants. For example, in an experiment measuring reaction times for word recognition under three different experimental conditions, e.g. words, pseudo-words, and words from a known second-language. In the case that each condition included 15 stimuli, each participant would contribute with 45 different observations that are non-independent.

One way to handle such correlated data is by explicitly modelling this de-

pendency. For example, it is possible to explicitly declare in the model that some responses were given by the same individuals. *Linear mixed-effects models* (henceforth LMMs) (e.g. Baayen et al., 2008), also known simply as ‘mixed models’, include these correlated data into the model via stating a random effect for participants (typically known as ‘by-subject’ random effects). This is also true of experiments in which some variables are clustered or nested within other variables; that is, when units of observations are clustered in groups such as students grouped within schools, or patients within hospitals. In such situations, it is also logical to expect observations within each group to be more similar to each other than observations across groups.

Following the example mentioned before, if taking words as the unit of observation¹, words in the first experimental condition (native words) are expected to be recognised much faster than those of the second (pseudo-words) and third experimental conditions (second-language words). Therefore, observations for those stimuli belonging to the same experimental condition are expected to be more similar to each other. As a result, once more, the assumption of independence is not met. In this case, this between-group variability is controlled for by means of a second random effect (known as ‘by-item’ random effects), which corresponds to stimuli nested in three different experimental conditions in our hypothetical example: native words, pseudo-words, and second-language words (Table 9).

It was noted long ago that the sample of linguistic items from which generalisation to the larger population of linguistic material is made also possesses an inherent variability that breaks the assumption of independence (Clark, 1973)—even if stimuli are not necessarily grouped within different experimental condi-

¹ Differently from many psychology experiments, in which the unit of observation is the participant.

Reaction Times	Participant	Condition	Item
0.108	1	Word	1
0.456	1	Pseudo-word	1
0.353	1	Second Language-word	1
0.110	1	Word	2
0.432	1	Pseudo-word	2
0.310	1	Second Language-word	2
...
0.112	1	Word	15
0.390	1	Pseudo-word	15
0.311	1	Second Language-word	15
0.187	2	Word	1
0.403	2	Pseudo-word	1
0.315	2	Second Language-word	1
0.120	2	Word	2
0.412	2	Pseudo-word	2
0.229	2	Second Language-word	2
...
0.115	2	Word	15
0.385	2	Pseudo-word	15
0.210	2	Second Language-word	15
...
0.150	20	Word	15
0.350	20	Pseudo-word	15
0.268	20	Second Language-word	15

Table 9: Fictitious data from a study on word recognition. Responses are non-independent, i.e. each participant provides 3 responses per item over three conditions; and every item in each condition receives 20 responses, one per participant, adding up to a total of 900 responses (15 items x 3 conditions x 20 participants). The respective by-subject and by-item random effects capture these two sources of variability.

tions, as in the provided example. As a solution, Clarke proposed to account, in two different analyses, for item variability (F2) across subjects (F1) by computing the quasi- F (F') and min- F' statistics (see [Quené & van den Bergh, 2008](#), for

a review).

In the case of LMMs, which are a generalisation of ordinary regression models, such by-subject and by-item dependencies of observations can be explicitly declared as random effects². Thus, mixed models (also known as ‘hierarchical regression’, ‘multi-level regression’, or ‘variance component model’), include three elements: fixed effects, random effects, and a last term for unaccounted variability, which contributes linearly to the dependent variable. An ordinary regression model can be specified by the following expression:

$$\begin{aligned} y_{i,j,k} &= \beta_0 + \beta_1 X_{i,j,k} + \epsilon_{i,j,k}, \\ \epsilon_{i,j,k} &\sim \mathcal{N}(0, \sigma^2). \end{aligned} \tag{1}$$

In this model (1), $y_{i,j,k}$ is the outcome to be predicted by the model, and the $X_{i,j,k}$ is the *predictor* corresponding to the observation made by a subject i for an item j in a certain condition k . Predictions are summarised by a straight line, which is defined by two parameters known as *regression coefficients*: the *intercept*, β_0 , the point at which the line crosses the Y axis; and the *slope*, β_1 , the gradient of the straight line fitted to the data. Any model includes an observation-level error $\epsilon_{i,j,k}$, which is normally distributed with mean 0 and variance σ^2 .

However, model (1) is not a mixed-effects model because it does not capture the violation of independence that has just been mentioned and that corresponds to the several observations made by each subject. This can be accounted for by including a new term u_{0i} corresponding to the deviation from β_0 for each subject i .

² In this section I mainly follow the detailed expositions made by Barr et al., 2013 and Singmann & Kellen, in press.

$$\begin{aligned}
 y_{i,j,k} &= (\beta_0 + u_{0i}) + \beta_1 X_{i,j,k} + \epsilon_{i,j,k}, \\
 \epsilon_{i,j,k} &\sim \mathcal{N}(0, \sigma^2), \\
 u_{0i} &\sim \mathcal{N}(0, \tau^2).
 \end{aligned}
 \tag{2}$$

In this case, the parameters β_0 and β_1 in model (2) are fixed effects, i.e. terms of interest to the researcher, which reveal the effects of the independent variable(s) on the population-level average, either as main effects or as interactions. Conversely, u_{0i} is a random effect and reflects the stochastic variability of a categorical variable. In this case, the random factor ‘subject’ corresponds to the random sample extracted from a population in order to eventually generalise over it. This implies that the sample would certainly have a different composition if the same experiment were carried out another time. Due to the fact that this random effect estimates the population distribution from which the u_{0i} effects were drawn, it is assumed that it follows a normal distribution with mean 0 and variance τ^2 .

If we take our previous example, we could see differences among participants in their reaction times for word recognition, some being slower and some being faster than the average. Nevertheless, such differences are not reflected in model (1), which assumes a single intercept β_0 for all of them. Model (2), on the contrary, allows for idiosyncratic averages per participant by introducing the term u_{0i} . This term captures the displacement of each participant from the *grand mean*, i.e. the intercept β_0 . For this reason, this model is known as a *random-intercept model*, and ‘subject’ is the grouping factor for which random intercepts are estimated.

Furthermore, although model (2) assumes that differences in reaction times can be due to variability between subjects, it does not include a second source of variability, i.e. the difficulty inherent to each condition. For example, even if

all subjects were beginner learners of Dutch as a second language, recognition of second-language words in Dutch might be more challenging for some participants depending on their second-language competence. In such a case, a new term capturing this variability within participants for the condition effect of β_1 , word type, is introduced into the model:

$$\begin{aligned} y_{i,j,k} &= (\beta_0 + u_{0i}) + (\beta_1 + u_{1i})X_{i,j,k} + \epsilon_{i,j,k}, \\ \epsilon_{i,j,k} &\sim \mathcal{N}(0, \sigma^2). \end{aligned} \tag{3}$$

As previously seen in model (2), u_{0i} represents the displacement of the i^{th} participant from β_0 , while now, in model (3), u_{1i} corresponds to the displacement of that same participant i from the mean associated to the effect of the experimental condition represented by β_1 . As repeated observations from the same participants are correlated, the several observations made for the same experimental condition are correlated. The two random effects u_{0i} and u_{1i} are assumed to have a mean 0 and a variance-covariance matrix Σ that lists both the variances (on the diagonal) and the covariances between conditions (off the diagonal):

$$(u_{0i}, u_{1i}) \sim \mathcal{N}\left(0, \Sigma = \begin{bmatrix} \tau_0^2 & \rho_{u_0, u_1} \\ \rho_{u_0, u_1} & \tau_1^2 \end{bmatrix}\right) \tag{4}$$

The resulting variance-covariance matrix in (4) reflects the dependencies that emerge when both random effects by-subject are correlated. For example, some participants with lower reaction times in word recognition might also have a better command of Dutch as a second language and, thus, show less differences, for example, between the ‘word’ and ‘second-language word’ conditions. This variability is reflected in regression parameters varying in β_1 for each participant; that is, varying in slope. In other words, in model (3) participants have different

random slopes next to different random intercepts. In Equation (4), τ_0^2 corresponds to random intercept variance, τ_1^2 , to random slope variance, and ρ_{u_0, u_1} , to the intercept-slope covariance.

By the same token, as repeated measures break the assumption of independence and are a source of stochastic variability, repetition of words across observations is a source of variability that must be accounted for in the form of the random effects produced by items, as previously done for subjects.

$$\begin{aligned}
 y_{i,j,k} &= (\beta_0 + u_{0i} + u_{0j}) + (\beta_1 + u_{1i})X_{i,j,k} + \epsilon_{i,j,k}, \\
 \epsilon_{i,j,k} &\sim \mathcal{N}(0, \sigma^2), \\
 (u_{0i}, u_{1i}) &\sim \mathcal{N}\left(0, \Sigma = \begin{bmatrix} \tau_0^2 & \rho_{u_0, u_1} \\ \rho_{u_0, u_1} & \tau_1^2 \end{bmatrix}\right), \\
 u_{0j} &\sim \mathcal{N}(0, \omega^2).
 \end{aligned} \tag{5}$$

In this new model (5), u_{0j} is the term capturing the random effects of the j^{th} item, which allows each word to have a random intercept different from the grand mean, showing that words vary in each experimental condition. As in previous models, u_{0j} is assumed to have a normal distribution with mean 0 and a variance ω^2 . Now, although declaring by-item random slopes (e.g. Baayen et al., 2008; Barr et al., 2013) in the model might be convenient—especially in experiments where a certain items vary across conditions, while others are held constant—, this would not make sense in our experiment example: reaction times for word recognition depend on word type, so there is no word that appears simultaneously in more than one condition. It would certainly be useless to specify a random effect to control for words varying according to word type (see Barr et al., 2013, for details).

Along these lines, Barr et al. (2013, p. 261) claimed in their influential paper

that the declaration of random effects must include all predictors specified as fixed effects, i.e. the ‘maximal’ random effects structure justified by the experimental design. However, they go on to suggest that:

“Although the maximal model best captures all the dependencies in the sample, sometimes it becomes necessary for practical reasons to simplify the random effects structure. Fitting LMEMs [(Linear Mixed Effects Models)] typically involves maximum likelihood estimation, where an iterative procedure is used to come up with the ‘best’ estimates for the parameters given the data. As the name suggests, it attempts to maximize the likelihood of the data given the structure of the model. Sometimes, however, the estimation procedure will fail to ‘converge’ (i.e., to find a solution) within a reasonable number of iterations. The likelihood of this convergence failure tends to increase with the complexity of the model, especially the random effects structure” (Barr et al., 2013, p. 261).

It is in this sense that Bates and his colleagues (2015a) later criticised Barr et al.’s position by stating that:

“The advice to ‘keep it maximal’ often creates hopelessly over-specified random effects” (2015a, p. 24) and “failure to converge is not due to defects of the estimation algorithm, but is a straightforward consequence of attempting to fit a model that is too complex to be properly supported by the data” (Bates et al., 2015a, p. 25).

Furthermore, one important feature of mixed models is that they permit to declare different random effects according to the way they are hierarchically grouped when more than two levels exist. Thus, *single* random effects (Figure 26a) can be *crossed* (Figure 26b) or *nested* (Figure 26c) when there is a higher

grouping factor (see Schielzeth & Nakagawa, 2012, for details). For example, as previously seen, words in by-item random effects were grouped according to word type, thus having only one factor and three levels: ‘word’, ‘pseudo-word’, and ‘second language-word’. However, these levels may also be grouped within an upper level factor, such as word length, e.g. 1-syllable, 2-syllable, and 3-syllable words. In this case random effects are said to be *nested* (Figure 26c).

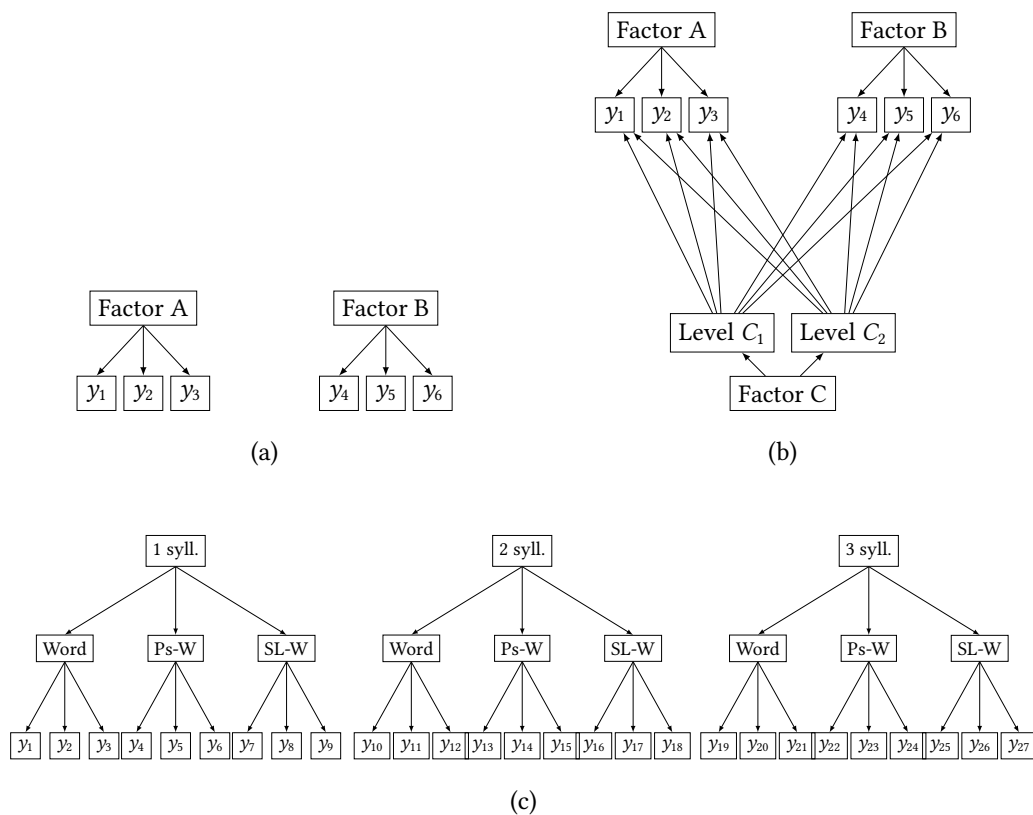


Figure 26: Different types of random effects: (a) single random effects, (b) crossed random effects, (c) nested random effects with the example experiment of word recognition used earlier for three levels: syllable length, word type, and word item. Observations are labelled as y_n . In nested random effects, observations made for items appear only once within the lowest level of the grouping factors.

Reasons for avoiding ANOVAs

Mixed models present more advantages compared to repeated measures analyses of variance (RM-ANOVAs), between-subjects ANCOVAs or mixed-model ANOVAs (e.g. Baayen et al., 2008; Gueorguieva & Krystal, 2004; Quené & van den Bergh, 2008). Apart from the fact that mixed models can handle designs that break the assumption of independence, they are also suitable to deal with unbalanced designs such as *Latin-square* designs, especially because many experiments require counterbalancing of subjects across experimental conditions in order to control for confounding variables and carry-over effects. Another important advantage of mixed models is their potential for integrating numeric predictors without needing to convert them previously into factors, such as low, medium, and high levels of a hypothetical factor, for example.

Finally, mixed models present also many advantages for all behavioural sciences, including linguistics (Figure 27). This is especially the case for analyses of categorical data, which are extremely common in behavioural sciences, whether they include a binary dependent variable or any other outcome that results in categorical grouping. In the case of a binary outcome, ordinary logistic regression has traditionally been conducted, while ANOVAs analyses have been applied with categorical outcomes after an arcsine-square-root transformation (known as ‘arcsine transformation’) of the dependent variable (e.g. Loftus et al., 1978; Mirković & Gaskell, 2016; see Warton & Hui, 2011, for a review). However, it has been noted that mixed models perform much better analyses than ANOVA by combining the strengths of logistic regression with random effects, at the same time as they benefit from the advantages of ordinary regression models (see Jaeger, 2008, for details). In this case, mixed models are said to be part of the generalised linear mixed model (henceforth GLMM) framework that makes possible to analyse different types of outcomes.

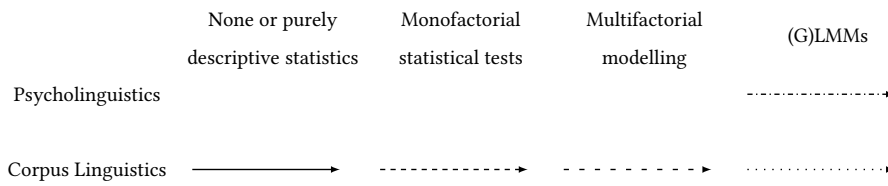


Figure 27: Evolution of statistical methods in psycholinguistics and corpus linguistics. Solidity of lines represents frequency of use. Adapted from Gries, 2015, p. 98.

3.2.2 LMMs and non-normal distributions: Generalised LMMs

As seen in the previous section, linear mixed models are just standard linear regression models that allow to declare random effects to account for stochastic variability and non-independence due to repeated measures. Apart from handling unbalanced designs, it allows to model non-normal distributions. Generally speaking, the two main types of distributions are discrete and continuous distributions, usually notated as ‘ $X \sim$ ’ to mean ‘ X is distributed as’. The *Normal* (also known as ‘Gaussian’) distribution is a continuous distribution with a symmetric bell-shaped curve in which the probability for an event to fall beyond 2 or 3 standard deviations from the mean is really small, i.e. 5% and 0.3%, respectively (Figure 28).

Differently, the binomial distribution is a discrete distribution that represents the probability of success p of an number of events in a sequence of n independent trials. In the case of n being equal to 1, i.e. one single trial, it is a *Bernoulli* distribution. If n is bigger than 1, the binomial distribution is equivalent to the sum of a number of independent Bernoulli experiments. For example, the probability of getting a number from one to six when tossing a dice is $1/6$ (a Bernoulli experiment, i.e. a single roll of the dice). However, if we wish to determine the probability of getting a certain number over 20 trials, it follows a binomial distribution:

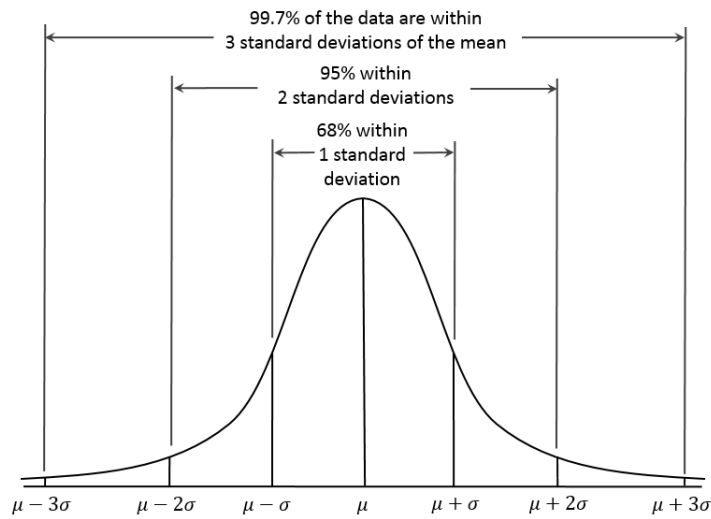


Figure 28: Normal or Gaussian density curve. The area under the curve corresponds to the probability for an event to occur, which is extremely low beyond 3 standard deviations from the mean. Adapted from Wikimedia Commons.

$$X \sim \text{Binomial}(n, p), \tag{6}$$

where n is the number of consecutive successes and p , the probability of success in each trial. Thus, the probability of getting, let's say, a six, in two separate rollings of the dice $P(x=2)$ is 19%, given 20 trials with a probability of 1/6 per trial, as seen in:

$$P(x|n, p) = \binom{n}{x} p^x (1-p)^{n-x}, \tag{7}$$

$$P(2|20, 1/6) = \frac{20!}{2!(20-2)!} 0.16^2 (1-0.16)^{2-0.16} = 0.19$$

More on binomial distributions

Statistical models dealing with binary outcomes—that is, data following a binomial distribution and therefore taking on the values 0 (no event) or 1 (success of event)—are known as *logistic regression* models. For the categorical outcome to be expressed in a linear way, a logarithmic transformation is needed. For this reason the logistic regression equation is expressed in logarithmic terms—i.e. log-odds, called the *logit* of the outcome—in order to overcome the violation of the assumption of linearity. For example, differently from linear regression, logistic regression predicts the value of y from a certain predictor variable X_1 (or several X s) by transforming the calculated probability of y occurring (8) into log-odds (9):

$$P(y) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_n X_n)}} = \text{logit}^{-1}(X_n \beta) \quad (8)$$

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_n X_n \quad (9)$$

Furthermore, the outcome expressed in log-odds (Equation 9), which are the natural logarithm of the odds ratio, can be plotted as a regression line that estimates the occurrence of an outcome y as a function of a certain predictor X_n ; and its cumulative predicted probability, i.e. the inverse of the logit function (Equation 8), can be plotted as an s-shaped curve (29).

In logistic regression, the regression coefficient β_1 amounts to the estimated increase of the outcome per unit increase in the predictor X_1 . As just mentioned, this is expressed in log-odds, and it reflects a linear relationship. However, odds ratio—the exponential of log-odds—is more commonly used as an indicator of the change in odds resulting from a unit change in the predictor and is also usually

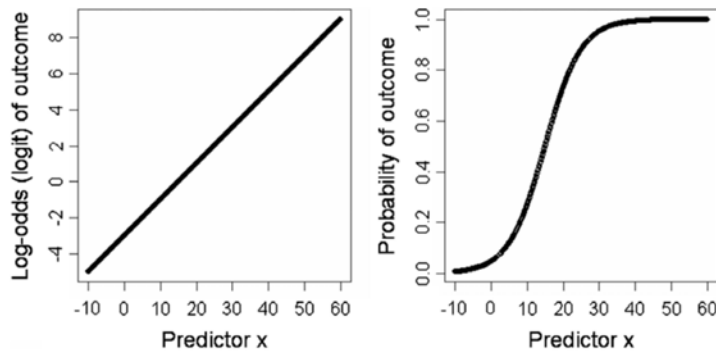


Figure 29: Example of the effect of a given predictor X on a categorical outcome y both as a linear logit function (left) and as an s -shaped curve reflecting cumulative predicted probability (right). Adapted from Jaeger, 2008, p. 438.

reported as an indicator of effect size.

As Andy Field et al. concisely explain in their brilliant effort to make statistics clear and straightforward, “the odds of an event occurring are defined as the probability of an event occurring divided by the probability of that event not occurring” (Field et al., 2012, p. 320):

$$\text{odds} = \frac{P(\text{success of event})}{P(\text{no event})} \quad (10)$$

Thus, odds ratio reflects this increase in odds resulting from a unit increase in the predictor, so that if the value is greater than 1, the odds of the outcome occurring increase as the predictor increases. Conversely, if the odds ratio is inferior to 1, it indicates that, as the predictor increases, the odds of the outcome occurring decrease (Field et al., 2012, p. 320):

$$\Delta \text{odds} = \frac{\text{odds after a unit change in the predictor}}{\text{original odds}} \quad (11)$$

As previously discussed, GLMMs are an extension of linear mixed models that

allow to handle non-normal distributions. They include a link function relating a linear predictor to a given type of outcome, and to a specific data distribution. In the case of a binary outcome:

$$\begin{aligned} y_{i,j,k} &\sim \text{Binomial}(n_{i,j,k} p_{i,j,k}) \\ p_{i,j,k} &= \text{logit}^{-1}(X_{i,j,k} \beta) \end{aligned} \tag{12}$$

In GLMMs the linear combination of fixed effects and random effects for subjects and items is achieved by declaring a term Z containing the values of the explanatory variables for the random effects, which are associated to the random-effect vector b corresponding to their coefficients. The latter term, b , is characterised by a multivariate normal distribution, with mean 0 and a variance-covariance matrix Σ :

$$\begin{aligned} y_{i,j,k} &\sim \text{Binomial}(n_{i,j,k} p_{i,j,k}) \\ \text{logit}(p) &= X\beta + Zb \\ b &\sim \mathcal{N}(0, \Sigma) \end{aligned} \tag{13}$$

The parameters obtained for GLMMs must fit the data so that the resulting model describes the data in an optimal way; however, differently from LMMs, there are not perfect solutions for the exact optimisation in GLMMs. In order to find the optimal parameters for the likelihood of data different estimation methods can be applied (see [Bolker et al., 2009](#), for details).

3.2.3 Parameters estimation and statistical inference

In the case of GLMMs, optimisation involves integrating the likelihoods of estimated parameters over all possible values of the random effects. Due to the

practical difficulties of carrying out such an optimisation, various ways have been proposed to approximate the likelihood of GLMMs' estimated parameters. In their useful review of best practices for GLMMs, Bolker and his colleagues (2009) point out the fundamental difference between two ways to achieve such a goal: standard *maximum likelihood* (ML) and *restricted maximum likelihood* (REML) estimation. In the first case, maximum likelihood in GLMMs “estimates the standard deviations of the random effects assuming that the fixed-effect estimates are precisely correct”, while REML is “a variant that averages over some of the uncertainty in the fixed-effect parameters” (Bolker et al., 2009, p. 128).

Maximum likelihood estimation in GLMMs is based on Fisher's (1922) work, which establishes that the best estimation of an unknown parameter is obtained through the logarithm of the following likelihood function:

$$\mathcal{L}(\theta|\text{data, model}) = \mathcal{L}(p|x, n, \text{binomial}) = \binom{n}{x} p^x (1-p)^{n-x} \quad (14)$$

where the likelihood \mathcal{L} of a particular numerical value of the unknown parameter θ (expressed as p in the actual parameter estimation) can be obtained ‘given’ (this is what the vertical line means) the empirical data (x observations in a sample n), and an approximating binomial model. Thus, maximum likelihood estimation would become $\log(\mathcal{L}\theta|\text{data, model})$.

Estimation methods for GLMMs were introduced in the 1990's (Aitkin, 1996; Breslow & Clayton, 1993; Schall, 1991, e.g.) based on maximum likelihood estimation. In a nutshell, they include: (a) analytic optimization of approximations of the true log-likelihood (called ‘quasi-log-likelihood’) such as: *pseudo-* and *penalised quasi-likelihood* (PQL) (Schall, 1991), *Gauss-Hermite quadrature* (GHQ) (Aitkin, 1996), and *Laplace* (Breslow & Clayton, 1993) and *nested Laplace approximation* (Rue et al., 2009); as well as (b) numerical simulations such as *Bayesian*

methods based on the *Markov chain Monte Carlo* (MCMC) technique (Zeger & Karim, 1991) (Table 10).

Estimation	Advantages	Disadvantages	Software – Function
PQL	<ul style="list-style-type: none"> • Flexible • Widely implemented 	<ul style="list-style-type: none"> • Likelihood inference inappropriate • Biased in random effects for large variances or small means 	<ul style="list-style-type: none"> • SAS[®] – PROC GLIMMIX • GenStat[®] – GLMM • R (2018) – glmmPQL – glmer
Laplace	<ul style="list-style-type: none"> • It approximates the true likelihood rather than quasi-likelihood • It allows the use of inference based on true likelihood 	<ul style="list-style-type: none"> • Slower and less flexible than PQL 	<ul style="list-style-type: none"> • SAS[®] – PROC GLIMMIX • R – glmer – glmm.admb • AD Model Builder (2012) • HLM[®]
GHQ	<ul style="list-style-type: none"> • More accurate than Laplace 	<ul style="list-style-type: none"> • Slower than Laplace • Limited to 2 or 3 random effects 	<ul style="list-style-type: none"> • SAS[®] – PROC GLIMMIX • R – glmer – glmmML
MCMC	<ul style="list-style-type: none"> • Highly accurate • Large number of random effects • Accurate 	<ul style="list-style-type: none"> • Very slow • Technically challenging • Bayesian framework 	<ul style="list-style-type: none"> • WinBUGS (2000) • JAGS (2003) • R – MCMCpack – MCMCglmm • AD Model Builder

Table 10: Overview of different estimation methods. Abbreviations: PQL (penalised quasi-likelihood); GHQ (Gauss-Hermite quadrature); MCMC (Markov chain Monte Carlo). Adapted from Bolker et al., 2009, p. 130.

The next step after estimating parameters in GLMMs involves carrying out statistical inferences whether for hypothesis testing, evaluation of goodness-of-fit among different models, or model selection. According to Bolker et al. (2009, pp. 131-132) the only test statistics appropriate for hypothesis testing in GLMMs are the Wald Z , χ^2 , t , and F tests. One of the problems for choosing a test statistic lies in assessing whether the model shows *overdispersion*, i.e. whether the variance of the response y is greater than the variance predicted by the model. This is partly due to the fact that the mean and the variance in models dealing with non-normal distributions are related and depend on the same parameter being predicted through the independent vector. Thus, the uncertainty in the estimates

as a result of overdispersion can be dealt with by Wald t and F tests, while Wald Z and χ^2 can only be applied to models without overdispersion.

Furthermore, the assessment of the goodness-of-fit of a statistical model can be carried out by comparing two models (a *null* model against an alternative model), often through the *likelihood ratio* (LR) test, which assesses the weight of a single effect, whether fixed or random. The LR test computes to what extent the likelihood of the alternative data is different from that of the null model. It yields a p -value that, compared to a critical value, allows to assess the statistical significance of the predictor under study and to reject the null model. This is achieved by comparing the fit of each model through the difference in their respective log-likelihoods:

$$\text{LR} = 2(\log\mathcal{L}_{\text{model null}} - \log\mathcal{L}_{\text{model 1}}) = -2(\log\mathcal{L}_{\text{model 1}} - \log\mathcal{L}_{\text{model null}}) \quad (15)$$

In addition, since a LR test yields a p -value, it has been widely used in model selection as well, via comparison of a series of nested models, in a procedure known as *stepwise multiple regression* (also ‘stepwise modelling’). However, Bolker et al. warn against the usage of the LR test in GLMMs³:

“Although widely used throughout statistics, the LR test is not recommended for testing fixed effects in GLMMs, because it is unreliable for small to moderate sample sizes (note that LR tests on fixed effects, or any comparison of models with different fixed effects, also require ML, rather than REML, estimates) [...]. We have found little

³ For an argument in favor of LR test in model selection using also non-nested models, see [Lewis et al., 2011](#).

guidance and no concrete rules of thumb in the literature on this issue, and would recommend against using the LR test for fixed effects unless the total sample size and numbers of blocks are very large [...]. The LR test is generally appropriate for inference on random factors, although corrections are needed [...]" (Bolker et al., 2009, p. 132).

Criticisms have suggested that model selection based on information criteria (IC) is a better practice than stepwise modelling based on LR testing, which is not limited to pairwise comparisons (e.g. Burnham & Anderson, 2002; Whittingham et al., 2006).

3.2.4 Akaike Information Criterion and model selection

Information criteria are often preferred to LR testing in model selection. According to Whittingham and his colleagues (2006), some of the problems arising in stepwise modelling include: biased parameters, model over-fitting, and incorrect significance tests (false positive results, i.e. Type I errors). Instead, stepwise modelling is replaced by model selection based on the principle of parsimony (Burnham & Anderson, 2002), which aims at identifying a model with the minimum number of predictors that satisfies a certain criterion. In this case, the minimal adequate model is that which best explains the data by containing only predictors that are significant at some specific probability level.

Procedures based on information criteria avoid the mentioned problems and permit to compare a large number of models that do not necessarily have to differ in one factor, i.e. nested models. Model selection based on information criteria, rather than estimating p -values, uses deviance as a measure of fit. Thus, multiple models are ranked based on their differences in predictive power. For

example, the Akaike Information Criterion (*AIC*) (Akaike, 1973)⁴ expresses the goodness-of-fit based on the natural logarithm of the likelihood function of the model:

$$AIC = 2K - 2\log\mathcal{L}(\hat{\theta}|y) \quad (16)$$

where $2K$ corresponds to a ‘penalisation for complexity’, actually an asymptotic bias-correction term based on the number of estimable parameters K (degrees of freedom); and $\mathcal{L}(\hat{\theta}|y)$ is the likelihood at its maximum point for the estimated model. Later on, Hurvich and Tsai (1989) introduced a correction for models with a small sample size (AIC_c) through the introduction of n in the following equation:

$$AIC_c = AIC + \frac{2K(K + 1)}{n - K - 1} \quad (17)$$

The fundamentals underlying *AIC* can be traced back to Kullback–Leibler’s (1951) ‘distance’ between two models. Kullback and Leibler established that the information loss I of a given model g when trying to approximate a certain phenomenon, a full reality f , is defined by the integral:

$$I(f, g) = \int f(x)\log\left(\frac{f(x)}{g(x|\theta)}\right) dx \quad (18)$$

where $I(f, g)$ is the distance from the model g to the perfect model or full reality f . This equation can be extended to discrete distributions, such as Poisson,

⁴ For the mathematical details of the information criterion developed by Akaike, I mainly follow the excellent review done by Burnham & Anderson, 2002, ch. 2.

binomial, or multinomial, thus becoming:

$$I(f, g) = \sum_{i=1}^k p_i \cdot \log \left(\frac{p_i}{\pi_i} \right) \quad (19)$$

where k are the possible outcomes of the underlying random variable, p_i is the true probability of the i^{th} outcome, and π_i , the approximating probability distribution, more precisely, the approximating model.

The novelty introduced by Akaike in the development of his information criterion was the estimation of Kullback–Leibler’s distance between two models by means of Fisher’s maximised log-likelihood (see Equation 14). Accordingly, the computed *AIC* values are not relevant *per se*, unless several *AIC* values from a set of models are compared. As Burnham and Anderson point out in their major revision of *AIC* and other information criteria:

“It is not the absolute size of the *AIC* value, it is the relative values over the set of models considered, [...] and particularly the *AIC* differences (Δ_i), that are important” (Burnham & Anderson, 2002, p. 63 and p.71).

The *AIC* value is only informative as a way to rank models. Therefore, model selection usually proceeds by building a set of models differing in the progressive introduction of factors relevant for, and justified by, the research. Each model candidate to be the model that best accounts for the data, i.e. the minimal adequate model, shows differences in their *AIC* values (AIC_i) respecto to the lowest *AIC* value in the set of models (AIC_{min}). These differences are expressed as Δ_i :

$$\Delta_i = AIC_i - AIC_{min} \quad (20)$$

Burnham and Anderson themselves explain the details of model ranking through differences in Δ_i in the following terms:

“As an example, candidate models g_1 , g_2 , g_3 , and g_4 have AIC values of 3,400, 3,560, 3,380, and 3,415, respectively. Then one would select model g_3 as the best single model as the basis for inference because g_3 has the smallest AIC value. Because these values are on a relative (interval) scale, one could subtract, say, 3,380 (the minimum of the 4 values) from each AIC value and have the following rescaled AIC values: 20, 180, 0, and 35. Of course, such rescaling does not change the ranks of the models, nor the pairwise differences in the AIC values [...]. We can say with considerable confidence that in real data analysis with several or more models and large sample size (say $n > 10 \times K$ for the biggest model) a model having $\Delta_i = 20$, such as model g_4 , would be a very poor approximating model for the data at hand” (Burnham & Anderson, 2002, p. 71).

If ordered from smallest to largest, the Δ_i values used in the example provided by Burnham and Anderson would be ranked as 0, 20, 35, and 180. According to this, the second Δ_i value (i.e. 20), corresponding to the difference between model g_1 and model g_3 reflects that some substantial variation is not expressed by model g_3 .

Burnham and Anderson consider that models with $\Delta_i > 10$ can be omitted from further consideration when compared to the model with the lowest AIC value, the minimal adequate model (Table 11), and they briefly conclude:

“The larger Δ_i is, the less plausible it is that the fitted model $g_i(x|\hat{\theta})$ is the Kullback–Leibler best model, given the data x ” (Burnham & Anderson, 2002, p. 70).

Δ_i	Close to the AIC_{\min}
0-2	Yes
4-7	Considerably less
> 10	No

Table 11: References for Δ_i values used in model selection when compared to the model with the lowest AIC value. Adapted from *Burnham & Anderson, 2002, p. 70*.

Arnold (2010), however, warns that in cases where a model g_i is within $\Delta_i < 2$ units respect to the AIC_{\min} model, it is possible that its additional uninformative parameters do not fairly represent a larger Δ_i difference from the minimal adequate model. Thus, it is possible to present erroneously a model g_i as being as good as the minimal adequate model. In order to avoid such a problem, he suggests reporting and discussing all models within $\Delta_i < 2$ (Arnold, T. W., 2010) (see § 3.2.4).

A step further in model selection is to quantify the plausibility of each model as being the model that best accounts for the data. For this, two additional values are computed: model likelihood and its normalisation as Akaike weights. In the first case, the concept of likelihood of the parameters $\mathcal{L}(\theta|x, g_i)$ (given both the data and the model, as previously seen in Equation 14), can be extended for a certain model given the observed data $\mathcal{L}(g_i|x)$. The details are expressed in the following equation, where ‘ \propto ’ means ‘proportional to’:

$$\mathcal{L}(g_i|x) \propto \exp\left(-\frac{1}{2}\Delta_i\right) \tag{21}$$

Model likelihood can be normalised through the assignment of so-called *Akaike weights* to each of the models being compared, which is a useful and effective way to interpret Δ_i values. In such a case, the Akaike weight of each model, w_i , can be seen as the evidence for each of them to be the one that best accounts for the

data. Overall, the computed likelihood of the set of models is normalised to add up to 1. Thus, each particular w_i depends on the entire set of models, i.e. if a model is either added or dropped from the set, w_i must be recalculated again for the new set of models (Table 12).

Model	Δ_i	$\mathcal{L}(g_i x)$	Akaike weight w_i
1	0	1	0.431
2	1.2	0.54881	0.237
3	1.9	0.38674	0.167
4	3.5	0.17377	0.075
5	4.1	0.12873	0.056
6	5.8	0.05502	0.024
7	7.3	0.02599	0.010

Table 12: Hypothetical ranking of a set of models given a certain data. Differences in AIC values are expressed as Δ_i . Akaike weights w_i , which result from the normalisation of the likelihood of models, offers evidence in favour of each model to be the one that best accounts for the data. Adapted from Burnham & Anderson, 2002, p. 77.

The relative likelihood of model pairs, i.e. $\mathcal{L}(g_i|x)/\mathcal{L}(g_j|x)$, or the equivalent ratio of Akaike weights w_i/w_j , allows to compare two models from the initial set of models, so that it “represent[s] the evidence about fitted models as to which is better in a Kullback–Leibler information sense” (2002, p. 78). For example, in the hypothetical ranking displayed in Table 12, the evidence ratio in favour of model g_1 respect to g_2 is just about 2 ($w_1/w_2 = 0.431/0.237 = 1.82$). In other words, the *evidence ratio* of model g_1 means that it is 1.82 times more likely to be the best model in Kullback and Leibler’s terms. Furthermore, evidence ratios can also be expressed in normalised probability through the expression:

$$ER = \frac{w_2}{w_1 + w_2}, \tag{22}$$

which in the example provided above between model g_1 and model g_2 would be equal to 0.645 (64.5%).

Practical challenges of AIC

Despite the insightful exposition made by Burnham and Anderson (2002) into the fundamentals of the Akaike Information Criterion, researchers are often faced with some common practical obstacles when modelling using an information theoretic approach. Grueber et al. (2011), for example, address some of the difficulties that may arise using *AIC* and suggest some solutions. However, even if Grueber et al.'s paper is mainly concerned with issues encountered in biological and evolutionary statistical modelling, their contribution is intended to encompass other fields also applying an IT approach.

Firstly, the actual predictors used in the model, which logically may include interactions and polynomial terms must be chosen from the possible input variables to include, i.e. the raw parameters as they are measured. For this, in order to avoid overparameterization, only predictors must be chosen that are justified by previous research or by the specific questions to be addressed. Thus, although there is a large number of second- and higher-order interactions, they should be included into the predictors set only if there is *a priori* reason for it. Additionally, in line with Bates et al. (2015a), Grueber and her colleagues recommend “attempting to fit both random intercepts and slopes unless the model does not converge, in which case fitting a random intercept only is preferable to not including the random variable at all” (Grueber et al., 2011, p. 703).

Secondly, after declaring predictors in the model, a set of models is created by deriving all possible submodels focusing on the predictors of interest, but not necessarily including all of them. In order to avoid non-convergence errors, it is advisable to avoid overparameterizing the initial global model.

Thirdly, a common problem encountered when models are ranked according to their *AIC* value is that there is no single best model. In this case, several of them may be within 2 *AIC* points (Burnham & Anderson, 2002)⁵ from the top-ranked model, differing by a small amount in their goodness-of-fit. Grueber and her colleagues then suggest averaging the subset of models within this cut-off of 2 *AIC* points. As a result, it may happen that a particular factor of interest is not present within this top model set, in which case they state that:

“Solutions in such cases are to either conclude that there is little evidence that the factor of interest explains variation in the response variable or extend the cut-off criteria to include at least one model that contains the factor of interest [...]. The latter solution may result in very large model sets, and/or inconsistent cut-off criteria for different response variables” (Grueber et al., 2011, p. 704).

Despite this, Burnham and Anderson remind that models that are not truly competitive with the best model and include predictors with poor explanatory power do appear within 2 *AIC* points of the model with the best goodness-of-fit:

“[These models] should be examined to see whether they differ from the best model by 1 parameter *and* have essentially the same values of the maximized log-likelihood as the best model. In this case, the larger model is not really supported or competitive, but rather is ‘close’ only because it adds 1 parameter and therefore will be within 2 Δ_i units, even though the fit, as measured by the log-likelihood value, is not improved” (Burnham & Anderson, 2002, p. 454).

⁵ Other authors set this cut-off criterion to 6 *AIC* points (Richards, 2008).

Furthermore, Arnold (2010), in his turn, insists that in such a case:

“If a truly limited set of a priori models are considered from the outset, then it probably makes sense to report and discuss all models, including those with one additional but uninformative parameter. However, the reporting should not be that these models are competitive with the higher ranked models, but rather that the additional variable(s) received little to no support, depending on the level of reduction in deviance versus the top-supported model” (Arnold, T. W., 2010, p. 1177)

And he goes on to state that:

“When a sequential modeling approach is used to evaluate a large suite of potential models, as is often done in exploratory context after first considering a more limited set of a priori models, some authors have adopted an a priori modeling approach that allows models with uninformative parameters to be discarded without further consideration” (2010, p. 1177)

In this sense, these works provide valuable insights into the role played by information criteria in model selection that will serve as a guideline for the statistical analyses conducted in the experimental part of this research. Most notably, the importance of not including second- and higher-order interactions unless there is a priori reason for it, and the additional reporting of a second-best model when it is found to lie within 2 Δ -points from the minimal adequate model.

3.2.5 Summary

The traditional use of different types of analyses of variance (ANOVAs) in psycholinguistics has experienced a paradigm change over the last decades, and

linear mixed-effects models (LMMs) have taken over as the main statistical analysis conducted, especially in cases of repeated measures, unbalanced designs, and categorical data (e.g. Gries, 2015; Jaeger, 2008). Linear mixed models are just standard linear regression models that allow to declare a set of fixed effects—coefficients of predictors not grouped by any factor—together with random effects capturing stochastic variability and non-independence due to repeated measures, i.e. coefficients grouped by a certain factor and therefore estimated with partial pooling.

Most common random effects include by-subject and by-item random effects to deal with the correlation of responses. Thus, random effects allow each subject and item to have their own intercept value differing from the grand mean, the model intercept. This results in so-called random-intercept models. In addition, similar to the deviation from the grand mean corresponding to the regression coefficient β_0 , each subject or item may also have different slope coefficients β_i for the effect of a given predictor X_i . This is what lies at the core of random-slope models.

Despite the fact that it has been suggested that random effects should be fully declared by including slopes for all fixed effects (Barr et al., 2013), it has been criticised on the grounds that this often leads to overspecifying the structure of random effects, and it often results in convergence failures, i.e. the impossibility to estimate parameters within a reasonable number of iterations) (Bates et al., 2015a).

LMMs are extended to deal with non-normal distributions, such as poisson or binomial distributions. In such a case, LMMs are often referred to as *generalised* linear mixed models (GLMMs). The estimation of parameters for GLMMs must fit the data so that the resulting model describes the data in an optimal way. Unlike LMMs, there are not perfect solutions for the exact optimisation

in GLMMs, and different estimation methods can be applied. Most are based on Fisher's (1922) maximum likelihood, and they have different advantages and disadvantages.

Traditionally, the goodness-of-fit of (G)LMMs have been tested through the *likelihood ratio* (LR) test. In the LR test, the fitted model is compared to a so-called *null* model to assess the weight of a single effect, whether fixed or random, and the resulting *p*-value allows to assess its statistical significance. However, it has been warned against this procedure (Bolker et al., 2009), and some of the problems associated to LR tests are biased parameters, model over-fitting, and incorrect significance tests (Burnham & Anderson, 2002).

A better practice based on information criteria (IC) has been suggested, which in addition is not restricted to pairwise comparisons with models differing in just one factor. Procedures based on information criteria avoid the mentioned problems and permit to compare a large number of models. The Akaike Information Criterion (*AIC*) (Akaike, 1973), for example, allows to compare a set of models based on the number of estimable parameters of each model. The grounds of (*AIC*) goes back to Kullback–Leibler's (1951) 'distance' between two models. Thus, the *AIC* value is only informative as a way to rank models, and model selection usually proceeds by building a set of models differing in the progressive introduction of factors relevant for, and justified by, the research. Finally, *AIC* differences among models are expressed as Δ_i values, which in turn can be standardised as Akaike weights w_i , i.e. the evidence for each model to be the one that best accounts for the data.

The practical obstacles that might arise when using *AIC* for model selection have been described on several occasions (e.g. Arnold, T. W., 2010; Burnham & Anderson, 2002; Grueber et al., 2011): it is important to declare an initial model containing a set of predictors justified by the research questions or by previous

studies; it is equally important to avoid overparameterization in the subset of models built from the initial global model; and finally, models within 2 *AIC* points from the top-ranked model should be also considered and reported or potentially averaged with the best ranked-model.

3.3 Methodology used in this study

3.3.1 Rationale

After having reviewed in detail the literature, a rationale logically emerges from the shortcomings of previous methodological approaches (§ 3.1.2). In order to overcome such limitations, a different methodology is applied using spontaneous speech in Castilian Spanish, aimed at addressing the two research questions addressed in this study: (1) how the different acoustic correlates of prominence relate to one another and to gestures, and (2) how gestures contribute to the perception of prominence.

On the one hand, results obtained so far using animated agents and elicited gestures in controlled experimental settings, although enlightening, have been limited in their explanatory potential. Some of the problems associated to both approaches lie in the difficulty of eliciting spontaneous gestures while keeping experimental control over the acoustic cues of prominence. Additionally, the range of body movements previously analysed have mostly been restricted to head and eyebrows.

On the other hand, and probably due to the methodological shortcomings previously discussed, the relation of visual cues—as in the gestures of hands, eyebrows, and head—to the different acoustic correlates of prominence has not been fully addressed yet. This is in line with the preponderant role traditionally given to f_0 , where, for example, the relation between eyebrow raise and f_0

has been largely studied. The lesser importance given to intensity and duration on the one hand, and the still controversial interaction of the different acoustic correlates of prominence on the other hand, have prevented from analysing how f_0 , intensity, and duration relate both to one another and to gestures in the perception of prominence.

Additionally, Castilian Spanish is the language of research in this study, differently from previous research, which has focused on several other languages (e.g. Dohen & Løevenbruck, 2009, for French; Granström et al., 1999; House et al., 2001, for Swedish; Krahmer & Swerts, 2007, for Dutch; Prieto et al., 2015, for Catalan).

The use of spontaneous speech in the study of prominence is not new. Both Ambrazaitis and House (2017) and Swerts and Krahmer (2010, experiment 1: ‘auditory mark-up’) have resorted to recordings of TV newsreaders, although their research questions were different from one another and also different from those addressed here. The former researchers analysed eyebrow and head movements in their relation to information structure and pitch accent distribution; the latter studied prominence perception in the auditory modality.

Krahmer and Swerts (2007, p. 411) themselves had previously noted the necessity for spontaneous speech material to complement current research in the study of the multimodal perception of prominence (§ 2.3.8.4). The research lying at the origin of this dissertation is partly based on Krahmer and Swerts’s call for spontaneous speech. Thus, the use of spontaneous speech material in this research is intended to overcome the experimental paradigm in which one stimulus typically contains one gesture, which is clearly very far from the natural interweaving of gestures performed with hands, head, and eyebrows found in everyday spoken language.

Apart from TV newsreaders, audiovisual recordings from reality television—

especially from television talent shows—are another possible source of spontaneous speech stimuli. Spontaneous speech from reality television has already been used, with a high degree of ecological validity, in discourse analysis, pragmatic studies, and linguistic variation (Bednarek, 2013; Eberhardt & Downs, 2015; Sonderegger, 2012). In this research, it is argued that within reality television, some television talent shows (e.g. *Factor X* in UK, *Operación Triunfo* in Spain), differently from TV newsreaders or interviews conducted on TV sets, offer an optimal source of audiovisual speech material: contestants of talent shows are likely to speak and act more naturally over the long periods of time while they are recorded with hidden cameras than other speakers in more conventional recording sessions; their conversations have a high degree of spontaneity; and the amount of recorded material is large and easily accessible.

Finally, as Dohen and Løevenbruck (2009) noted, a possible way of exploring audiovisual prosody in prominence perception involves creating audiovisual mismatches, i.e. incongruent stimuli in which the audio and visual cues convey contradictory information, as previously done by Swerts and Kraemer (2004; 2008). Yet, Kraemer and Swerts (2007) expressed skepticism about the possibilities offered by spontaneous speech stimuli in a mismatch paradigm. In their turn, Dohen and Løevenbruck, studying audiovisual prominence perception, opted for designing auditorily degraded prosodic stimuli and, consequently, elicited whispered speech in an experimental setting. This method proved useful in their study and paved the way for similar studies using auditorily degraded stimuli.

In this research, differently from both Kraemer and Swerts and Dohen and Løevenbruck, but following the same logic pointed out by the latter researchers, auditorily degraded stimuli are created and administered in three experimental conditions corresponding to the neutralisation of the acoustic correlates of prom-

inance: f_0 , intensity, and both f_0 and intensity. The multimodal perception of acoustic prominence is analysed in two different stages (Experiment I § 4 and Experiment II § 5) following Streefkerk's (2002) research by means of a binary prominence-marking task conducted at word level.

3.3.2 Speech material

The speech materials, publicly available on the website *youtube.com*, were extracted from audiovisual recordings captured with hidden cameras in the talent show *Operación Triunfo* (1st edition). The recordings show the contestants of the talent show engaged in spontaneous conversation. From these TV recordings, a small corpus of 50 videoclips was created. The criteria for the selection was that each videoclip showed one single speaker uttering a sentence without being interrupted while gesticulating with either hands, eyebrows, or head, or with several of these body parts simultaneously. Thus, from these criteria the collected corpus reflect very well the natural interweaving of gestures performed with hands, head, and eyebrows typically found in everyday spoken language. Actually, it was relatively difficult to find utterances containing only one or two distinct gestures performed with one articulator. Conversely, speakers were found to be engaged in conversation without needing to gesticulate and then they would suddenly engage in abundant gesticulation while uttering full meaningful sentences. This last example was the basis for our corpus. Therefore, the speaker in each videoclip is always seen throughout the utterance performing some gesture in any of its phases at all times, so that and their gestures follow each other in a continuous fashion. It is in this sense that the corpus could be described as a 'hypergestural', where hardly any words occur without the presence of a gesture.

Out of these 50 videoclips, 30 were manipulated and used as target sentences,

10 were used as stimuli without any manipulation, and the remaining 10 were used as trial sentences for participants to get familiar with the experimental task (Appendix A). Subsequently, the 30 audio tracks of the videoclips used as target sentences were extracted, and their speech signal was manipulated with PRAAT (Boersma & Weenink, 2018)⁶. The audio tracks resulting from the manipulation were used to replace the original speech signal in the videoclips by means of the iMovie[®] software of Apple[®]. The manipulation of the stimuli allowed to present them to participants in three conditions and in two modalities: audio-only and audiovisual.

The speech material for Experiment I were the 30 sentences, out of the 50 sentences collected in this small corpus, that corresponded to 10 speakers (5 men, 5 women) that uttered a total of 531 words. Sentences ranged between 9 and 32 words ($M = 17.7$, $SD = 5.83$), with a duration between 2.51 and 9.49 seconds ($M = 4.60$, $SD = 1.58$). The speech rate was calculated for each sentence as the relation between words and seconds ($M = 4.00$ words/sec, $SD = 1.15$), with a minimum of 1.15 and a maximum of 6.4 words per second (Appendix A1).

For Experiment II, however, 13 sentences were selected from the initial corpus. Out of these, four were manipulated and used as target sentences. These four sentences were chosen so that there were one female and one male speaker, each uttering two sentences—one shorter and one longer—that were equivalent for both speakers in length and speech rate. The four sentences ranged between 9 and 18 words ($M = 13.5$, $SD = 3.69$), with a duration between 2.51 and 3.38 seconds ($M = 2.81$, $SD = 0.38$). The mean speech rate was 4.68 words per second ($SD = 0.92$) (Appendix B2). Out of the 13 sentences, 4 were used for instructions and trials, 2 were non-manipulated stimuli, and the remaining 3 were used as filler sentences.

⁶ This material is publicly available at <http://dx.doi.org/10.17632/jkvftnpr5j.1>

3.3.3 Stimuli creation

Following the idea of Dohen and Lœvenbruck (2009) of using degraded speech as an alternative to a mismatch paradigm, the acoustic correlates of prominence were manipulated to neutralise the prominence-lending properties of both f_0 and intensity. This allows to analyse how the resulting available cues relate to each other in the perception of prominence. The available cues in each condition are: (0) all acoustic cues—control condition; (1) intensity and duration; (2) f_0 and duration; (3) only duration.

The manipulation of stimuli, conducted with PRAAT, was done in the following way: in the first experimental condition, f_0 was smoothed for each speaker within a frequency range of 20 Hz in the intonation curve in Experiment I, which corresponded to an average of 2.22 semitones (ST) ($SD = 0.68$ ST) for all sentences. In both experiments, the resulting smoothed intonation curve was resynthesised using linear predictive coding (LPC) (Figure 30).

Differently, in Experiment II, the manipulation of the intonation curve was made using a semitone scale instead. This seemed more adequate, since the perception of pitch intervals is not linear but logarithmic; for example, the difference between 100 and 200 Hz is not the same as that between 150 and 300 Hz, i.e. twice the frequency does not correspond to a twofold increase in pitch. Nevertheless, a precise standard of *just noticeable differences* (JNDs) for the perception of f_0 is not agreed upon. For example, one study with disyllabic structures [baba] determined that 2 ST is the threshold beyond which it is difficult not to perceive intonation differences (Pamies et al., 2002). However, 't Hart (1981) raises this threshold to 3 ST, while Rietveld and Gussenhoven (1985) reduce it to 1.5 ST. For the neutralization of f_0 in Experiment II, following Pamies et al. (2002), the intonation curve was strictly kept within a maximum difference of 2 ST.

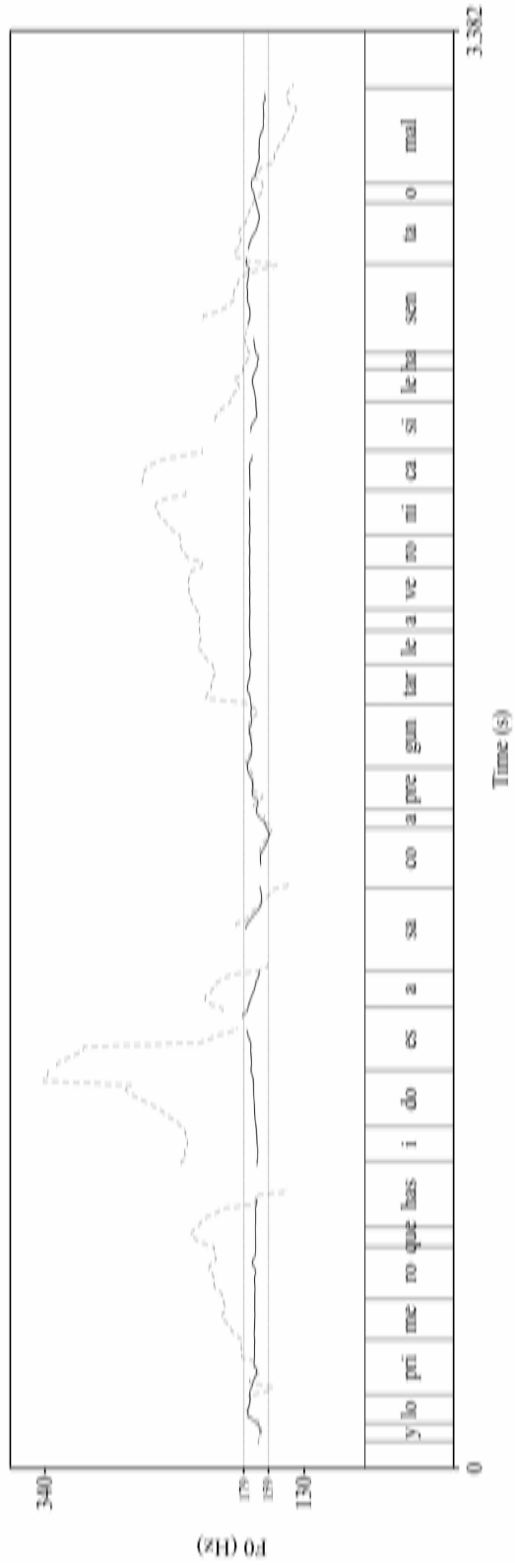


Figure 30: Example of f_0 manipulation for one of the sentences in Experiment I, showing the fundamental frequency of both: the original speech signal (dashed line) and after neutralization within a 20 Hz range (solid line), which in this example corresponded to 2.05 ST.

3.3.4 Participants

The sample size of participants in Experiment I was $N = 12$, and they were recruited among university students in Madrid. In an attempt to strike a gender balance, 6 male and 6 female participants were selected. None reported having any visual or hearing impairment. In Experiment II, due to the between-subjects design, the initial sample size of participants increased to $N = 320$, from which only 240—68 men and 172 women—met a certain set of criteria and whose responses were finally selected. In this case, participants were recruited via social media and several Internet forums. After conducting the online test, information about them was collected—such as place of birth, country of residence, mother tongue, age, and musical training. Additionally, in both experiments two trained listeners with academic background in phonetics independently also provided marks of prominence for both modalities relying on all acoustic cues of prominence, which served as a ‘gold standard’ against which to compare the responses given by participants. Specific details about participants in each experiment will be provided in the respective sections (§ 4.2.1 and § 5.2.1).

Additionally, the two experiments on prominence perception of this research have taken into consideration studies revealing important perceptive differences resulting from the musical competence of participants. For example, neuropsychological research has observed that both pitch contour and rhythmic grouping are aspects shared by both music and speech (Patel et al., 1998). Similarly, it has been reported a transfer of abilities between music and linguistic prosody for perceiving pitch and rhythmic variations (Thompson et al., 2004). Furthermore, several studies have explored the extent to which musical training might affect prosody perception. For example, Hutka et al. (2015) conducted a study using EEG and determined that pitch discrimination was enhanced in trained

musicians as well as in speakers of tone languages when compared to native English-speaking non-musicians. Similar results were obtained by e.g. Liang et al. (2016) and Zioga et al. (2016).

Therefore, information on the musical competence of participants was collected in both experiments. In Experiment I, as a pilot study, the small sample size, $N = 12$, did not include any participants having musical training. Conversely, in Experiment II, which had a much larger sample size, $N = 240$, the musical training of participants was included as a variable in the statistical models.

3.3.5 Gesture annotation

As for gestures, all sentences were annotated with ELAN (Brugman & Russel, 2004) according to the body part involved in the production of the gesture, both separately (hand, eyebrows, and head) and in combination (hand-eyebrows, hand-head, eyebrows-head, hand-eyebrows-head) (second tier in Figure 31). Since it was considered irrelevant in this research, any further categorisation of gestures as beats, deictic, iconic, metaphoric, etc. was omitted. Rather, their different phases were annotated: preparation, stroke, apex, hold, recoil, and retraction (third tier in Figure 31).

The annotation of the different gesture phases presented some difficulties, especially when gestures included several articulators. Thus, when the gesture was performed with hands together with any other articulator, the annotation followed the most visible movement of the hands, since the second articulator—or possibly the remaining two articulators, i.e. head and eyebrows—hardly presented any preparation or stroke phase, but the gesture was sudden and its apex coincided with that of the apex phase of the hand gesture. Alternatively, the annotation of a gesture performed uniquely with either head or eyebrows could

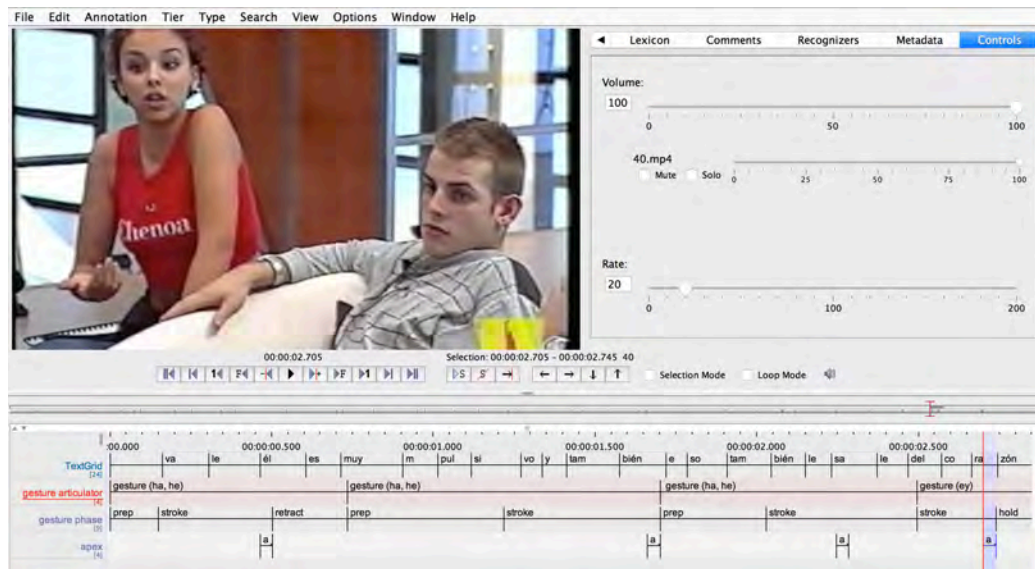


Figure 31: Sample screen of ELAN annotation. The still shows the speaker on the left performing an eyebrow raise in its effort peak (apex). The cursor position in the bottom right corner shows how the apex phase is embedded within the stroke of a gesture performed as an eyebrow raise. The four tiers used in annotation appear at the bottom: TextGrid with syllabic segmentation, gesture articulator, gesture phase, and apex.

include the preparation and stroke phases with the words they co-occurred with (Figure 31), although in most cases these two phases were relatively short and stretched over a small fraction of an uttered word. In this sense, for example, if a head nod or an eyebrow raise included a short preparation followed by a stroke co-occurring both in the same word, it was preferred to annotate the perceptually more relevant phase of the gesture—i.e. apex instead of stroke, and stroke instead of preparation.

As previously discussed, both strokes and apexes of gestures align with different intonational units (Esteve-Gibert & Prieto, 2013; Kendon, 1972; Loehr, 2012). The annotation of gestures into their different phases included preparation, stroke, apex, post-stroke hold (or simply ‘hold’), recoil, and retraction (see § 2.3.4.5). The segmentation of gesture into its different phases aims at better

understanding the role of each of them in the perception of audiovisual prominence, especially strokes, corresponding to the central unit of the gesture and spanning over an interval of time (Kendon, 1972, 1980); and apexes, the peak effort within the stroke, corresponding to an instant in time (e.g. Loehr, 2004, 2007) (fourth tier in Figure 31).

3.3.6 Data analysis

The data collected in this research was firstly annotated acoustically and visually. The acoustic measurements are conducted in all sentences for all acoustic correlates of prominence: (a) maximum f_0 of stressed vowels—on some occasions, the vowel of the adjacent syllable due to f_0 shift, as previously noted (e.g. Garrido et al., 1993, 1995); (b) mean intensity of the stressed vowel; and (c) mean duration of stressed syllable, in which all segments of the syllable are taken into account (Face, 2000). In Experiment II, next to these acoustic measures, spectral balance was also added, as the difference between the amplitude of the first harmonic and the second harmonic, H1–H2⁷ (e.g. Heldner, 2003) (Table 13).

Acoustic correlate	Acoustic measure	Measure unit	Experiment
Fundamental frequency (f_0)	Max. f_0 in stressed vowel (or vowel in adjacent syllable)	Hertz (Hz)	1 – 2
Intensity	Mean intensity of stressed vowel	Decibels (dB)	1 – 2
	Spectral balance: H1–H2	Decibels (dB)	2
Duration	Mean duration of stressed syllable	Seconds (s)	1 – 2

Table 13: *Acoustic measures conducted for the acoustic correlates of prominence in both Experiment I and Experiment II.*

In each experiment the results obtained from the data are then firstly analysed

⁷ I thank Dr. José María Lahoz Bengoechea for his helpful insights on this point and his generosity in sharing his PRAAT scripts.

in a *Descriptive statistics* section (4.3.1 and § 5.3.1). This provides a first approximation to the results and allows to understand how the marks of prominence given by participants are distributed. After such an initial description, an section devoted to *Inferential statistics* follows in each experiment (§ 4.3.2 and § 5.3.2). Participants' prominence marks are first assessed inferentially in their degree of agreement through averaged Cohen's kappa (1960) for all pairs of participants, as done in previous studies (e.g. Buhmann et al., 2002; Streefkerk, 2002). Cohen's kappa yields a coefficient for the agreement of a pair of raters taking into account agreements by chance. Alternatively, some authors have used Fleiss's kappa (1971), which goes beyond pairwise comparisons and provides a single value for agreement among more than two raters (e.g. Mo, 2008a; Cole et al., 2010). In this research, the Cohen's kappa statistic is preferred, since the responses of two trained listeners relying on all acoustic cues of prominence serve as a 'gold standard' in both experiments. Their marks are compared pairwise to those of each participant across experimental conditions and modalities. Thus, at a first stage, mean Cohen's kappa is computed separately for participants and for the two trained listeners. Then, Cohen's kappa is calculated comparing the responses of each participant to those of the trained listeners in order to assess how they deviate from this 'gold standard'. The resulting mean is offered and compared to the value calculated previously for only participants.

Subsequently, as explained above (§ 3.2.2), generalised linear mixed models (GLMMs) are computed for a certain number of variables predicting prominence perception. The reasons for avoiding a traditional analysis of variance (ANOVA) have been previously stated (§ 3.2.1). In both experiments, the binary outcome of the dependent variable y (prominent vs. non-prominent) is estimated through a logit link function. For this, statistical modelling is conducted using the open source statistical software R (2018)—and its integrated development environment

(IDE), RStudio (2016)—using the function *glmer* from the *lme4* package (Bates et al., 2015b). It is worth mentioning that, in the statistical models, the different gesture phases were introduced as separate variables using dummy coding, rather than as different levels of one single variable. This was done so partly due to the absence of a reliable baseline against which to compare each gesture phase, since very few words in the corpus were uttered in the absence of a gesture. (In this sense, as just mentioned (§ 3.3.2), the sentences uttered by speakers in the corpus were selected because they had a complete sense and were accompanied by an intertwined succession of gestures, which resulted in very few words co-occurring with no gestures). Additionally, by introducing each gesture phase as a separate variable, it was possible to compare the effect that the presence of each of the phases had in relation to their absence.

The use of GLMMs allows to account for the effects of fixed and random predictors on the binary outcome, which are estimated through maximum likelihood method (Laplace Approximation). The different fitted models are then compared in a model selection procedure using *Akaike information criterion (AIC)* (Akaike, 1973), as explained above (§ 3.2.4). Model selection by means of *AIC* permits to establish the model, from a set of estimated models, that best accounts for the data; this ‘minimal adequate model’ is the model that cannot be improved by either adding or dropping any predictor. *AIC*, differently from traditional *p*-values, simply provides an ordinal value assessing the quality of models respect to their complexity, thus lacking any meaning of its own except as a way to rank models from a set. Finally, details of the respective minimal adequate model resulting from the selection are discussed.

3.3.7 Summary

The methodology applied in this research intends to overcome some of the limitations inherent to methodologies used in previous studies in order to address the questions of how the different acoustic correlates of prominence relate to one another and to gestures, and how gestures contribute to the perception of prominence. For this, spontaneous speech samples were collected from the talent show *Operación Triunfo* (1st edition) that were publicly available on the website *youtube.com*, and which show close shots of speakers engaged in spontaneous conversations. From these TV recordings a small corpus of 50 videoclips was created. In the Experiment I, 30 of these videoclips were manipulated to be used as target sentences, while 10 were used as non-manipulated stimuli and the remaining 10, as trial sentences. In Experiment II, only 13 sentences were selected from this initial corpus: 4 were manipulated and used as target sentences, 4 were used for instructions and trials, 2 were non-manipulated stimuli, and the remaining 3 were used as filler sentences for an extra experimental task different from prominence marking.

The speech signal of target sentences was manipulated to reduce the prominence-lending properties of the f_0 and intensity, so that the available cues in each condition were: (0) all acoustic cues—control condition; (1) intensity and duration; (2) f_0 and duration; (3) only duration. The manipulation of intensity involved smoothing intensity to a constant value of 69 dB, and smoothing f_0 within a range of 20 Hz in Experiment I and within a range of 2 semitones in Experiment II.

The sample size of participants was different in each experiment. In Experiment I, as a pilot study, participants added up to 12 individuals (6 men and 6 women), while in Experiment II the sample size added up to 320 individuals, out

of which only 240 were selected (68 men and 172 women). In both experiments, information about participants was collected including place of birth, country of residence, mother tongue, age, and musical training. Additionally, in both experiments two trained listeners with academic background in phonetics independently provided marks of prominence for both modalities relying on all acoustic cues of prominence. This served as a ‘gold standard’ against which to compare the responses given by participants.

All gestures were annotated according to the body part involved in their production both separately (hands, eyebrows, and head) and in combination, as well as as their different phases: preparation, stroke, apex, hold, recoil, and retraction. Similarly, acoustic measurements for all acoustic correlates of prominence in each sentence of the corpus were conducted as: (a) maximum f_0 of stressed vowels—on some occasions, the vowel of the adjacent syllable due to f_0 shift; (b) mean intensity of the stressed vowel; and (c) mean duration of stressed syllable. In Experiment II, next to these acoustic measures, spectral balance was also added, as the difference between the amplitude of the first harmonic and the second harmonic, H1–H2.

The results were reported in a *Descriptive statistics* section, followed by a *Inferential statistics* section, in which generalised linear mixed models (GLMMs) were first estimated and then compared through the Akaike Information Criterion. Finally, the minimal adequate model from the set of estimated models was assessed in order to determine the weight of each predictor variable in predicting the marks of prominence given by participants.

Experiment I

4.1 Introduction

In a first attempt to address the two questions of this research (Question 1 and Question 2), a pilot study on the multimodal perception of prominence in Spanish¹ was initially envisaged.

Here a new methodology is presented, which uses spontaneous speech as a way to overcome some of the limitations of previous methodological approaches (§ 3.1). As stated there, the shortcomings of previous methods make it difficult to address how the different acoustic correlates of prominence relate to one another and also to gestures. On the one hand, studies on prominence perception with animated agents in the form of talking heads have limited themselves to reproduce eyebrow and head movements but have excluded hand movements (e.g. House et al., 2001; Prieto et al., 2011). On the other hand, stimuli used in experimental settings have mostly been created through audiovisual recordings in a controlled environment (e.g. Dohen & Løevenbruck, 2009; Krahmer & Swerts, 2007). For example, the participants of Krahmer and Swerts' study were instructed to utter a short sentence and produce concomitantly a quick visual gesture

¹ This section is to appear with some minor changes in Jiménez-Bravo and Marrero (submitted).

with either hand, eyebrows, or head on a specific target word. Participants were also allowed to train until they felt they could not improve the realisation of the gesture accompanying the uttered sentence. The audiovisual recordings so obtained were later used as stimuli in an experiment on the effect of gestures to prominence perception.

Differently from this practice, in this research spontaneous speech samples are used as stimuli in an experiment on multimodal prominence perception. In this sense, the goal of this study was precisely to assess the validity of the described methodology and obtain some initial results. For this, a small corpus was created with videoclips extracted from a Spanish talent show, and stimuli were neutralised in the prominence-lending properties of their acoustic cues. In this pilot study, stimuli were administered to 12 naïve listeners under three conditions in two modalities, i.e. audio-only and audiovisual. Participants marked words for binary prominence at word level (prominent vs. non-prominent). Additionally, two trained listeners with academic background in phonetics independently provided marks for both modalities relying on all acoustic cues of prominence, which were used as a control condition. The obtained responses were modelled in R using generalised linear mixed models (GLMMs) through a logit link function with different fixed and random effects. The resulting models were then compared using the Akaike Information Criterion (*AIC*) (Akaike, 1973). Finally the minimal adequate model was summarised in order to establish the effects of the tested variables on the prominence-marking task.

4.2 Methodology

4.2.1 Participants

Twelve naïve listeners rated a total of 6372 words available for marking (531 words per participant). All were Spanish native speakers (6 men and 6 women, $M_{\text{age}} = 21.5$ years, $SD = 2.31$) from Madrid, and none of them had formal musical education (§ 3.3.4). They were financially compensated with 10€, and all were unaware of the purpose of the study. Additionally, two trained listeners with academic background in phonetics also marked the same stimuli as participants and relied on all non-manipulated acoustic cues of prominence.

4.2.2 Stimuli

The recordings that served as the source of speech samples showed the contestants of the television talent show *Operación Triunfo* (1st edition) speaking spontaneously with each other (§ 3.3.2). From these recordings a small corpus of 50 videoclips was created. Stimuli included 30 manipulated videoclips², 10 non-manipulated videoclips, and the remaining 10 were used for instructions and trials. In each videoclip a single speaker was visible while uttering a sentence in a spontaneous conversation and gesturing with either hands, head, or face, or with several of these articulators simultaneously. As previously mentioned (§ 3.3.5), the basis for our corpus was examples of speakers engaged in active gesturing while uttering full meaningful sentences. So the speaker in each videoclip is always seen throughout the utterance performing some gesture in any of its phases at all times, so that and their gestures follow each other in a continuous fashion. It is in this sense that the corpus could be described as a ‘hypergestural’, where

² This material is publicly available at <http://dx.doi.org/10.17632/jkvftnpr5j.1>

hardly any words occur without the presence of a gesture.

As previously discussed, both strokes and apexes of gestures align with different intonational units (Esteve-Gibert & Prieto, 2013; Kendon, 1972; Loehr, 2012). The annotation of gestures into their different phases included preparation, stroke, apex, post-stroke hold (or simply ‘hold’), recoil, and retraction (see § 2.3.4.5). The segmentation of gesture into its different phases aims at better understanding the role of each of them in the perception of audiovisual prominence, especially strokes, corresponding to the central unit of the gesture and spanning over an interval of time (Kendon, 1972, 1980); and apexes, the peak effort within the stroke, corresponding to an instant in time (e.g. Loehr, 2004, 2007) (fourth tier in Figure 31). The annotation of the different gesture phases presented some difficulties, especially when gestures included several articulators (see § 3.3.5 and Figure 31).

In the first experimental condition (C1), the acoustic cue of f_0 was neutralised, and intensity ranged between 50-82 dB ($M = 71.20$, $SD = 6.01$) for stressed vowels, while duration ranged between 0.023-0.454 seconds for stressed syllables ($M = 0.120$, $SD = 0.061$). In the second experimental condition (C2), intensity was homogenised at an average of 69 dB ($SD = 0.11$ dB) for all speakers, while f_0 ranged between 93-344 Hz ($M = 194.65$, $SD = 48.89$) for maximum f_0 in stressed vowels (or the vowel in the adjacent syllable), and duration kept the same values as in C1. In the third condition (C3), both f_0 and intensity were neutralised with the same criteria as in C1 and C2, while duration maintained the same values as in C1. In each of the conditions, the acoustic cues available to participants were intensity and duration in C1, f_0 and duration in C2, and just duration in C3.

4.2.3 Experiment design

The experiment had a within-subjects design in which the 30 target sentences were first grouped in six blocks (from A to F) containing five clips each. Blocks were sorted following a Latin square pattern. Additionally, non-manipulated stimuli were included and randomised across all blocks. The 30 target sentences were administered in two parts corresponding to each modality, first audio-only and then audiovisual, or vice versa. Initially, participants received the first half of the stimuli, which corresponded to one modality, in three blocks (a total of 15 clips, i.e. five clips per condition, plus 5 non-manipulated sentences). Later, the next three blocks (a total of 15 clips plus 5) for the remaining modality were similarly presented. Since the experiment was designed for 6 participants, the same procedure was repeated twice until collecting responses from 12 participants. All of them marked all stimuli but never the same stimulus in different conditions. The order of the three conditions and the two modalities was counterbalanced to avoid carry-over effects (Figure 14).

4.2.4 Hypotheses

1. Methodology

The spontaneous nature of the speech samples used in this research will prove adequate to the experimental task, and the manipulations conducted on them will not hinder participants to detect prominence beyond chance, i.e. participants will show sufficient agreement.

2. Acoustic correlates of prominence

2.1 Significant differences will be found for the experimental conditions when compared to the control condition.

Experiment I

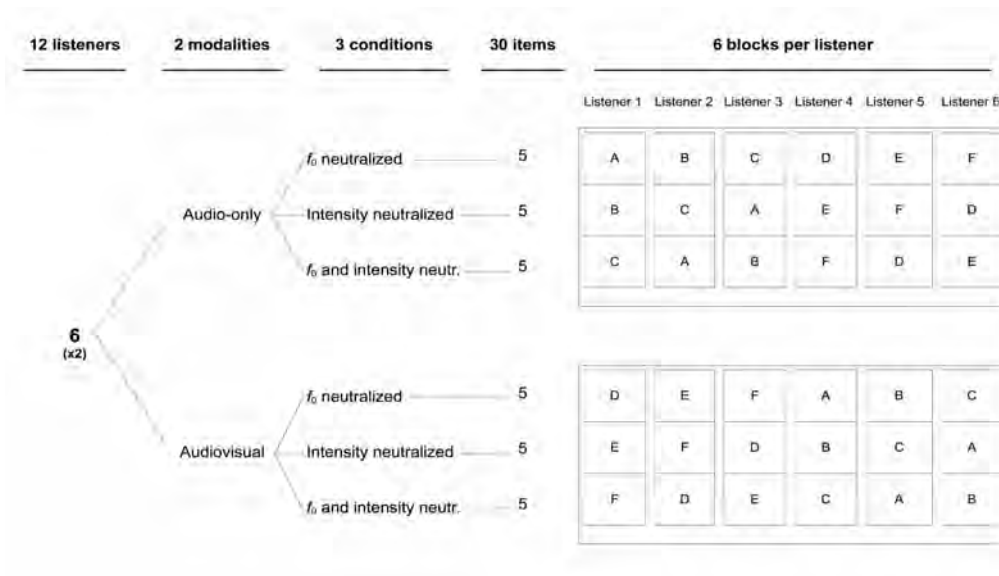


Table 14: Scheme showing the experiment design with a Latin square to reorder stimuli blocks corresponding to 30 target sentences. 10 non-manipulated sentences were randomly placed across all blocks.

2.2 The perception of prominence will be rendered through a combination of acoustic correlates, i.e. no single correlate will suffice for participants to detect prominence.

3. Gestural correlates of prominence

3.1 There will be significant differences between the audio-only modality and the audiovisual modality.

3.2 Prominence perception will mostly be driven by the apex phase of gestures.

4.2.5 Procedure

Participants were briefly instructed about the nature of the task so as to determine the word or words in each sentence that were prosodically emphasised.

For this, prominence was informally defined as ‘words which were produced with clear auditory emphasis’. Participants were provided with the minimal necessary information to conduct the task successfully.

After this, participants were individually presented with the stimuli through headphones in a sound-proof cabin. The stimuli were administered under supervision at the phonetics laboratory of UNED by means of the online software *SurveyGizmo*³ on a MacBook Pro. Following Streefkerk and her colleagues (1997), a binary-marking experimental task was used, in which all words of the sentence were available for marking. Participants could mark any number of words they perceived as prominent in the sentence and leave non-prominent words unmarked. For participants to become familiar with the task and to ensure they had understood it properly, they were given a practice (2 trials) before each modality. After this practice, the marking task involved two stages: when presented with a stimulus, participants could watch or listen to it by playing back the clip up to three times. Then, the corresponding words of the sentence were displayed on the computer screen together with check-boxes for marking. At that moment, they were allowed to play back the stimulus again up to four more times until they made a final decision with the prominent words they had marked in the sentence (Figure 32). Participants needed approximately 45 minutes to complete the task.

³ <https://www.surveygizmo.com>.



Figure 32: Sample screen of the experimental task that was conducted by participants in the audiovisual modality. After an initial screen displaying only the instructions and the audiovisual stimulus, check-boxes appeared at the bottom for participants to conduct marking.

4.3 Results

4.3.1 Descriptive statistics

4.3.1.1 Prominence marks

The 531 words available for marking per participant added up to 8496 words (including also the words marked by two trained listeners). Out of these, 2058 (24.22%) were marked as prominent. Prominence marks per sentence ranged between 2.75 and 6.50 ($M = 4.51$, $SD = 1.01$). The 10 speakers received on average 165.2 ($SD = 104.5$) marks of prominence, while the 12 participants marked on average 135.41 words as prominent ($SD = 38.76$) (Table A2). Additionally, interjections (64.5%), adjectives (48.9%), and nouns (45.2%) were the word classes that received the highest percentage of prominence marks (see Appendix A3 for details).

4.3.1.2 Inter-rater agreement

Similar to other studies computing inter-rater agreement (Buhmann et al., 2002, e.g), cross-tables for all pairs of participants were made (Table 15a). Then, overall inter-rater agreement through mean Cohen’s kappa (1960) was computed in order to take into account agreement of participants beyond chance, as explained above (§ 3.3.6).

Inter-rater agreement for all paired combinations of participants for all stimuli across conditions and in both modalities had a value of $\kappa = .39$. When agreement was computed for each participant respect to the ‘gold standard’ provided by two trained listeners, mean Cohen’s kappa was slightly lower, $\kappa = .38$. This lower mean reflects the extent to which participants’ responses diverged from those of the two trained listeners, whose independent agreement was high ($\kappa = .83$) (Table 15b).

Participant 11	Participant 12			Agreement
	Non-prom	Prom	Total	
Non-prom	364	85	449	79%
Prom	22	60	82	
Total	386	145	531	$\kappa = .41$

(a) Example of cross-table with non-prom(inent) and prom(inent) marks given by participants 11 and 12. Agreement shows percentage of coincidental responses for both participants and Cohen’s kappa coefficient.

Experiment I

Participant	1	2	3	4	5	6	7	8	9	10	11	12	TL 1	TL 2
1	-	.42	.35	.28	.42	.28	.40	.38	.29	.33	.40	.44	.37	.35
2	-	-	.34	.26	.39	.38	.40	.48	.24	.39	.26	.39	.34	.31
3	-	-	-	.29	.33	.28	.21	.34	.27	.31	.32	.37	.30	.24
4	-	-	-	-	.26	.25	.26	.24	.33	.18	.32	.33	.34	.33
5	-	-	-	-	-	.48	.47	.35	.44	.39	.49	.40	.43	.42
6	-	-	-	-	-	-	.36	.35	.27	.33	.36	.40	.33	.31
7	-	-	-	-	-	-	-	.36	.42	.42	.42	.44	.40	.37
8	-	-	-	-	-	-	-	-	.25	.34	.30	.45	.31	.32
9	-	-	-	-	-	-	-	-	-	.31	.38	.28	.38	.38
10	-	-	-	-	-	-	-	-	-	-	.29	.40	.37	.36
11	-	-	-	-	-	-	-	-	-	-	-	.41	.37	.36
12	-	-	-	-	-	-	-	-	-	-	-	-	.41	.39
TL 1	-	-	-	-	-	-	-	-	-	-	-	-	-	.83
TL 2	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Mean Cohen's kappa													.39	.38

(b) Inter-rater agreement expressed as mean Cohen's kappa for all pairs of participants including two trained listeners (TL).

Table 15: Cross-table of prominence marks and inter-rater agreement.

4.3.1.3 Prominence and acoustic cues

As it is apparent from Table 16, words considered prominent regardless of modality increased across conditions respect to the control condition.

Modality	Words for marking per condition and modality	Prominent words (%)			
		C0	C1	C2	C3
	2124	434 (20.4)	539 (25.3)	530 (24.9)	555 (26.1)
A	1062	198 (18.6)	269 (25.3)	259 (24.4)	267 (25.1)
AV	1062	236 (22.2)	270 (25.4)	271 (25.5)	288 (27.1)
AV ($\pm\%$)		+3.6	+0.1	+1.1	+2.0

Table 16: Distribution of marks of prominence per experimental condition for both modalities combined ($n = 2124$) and for each modality separately ($n = 1062$): audio-only (A) and audiovisual (AV). Acoustic cues were intensity and duration (C1), f_0 and duration (C2), and only duration (C3). All three acoustic cues of prominence were available in the control condition (C0). Total words for marking added up to $N = 8496$.

The total 2124 words for marking were teased apart to better understand the distribution of prominence marks per condition in both modalities: the audio-only and audiovisual modalities, each with a total of 1062 words available for marking. The audiovisual modality consistently showed a higher number of words marked as prominent. The difference between modalities was highest precisely in C3, with a relative number of prominence marks of 27.12%, increasing by 1.98% respect to the audio-only modality in the same condition. The largest difference between modalities (3.6%) was found in C0. On the contrary, modalities hardly differed in both C1 and C2.

Overall, all experimental conditions showed a higher number of marks of prominence, and the highest difference with C0 was found in C3 (5.7%); however, when compared to C1 and C3, the difference was smaller (0.75% and 1.13%, respectively).

4.3.1.4 Prominence and visual cues

In the case of gestures, sentences were annotated with ELAN (Brugman & Russel, 2004), as discussed earlier (§ 3.3.5). Out of the 170 annotated gestures of the 30 target sentences (see Appendix A4 for details), those performed uniquely with hands were more numerous (40 – 23.5% of occurrences) than those performed with either head (17 – 10%), or eyebrows (4 – 2.3%). However, the highest number of gestures were those performed simultaneously by the joint movement of hands and head (87 – 51.1% of the occurrences). Conversely, the minimum number of occurrences corresponded to eyebrows and head (2 – 1.1%) and hands and eyebrows (1 – 0.6%). Gestures performed simultaneously with hands, eyebrows, and head were (19 – 11.1%) (Figure 33).

Experiment I

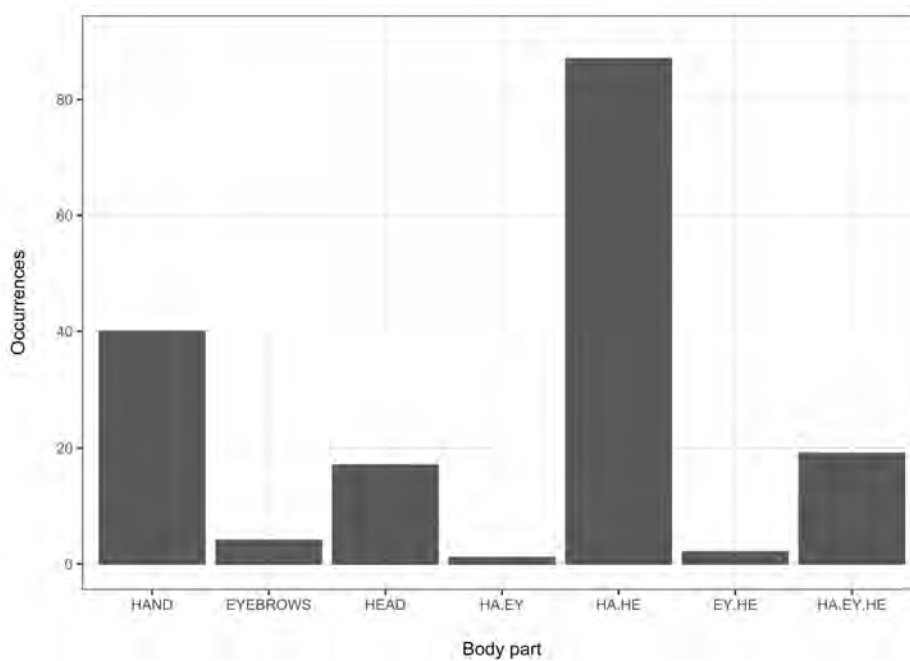


Figure 33: Occurrences of gestures according to the body part involved in their production, separately: hand, eyebrows, head; and combined: hand and eyebrows (HA.EY), hand and head (HA.HE), eyebrows and head (EY.HE), and hand, eyebrows, and head (HA.EY.HE).

It was also observed that gesture phases tended to be synchronised for the different articulators involved in the production of gestures. Thus, the apex of a simultaneous gesture of hand and head, for example, tended to coincide for both gesture articulators (Figure 34).

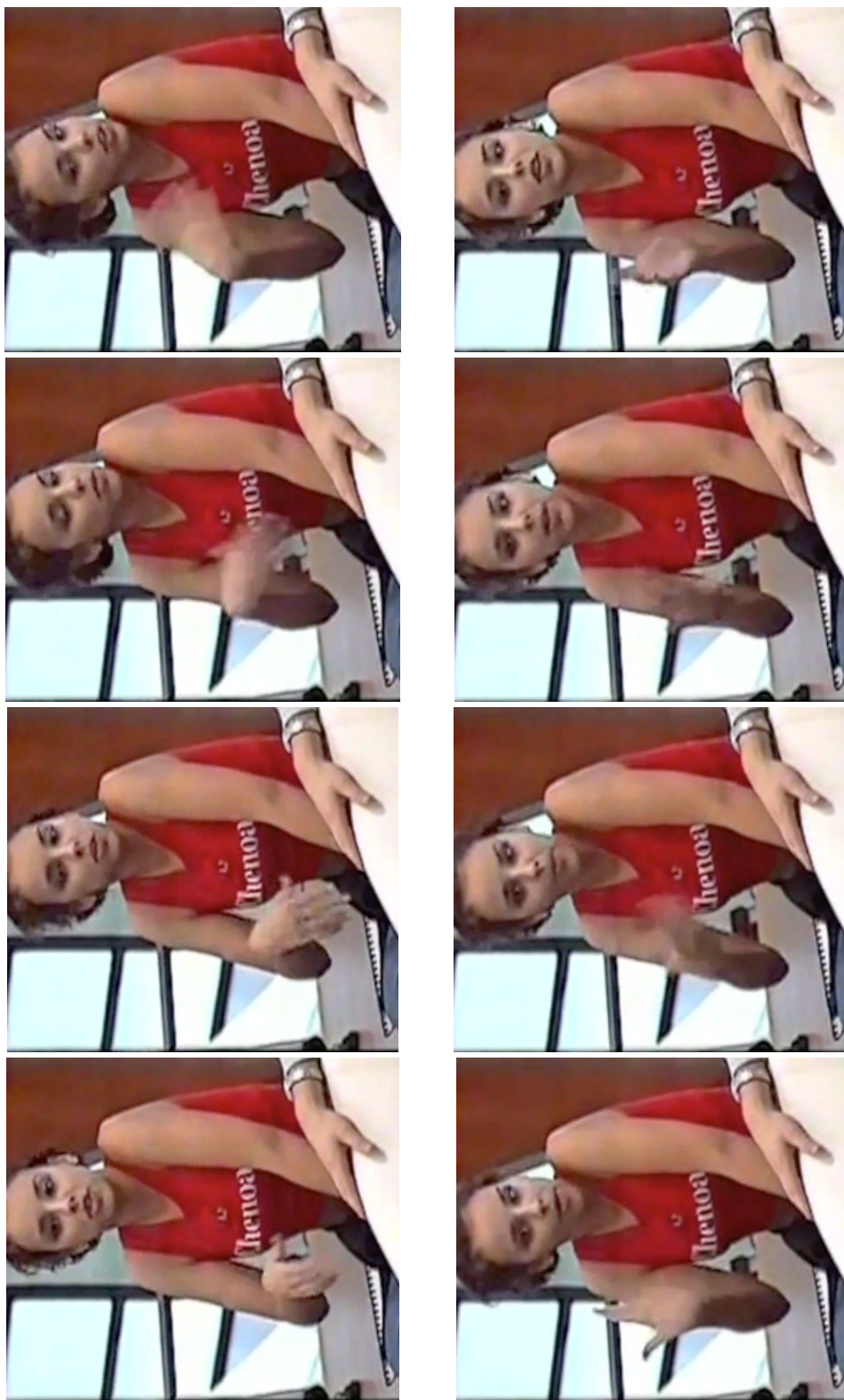


Figure 34: Sample of gesture sequence performed simultaneously as a hand and head movement.

Experiment I

As it can be seen in Table 17, among phases coinciding with words given a mark of prominence, apexes consistently obtained the highest number of prominence marks in C3, with a maximum 128 occurrences out of 270 (47.4%).

Gesture phase	Words for marking per condition and modality	Prominent words (%)					
		C0			C1		
		A	AV	AV ($\pm\%$)	A	AV	AV ($\pm\%$)
Preparation	258	19 (7.3)	25 (9.6)	+2.3	41 (15.8)	48 (18.6)	+2.8
Stroke	174	6 (3.4)	17 (9.7)	+6.3	13 (7.5)	20 (11.5)	+4.0
Apex	270	109 (40.4)	126 (46.6)	+6.2	113 (41.9)	110 (40.7)	-1.2
Hold	164	29 (17.7)	28 (17.0)	-0.7	54 (32.9)	45 (27.4)	-5.5
Recoil	78	19 (24.4)	19 (24.3)	-0.1	23 (29.5)	24 (32.1)	+2.6
Retraction	118	16 (13.6)	21 (17.8)	+4.2	25 (21.2)	23 (19.5)	-1.7
Total	1062	198 (18.6)	236 (22.2)	+ 3.6	269 (25.3)	270 (25.4)	+0.1

(a) Prominent words according to gesture phase for C0 and C1

Gesture phase	Words for marking per condition and modality	Prominent words (%)					
		C2			C3		
		A	AV	AV ($\pm\%$)	A	AV	AV ($\pm\%$)
Preparation	258	45 (17.4)	36 (13.9)	-3.5	40 (15.5)	42 (16.2)	+0.7
Stroke	174	18 (10.3)	31 (17.8)	+7.5	12 (6.9)	26 (14.9)	+8.0
Apex	270	114 (42.2)	122 (45.2)	-3.0	126 (46.7)	128 (47.4)	+0.7
Hold	164	45 (27.4)	46 (28.0)	+0.6	38 (23.2)	43 (26.2)	+3.0
Recoil	78	19 (24.4)	20 (25.6)	+1.2	23 (29.5)	25 (32.1)	+2.6
Retraction	118	18 (15.3)	16 (13.6)	-1.7	28 (23.7)	24 (20.3)	-3.4
Total	1062	259 (24.4)	271 (25.5)	+1.1	267 (25.1)	288 (27.1)	+2.0

(b) Prominent words according to gesture phase (cont.) for C2 and C3

Table 17: Different gesture phases coinciding with words marked as prominent per condition and modality ($n = 1062$; $N = 8496$). Values for the audio-only modality, where no visual information was available, served to compare the marks given by participants in the audiovisual modality. This difference is expressed as AV ($\pm\%$). Acoustic cues were intensity and duration (C1), f_0 and duration (C2), and only duration (C3). All three acoustic cues of prominence were available in the control condition (C0).

Conversely, strokes had the lowest percentage of prominence marks in C0, with a minimum of 6 out of 174 (3.4%). However, words accompanied by strokes increased across conditions respect to C0, as they also increased consistently in the audiovisual respect to the audio-only modality, with a maximum difference between modalities of 8% in C3. Interestingly, this difference is the largest for any gesture phases across conditions between both modalities. It is also worth mentioning that those words co-occurring with a recoil phase received numerous prominence marks, with a maximum of 25 out of 78 (32.1%) in the audiovisual modality of C3. In the case of the hold phase, it received fewer marks in the audiovisual modality in C0 (-0.7%) and C1 (-5.5%), and slightly more in the audiovisual modality of C2 (+0.6%) and C3 (+3.0%). In total, the number of words given a mark of prominence in the audiovisual modality increased in all conditions, with a minimum difference of 0.1% in C1, and maxima of 3.6% and 2.0% in C0 and C3, respectively.

4.3.2 Inferential statistics

4.3.2.1 Correlation

Firstly, a correlation test was conducted in relation to hypothesis 1 in order to assess whether the difficulty of the task increased with sentence length so as to make impossible for participants to mark prominence above chance, especially in longer sentences given the variability of to the spontaneous speech stimuli and the manipulations on the acoustic correlates of prominence. The correlation showed a significant positive relationship between words considered prominent and the number of words per sentence $r = .71, p < .01$ (Figure 35).

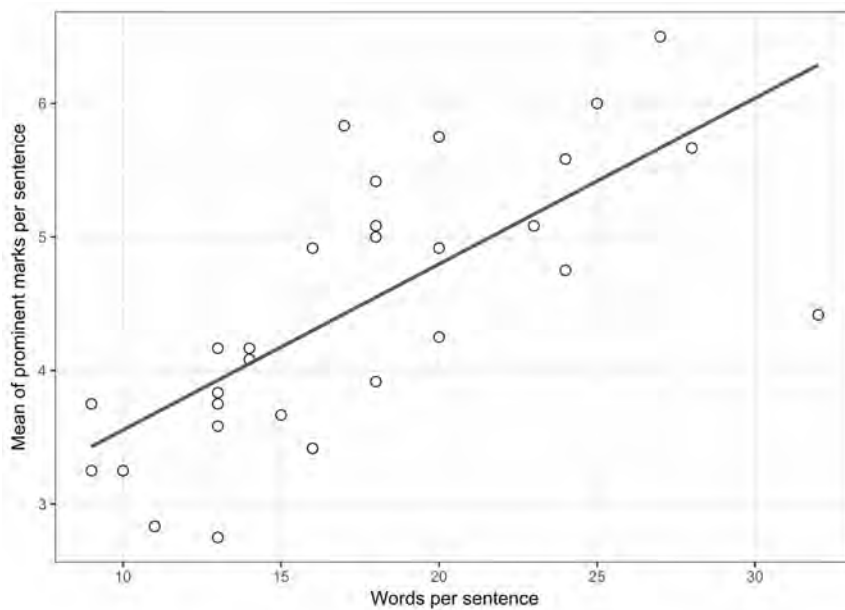


Figure 35: Correlation between sentence length and marks of prominence given by participants.

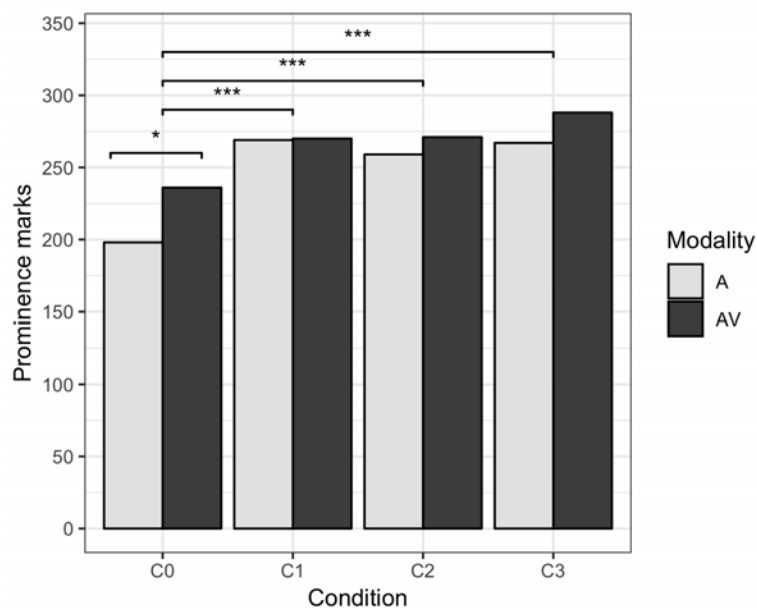
4.3.2.2 Number of prominence marks

Furthermore, the differences observed in the number of marks of prominence given per condition and modality were assessed through chi-square tests. The

Experiment I

results revealed that the higher number of marks given in the audiovisual modality were significantly different from those of the audio-only modality in C0, $\chi^2(1) = 3.96, p < .05$ (Figure 36a). When the number of marks were compared among conditions, all experimental conditions contrasted with C0, and significant differences were found between C0 and C1, $\chi^2(1) = 14.41, p < .001$; between C0 and C2, $\chi^2(1) = 12.11, p < .001$; and between C0 and C3, $\chi^2(1) = 18.79, p < .001$. No significant differences were found among experimental conditions.

Additionally, these differences were analysed per modality: in the audio-only modality the control condition contrasted significantly with the higher number of prominence marks given in C1, $\chi^2(1) = 13.45, p < .001$; in C2, $\chi^2(1) = 10.03, p < .01$; and C3, $\chi^2(1) = 12.73, p < .001$; while no differences were found among experimental conditions (Figure 36b). Conversely, in the audiovisual modality, differences were found only between the control condition and C3, $\chi^2(1) = 6.58, p < .01$ (Figure 36c).



(a) All conditions in both modalities

Experiment I

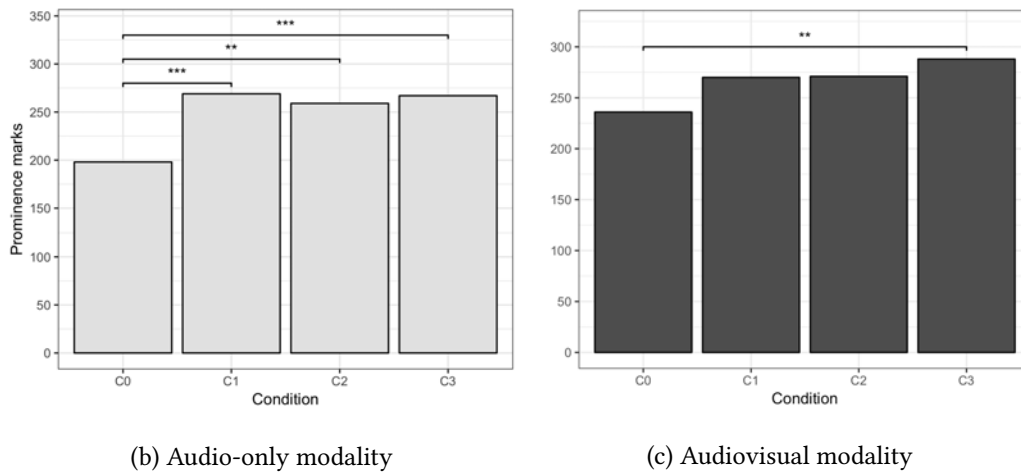
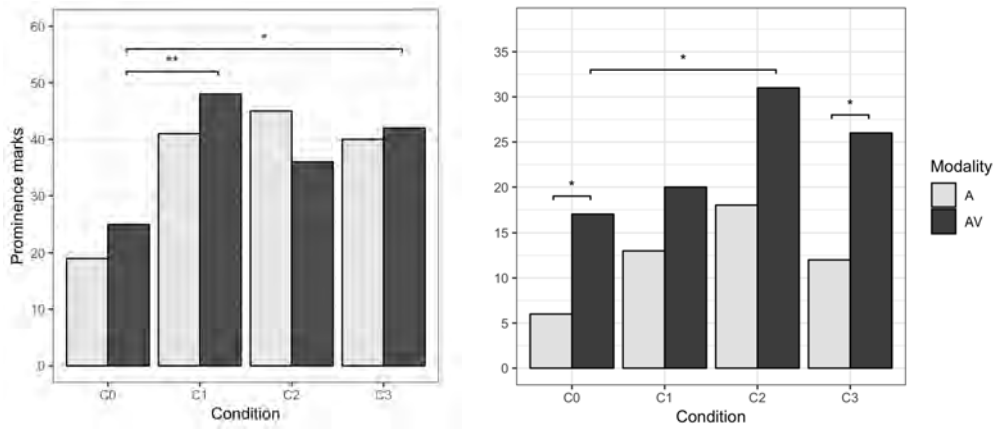


Figure 36: Significant differences in the number of prominence marks per modality and condition. Acoustic cues were intensity and duration (C1), f_0 and duration (C2), and only duration (C3). All three acoustic cues of prominence were available in the control condition (C0).

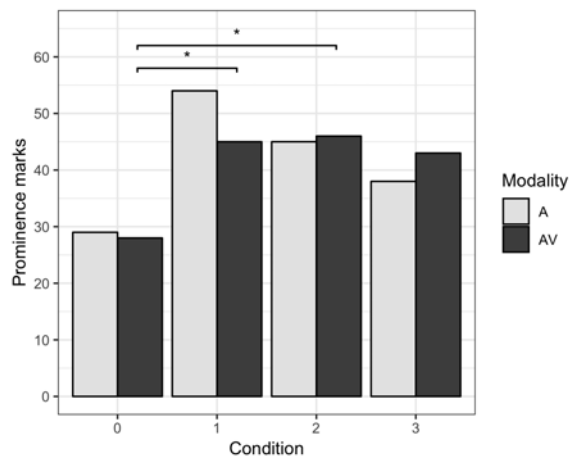
As for gesture phases, differences between the audio-only and the audiovisual modality were found only for words considered prominent that were accompanied by a stroke in C0, $\chi^2(1) = 4.65, p < .05$ and C3, $\chi^2(1) = 4.99, p < .05$ (Figure 37b). Additionally, differences for all gesture phases were analysed only in the audiovisual modality among conditions. In the case of the preparation phase, differences were observed between C0 and C1, $\chi^2(1) = 7.72, p < .01$; and between C0 and C3, $\chi^2(1) = 4.08, p < .05$ (Figure 37a). For strokes, differences were only found between C0 and C2, $\chi^2(1) = 9.68, p < .01$ (Figure 37b). For holds, words given a mark of prominence differed significantly between C0 and C1, $\chi^2(1) = 4.51, p < .05$; and between C0 and C2, $\chi^2(1) = 5.04, p < .05$ (Figure 37c). No differences were found between modalities nor conditions for the gesture phases of apex, recoil, and retraction.

Experiment I



(a) Preparation

(b) Stroke



(c) Hold

Figure 37: Graphs showing significant differences in the number of prominence marks with different gesture phases per modality and per condition in the audiovisual modality. No differences were found for the gesture phases of apex, recoil, and retraction. Acoustic cues were intensity and duration (C1), f_0 and duration (C2), and only duration (C3). All three acoustic cues of prominence were available in the control condition (C0).

4.3.2.3 Model building and selection

In order to determine which variables influenced participants' marks for binary prominence at word level, different generalised linear mixed models (GLMMs) with a logit link function were estimated for a binomial distribution (see Jaeger, 2008, for details). As previously seen (§ 3.2.2), GLMMs are an extension of linear mixed models (LMMs) for non-normal distributions, and both have been previously applied in linguistic research (e.g. Adamou et al., 2018; Gries, 2015; Masson-Carro et al., 2017; Quené, 2008). Mixed-models are considered more robust statistical tools than conventional analysis of variance (ANOVAs) (Jaeger, 2008), especially in cases of repeated measures, unbalanced designs, nested data, or categorical variables (see Singmann & Kellen, *in press*, for a review).

One important advantage of mixed-models is that they allow to include random variables capturing the variability inherent to participants and stimuli (Baayen et al., 2008). In this pilot study, this was done using the *lme4* package (Bates et al., 2015b) in R (2018). All models were built using maximum likelihood (Laplace Approximation) and optimised with *BOBYQA* (Powell, 2009) to increase iterations and avoid convergence errors, in a process explained earlier (Barr et al., 2013) (§ 3.2.1).

In this pilot study, the binary dependent variable *prominence* was modelled through categorical independent variables: *speaker* ($n = 12$), *sentence* ($n = 30$), *word* ($n = 531$), *modality* ($n = 2$), and *condition* ($n = 4$). Also, as categorical independent variables, each gesture phase was introduced separately, so as to indicate their co-occurrence with the words being marked by participants: *preparation* ($n = 2$), *stroke* ($n = 2$), *apex* ($n = 2$), *hold* ($n = 2$), *recoil* ($n = 2$), and *retraction* ($n = 2$). Finally, as independent continuous variables, *fundamental frequency*, *intensity*, and *duration* were fed into the model, which corresponded to the acoustic values of the 531 words available for marking, as explained above (see Table 13 in §

3.3.6). The different gesture phases were introduced as separate variables using dummy coding, rather than as different levels of one single variable. This was done so partly due to the absence of a reliable baseline against which to compare each phase, since very few words in the corpus were uttered in the absence of a gesture. (As mentioned earlier § 3.3.2, all full meaningful sentences uttered by speakers in the corpus were accompanied by an intertwined succession of gestures). On the other hand, by introducing each gesture phase as a separable variable, it was possible to compare the effect that the presence of each of the phases had in relation to their absence.

In addition, several adjustments were conducted to both control for confounding variables and avoid convergence errors in the estimation of parameters due to the different scales of the acoustic correlates of prominence. The continuous variables of f_0 , intensity, and duration were fed into the model as z-scores, a scale that keeps the same relationship of each data point respect to their mean. The standardisation for each acoustic correlate was conducted according to the specific sentence in which they occurred, especially since words are assumed to be perceived as prominent relative to the environment in which they were uttered. Thus, intensity and duration were standardised per sentence, as well as f_0 . Additionally, in this case of f_0 , pitch bias resulting from speaker gender was eliminated through standardisation based on the gender of the speaker (see Appendix A1, A2, and A3 for details of their respective distributions).

Model building proceeded by specifying fixed effects in model 1 (*M1*) with the predictors: *modality*, *condition*, *preparation*, *stroke*, *apex*, *hold*, *recoil*, *ff* (*fundamental frequency*), *intensity*, and *duration*. By-subject random effects were declared through the variable *participant*, and by-item random effects were specified through the nested variables of *word* within *sentence* within *speaker*, as previously discussed (§ 3.2.1, Figure 26c). Using the formula notation of the *lme4*

package developed for R by Bates et al (2015b):

```
prominence ~ modality + condition + preparation + stroke + apex + hold + recoil + retract +  
+ z_intensity + z_ff + z_duration + (1|participant) + (1|speaker/sentence/word)
```

In this model, *M1*, intercepts in random effects are signalled through the notation (*1* / ...). In the case of by-subject and by-item random effects, this becomes (*1* | *participant*) and (*1* | *speaker/sentence/word*), respectively.

The set of models was then built using the Akaike Information Criterion (*AIC*) (as explained in § 3.2.4) using R (2018) and compared through the package *AICcmodavg* (Mazerolle, 2017). The Akaike Information Criterion (*AIC*) is an ordinal score assessing the quality of the model with no meaning of its own except as a way to rank models (Akaike, 1973; see Burnham & Anderson, 2002, for details). From models *M1* to *M5*, different non-significant predictors were progressively removed. Model *M6* included a first-order interaction of *modality* with all the predictors of all gesture phases and all acoustic correlates:

```
prominence ~ modality * (condition + preparation + stroke + apex + hold + recoil + retract +  
+ z_intensity + z_ff + z_duration) + (1|participant) + (1|speaker/sentence/word)
```

The resulting *AIC* value in *M6*, which included a first-order interaction, decreased by more than 30 points (Table 19). Subsequently, from models *M6* to *M9*, different non-significant predictors were removed from the interaction. The lowest *AIC* value (*AIC* = 6776.1) was achieved by model *M9*. In this model, the predictors *preparation* and *z-intensity* proved non-significant and were omitted. As an alternative to *M6*, a second interaction was declared in *M8* between *condition* and the continuous variables of *z-fundamental frequency*, *z-intensity*, and *z-duration*, but the resulting *AIC* value (*AIC* = 6789.05) did not improve the fit of model *M6*:

```
prominence ~ modality * (condition + preparation + stroke + apex + hold + recoil + retract + z_ff + z_duration) +  
+ condition * (z_intensity + z_ff + z_duration) + (1|participant) + (1|speaker/sentence/word)
```


Subsequently, the *AIC* value of *M9* became still lower when *speaker* was omitted from the nested by-item random effects in *M10* (*AIC* = 6774.1) (Table 18).

Random effects	Variance	Std. Deviation	<i>AIC</i>
Model 9 (<i>M9</i>)			6776.1
participant	0.313	0.560	
word/sentence/speaker	2.385	1.544	
sentence/speaker	0.003	0.059	
speaker	0.002	0.046	
Model 10 (<i>M10</i>)			6774.1
participant	0.313	0.560	
word/sentence	2.385	1.544	
sentence	0.005	0.071	

Table 18: Intercept random effects for *M9* and *M10*, showing the decrease of the *AIC* value when the redundant upper level factor *speaker* was omitted from the nested random effects is omitted.

The variance of by-item random effect *M9* was almost entirely captured by the subsequent two levels of *sentence* and *word*. Its omission resulted in a decrease of precisely 2 *AIC* points, as mentioned by Burnham and Anderson (2002, p. 454), equivalent to one estimable parameter (§ 3.2.4). This was important for the later declaration of random slopes (Table 19).

In the subsequent models, from *M11* to *M25*, different random structures were explored. Logically, variability in participants' responses captured by the intercepts of by-subject random effects—expressed by the number one in (*1* | *participants*)—may also vary in slope. Different slopes were expected to result from the levels of *modality*, since each participant rated stimuli in both modalities⁴. Similarly, items, i.e. the words marked by participants, declared as by-item nested

⁴ In the case of the variability associated to the levels of *condition*, this is more problematic, since the unbalanced design of the experiment made that the two trained listeners in the control condition marked words only in one of the three manipulated conditions. Thus, *AIC* values with maximal by-subject random effects (i.e. *condition* and *modality*) did not reduce the Δ_i difference respect to the AIC_{\min} .

random effects ($1 | speaker/sentence/word$), were expected to vary among conditions and modalities.

Additionally, the difference observed for models $M9$ and $M10$ expressed as Δ_i was less than 2 *AIC* points (see § 3.2.4, Table 11). This fact was explored in all subsequent models, $M11$ to $M25$, which were fitted alternatively with and without the upper level of the nested random effects *speaker*, as it can be seen in models $M11$ and $M12$, $M13$ and $M14$, etc. Actually, in these two pairs of models, when trying to declare slopes in by-subjects random effects, the removal of *speaker* in by-items random effects resulted in a non-convergence error. A convergence problem also appeared for a similar reason in models $M24^*$ and $M25^{*5}$. Conversely, in the cases where *speaker* was declared, the goodness-of-fit of the model worsened considerably, as in model $M17$ respect to model $M18$ (Table 19).

As it is apparent from Table 19, the minimal adequate model in Burnham and Anderson's (2002) terms was $M18$, with an *AIC* value of 6697.18. This model, when compared to the rest of the fitted models of the set, had the highest Akaike weight $w_i = 0.51$, closely followed by model $M19$ ($w_i = 0.37$), and $M20$ ($w_i = 0.08$).

On the one hand, the difference between $M18$ and $M19$ was due to the removal of the non-significant predictor *gesture-recoil* in $M19$, but this did not lead to a lower *AIC* value ($AIC = 6697.83$) in $M19$. On the other hand, the *AIC* difference between both $M18$ and $M20$, which corresponded to 3.68 points, was due to their different by-subject structures. Although slopes for the random effects of *modality* were specified for participants in $M20$, the intercept by-subject random effects of $M18$ —as signalled by ($1 | participants$)—resulted in a lower *AIC* value. In addition, the single predictors of *modality* and *condition* in the random structure of by-item random effects of model $M18$ were extended to be included in the structure of by-subject random effects, but this random slope model, $M18'$ failed

⁵ Models labelled with an asterisk failed to converge.

Model	Prominence as a function of (-) fixed effects, with by-subject and by-item random effects	LogLikelihood	K	AIC	Δ_i	w_i
M1	modality + condition + prep + stroke + apex + hold + recoil + retraction + z-ff + z-intensity + z-duration + (1 participant) + (1 speaker/sentence/word)	-3384.70	18	6805.40	108.22	0.00
M2	modality + prep + stroke + apex + hold + recoil + retraction + z-ff + z-intensity + z-duration + (1 participant) + (1 speaker/sentence/word)	-3391.77	15	6813.53	116.35	0.00
M3	condition + prep + stroke + apex + hold + recoil + retraction + z-ff + z-intensity + z-duration + (1 participant) + (1 speaker/sentence/word)	-3387.41	17	6808.82	111.65	0.00
M4	prep + stroke + apex + hold + recoil + retraction + z-ff + z-intensity + z-duration + (1 participant) + (1 speaker/sentence/word)	-3395.83	11	6813.67	116.49	0.00
M5	apex + hold + recoil + z-ff + z-duration + (1 participant) + (1 speaker/sentence/word)	-3396.44	10	6812.87	115.69	0.00
M6	modality x (condition + prep + stroke + apex + hold + recoil + z-ff + z-intensity + z-duration) + (1 participant) + (1 speaker/sentence/word)	-3362.73	28	6781.47	84.29	0.00
M7	modality x (condition + stroke + apex + hold + recoil + z-ff + z-intensity + z-duration) + (1 participant) + (1 speaker/sentence/word)	-3362.83	26	6777.66	80.48	0.00
M8	modality x (condition + stroke + apex + hold + recoil + z-ff + z-duration) + condition x (z-ff + z-intensity + z-duration) + (1 participant) + (1 speaker/sentence/word)	-3360.52	34	6789.05	91.87	0.00
M9	modality x (condition + stroke + apex + hold + recoil + z-ff + z-duration) + (1 participant) + (1 speaker/sentence/word)	-3364.04	24	6776.09	78.91	0.00
M10	modality x (condition + stroke + apex + hold + recoil + z-ff + z-duration) + (1 participant) + (1 sentence/word)	-3364.05	23	6774.10	76.92	0.00

(a) Summary of AIC results for the GLMM random-intercept models.

Model	Prominence as a function of (-) fixed effects with by-subject and by-item random effects	LogLikelihood	K	AIC	Δ_i	w_i
M11	modality x (condition + stroke + apex + hold + recoil + z-ff + z-duration) + (condition participant) + (1 speaker/sentence/word)	-3359.69	33	6785.37	88.20	0.00
M12*	modality x (condition + stroke + apex + hold + recoil + z-ff + z-duration) + (condition participant) + (1 sentence/word)	-3359.69	32	6783.37	86.20	0.00
M13	modality x (condition + stroke + apex + hold + recoil + z-ff + z-duration) + (modality + condition participant) + (1 speaker/sentence/word)	-3359.39	38	6794.78	97.60	0.00
M14*	modality x (condition + stroke + apex + hold + recoil + z-ff + z-duration) + (modality + condition participant) + (1 sentence/word)	-3359.39	37	6792.78	95.60	0.00
M15	modality x (condition + stroke + apex + hold + recoil + z-ff + z-duration) + (condition participant) + (condition sentence/word)	-3307.82	50	6715.63	18.45	0.00
M16	modality x (condition + stroke + apex + hold + recoil + z-ff + z-duration) + (condition participant) + (condition speaker/sentence/word)	-3307.69	60	6735.38	38.20	0.00
M17	modality x (condition + stroke + apex + hold + recoil + z-ff + z-duration) + (1 participant) + (modality + condition speaker/sentence/word)	-3296.42	66	6724.84	27.66	0.00
M18	modality x (condition + stroke + apex + hold + recoil + z-ff + z-duration) + (1 participant) + (modality + condition sentence/word)	-3297.59	51	6697.18	0.00	0.51
M19	modality x (condition + stroke + apex + hold + z-ff + z-duration) + (1 participant) + (modality + condition sentence/word)	-3229.91	49	6697.83	0.65	0.37
M20	modality x (condition + stroke + apex + hold + recoil + z-ff + z-duration) + (modality participant) + (modality + condition sentence/word)	-3297.43	53	6700.86	3.68	0.08
M21	modality x (condition + stroke + apex + hold + recoil + z-ff + z-duration) + (modality participant) + (modality + condition speaker/sentence/word)	-3296.22	68	6728.43	31.25	0.00
M22	modality x (condition + stroke + apex + hold + recoil + z-ff + z-duration) + (condition participant) + (modality + condition speaker/sentence/word)	-3290.23	75	6730.45	33.27	0.00
M23	modality x (condition + stroke + apex + hold + recoil + z-ff + z-duration) + (condition participant) + (modality + condition sentence/word)	-3291.38	60	6702.76	5.58	0.03
M24*	modality x (condition + stroke + apex + hold + recoil + z-ff + z-duration) + (modality + condition participant) + (modality + condition speaker/sentence/word)	-3290.11	80	6740.23	43.05	0.00
M25*	modality x (condition + stroke + apex + hold + recoil + z-ff + z-duration) + (modality + condition participant) + (modality + condition sentence/word)	-3295.51	65	6721.02	23.84	0.00

(b) Summary of AIC results for the GLMM random-slope models.

Table 19: Summary of AIC results for the GLMMs modelling the marking of prominence by participants as a function of predictors declared as fixed effects. All models include by-subject and by-item random effects. Table (a) shows random-intercept models; Table (b) shows different structures in random effects. K indicates the estimated parameters in the model. The statistics associated to each model are Akaike Information Criterion (AIC), increase (Δ_i) of each model respect to the minimum AIC value, the log-likelihood of the model, and the Akaike weight (w_i) for each candidate model to be the minimal adequate model. Models labelled with an asterisk failed to converge.

to converge and was not included in the set of models.

Following Arnold's (2010) guidelines, both models *M18* and *M19* are reported, which differed in 2 estimated parameters that corresponded to the interaction of predictor *recoil*. Firstly the results of the top-ranked model *M18* ($AIC = 6697.18$) are offered; then the β -values for *M19* ($AIC = 6697.83$) are displayed for comparison, showing that the difference in their AIC value did not result in a difference in the estimates for the predictors included in *M18*.

4.3.2.4 Details of minimal adequate model *M18*

The minimal adequate model, *M18*, did not show overdispersion ($\Phi_{\text{Pearson}} = 0.55$, $p > .05$; see Appendix A4 for details)⁶. It revealed a strong effect of the acoustic correlates of f_0 ($\beta = 0.46$, $SE = 0.07$, $z = 2.27$, $p < .001$, $OR = 1.59$) and syllable duration ($\beta = 0.98$, $SE = 0.09$, $z = 10.9$, $p < .001$, $OR = 2.68$) in the perception of prominence for all three conditions as well as for the marks of two trained listeners in the control condition (Figure 38, Table 20).

Following the standard procedure, the effect size of these variables is expressed as the odds ratio (OR) coefficient. Thus, as standardized variables, 1 standard deviation increase in the continuous variables of f_0 ($SD = 23.85$ Hz) and in duration ($SD = 0.055$ s) made words 1.59 and 2.68 times more likely, respectively, to be considered prominent.

When compared to the control group, all experimental conditions showed a significant increase in the probability of marking words as prominent. A word marked in C1, where f_0 had been neutralised, was 6.30 times more likely to be given a mark of prominence ($\beta = 1.84$, $SE = 0.52$, $z = 3.49$, $p < .001$), although

⁶ The occurrence of more variance in the data than predicted by a statistical model, a phenomenon known as overdispersion, is reported here and in the subsequent fitted models that are discussed later.

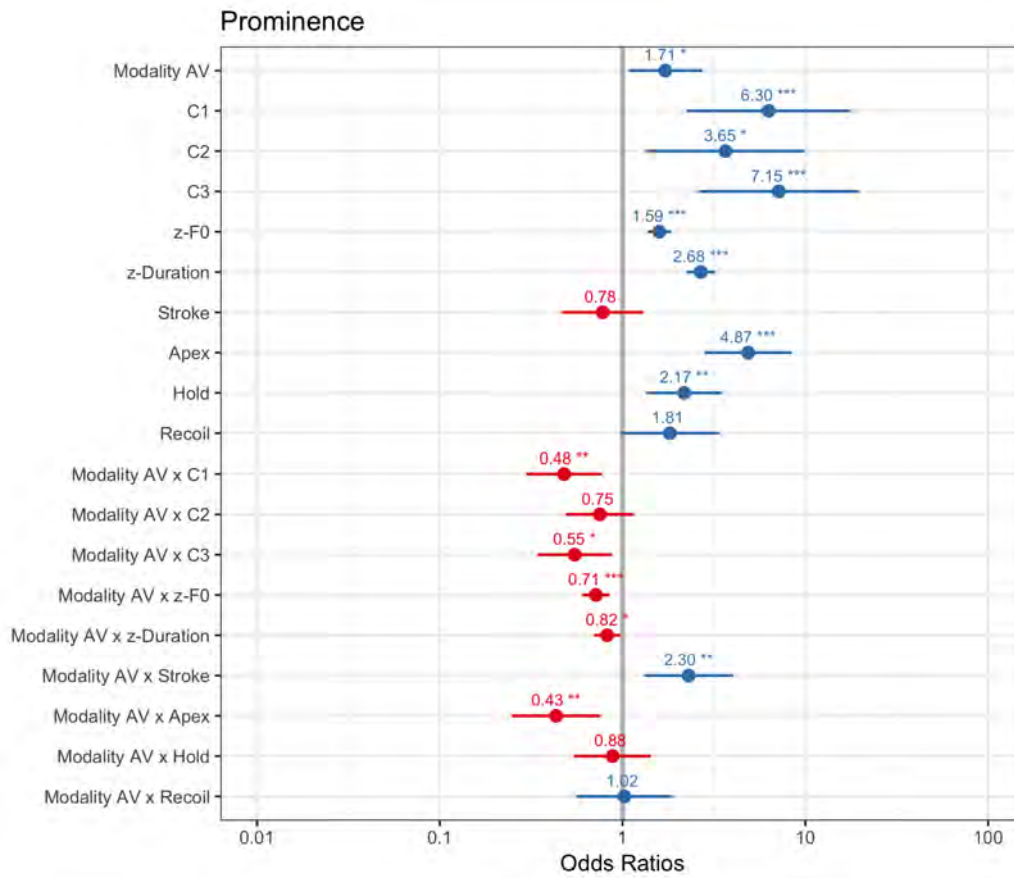


Figure 38: Forest plot showing odds ratios for main effects and interactions predicting prominence in M18 (AIC = 6697.18). For OR < 1, effect size equals 1/OR. Error bars are 95% CI.

odds reached a maximum in C3, so that words were 7.15 times more likely to be considered prominent in this condition ($\beta = 1.96$, $SE = 0.51$, $z = 3.81$, $p < .001$), where both f_0 and intensity had been neutralised and duration served as the only acoustic cue of prominence. In the case of C2, the condition where intensity had been neutralised, and both f_0 and duration were available to listeners, the effect size was slightly less strong, and odds decreased to 3.65 ($\beta = 1.29$, $SE = 0.50$, $z = 2.54$, $p < .05$) (Figure 39).

Also, marking a word as prominent coincided significantly with the occurrence of the apex phase ($\beta = 1.58$ $SE = 0.27$, $z = 5.67$, $p < .001$) and the hold phase of

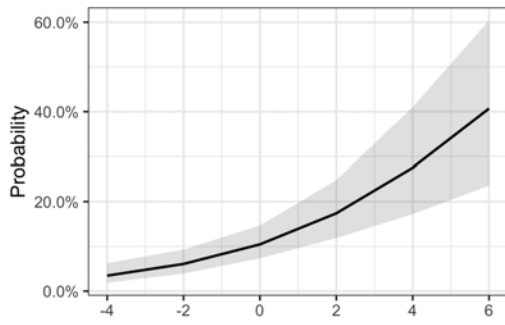
Experiment I

gestures ($\beta = 0.77$, $SE = 0.24$, $z = 3.20$, $p < .001$). Interestingly, the odds for marking words as prominent were 1.71 higher in the audiovisual than in the audio-only modality ($\beta = 0.53$, $SE = 0.23$, $z = 2.27$, $p < .05$), although this effect was nuanced for all three experimental conditions, which seemed to be more conservative in marking prominent words in the audiovisual modality.

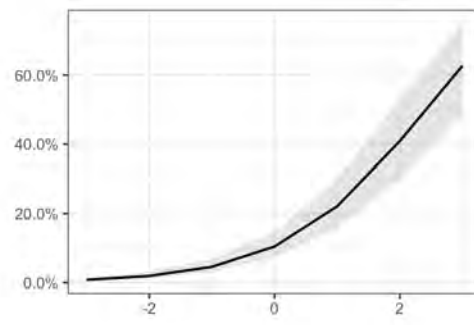
Predictor	β (SE)	p	95% CI for odds ratio		
			Lower limit	Odds ratio	Upper limit
Intercept	-4.00 (0.50)	< .001***	0.01	0.02	0.05
Modality (0=A, 1=AV)	0.53 (0.23)	< .05*	1.08	1.71	2.73
Condition 1	1.84 (0.52)	< .001***	2.24	6.30	17.70
Condition 2	1.29 (0.50)	< .05*	1.35	3.65	9.91
Condition 3	1.96 (0.51)	< .001***	2.60	7.15	19.67
z-F0 (standardized f_0)	0.46 (0.07)	< .001***	1.37	1.59	1.84
z-Duration (standardized duration)	0.98 (0.09)	< .001***	2.25	2.68	3.20
Stroke (0=no, 1=yes)	-0.25 (0.26)	.33	0.47	0.78	1.30
Apex (0=no, 1=yes)	1.58 (0.28)	< .001***	2.82	4.87	8.42
Hold (0=no, 1=yes)	0.77 (0.24)	< .01**	1.35	2.17	3.50
Recoil (0=no, 1=yes)	0.59 (0.31)	.06	0.97	1.81	3.38
Modality AV x Condition 1	-0.73 (0.24)	< .01**	0.30	0.48	0.77
Modality AV x Condition 2	-0.28 (0.22)	.19	0.49	0.75	1.16
Modality AV x Condition 3	-0.60 (0.23)	< .05*	0.34	0.55	0.87
Modality AV x z-F0	-0.33 (0.08)	< .001***	0.60	0.71	0.85
Modality AV x z-Duration	-0.19 (0.08)	< .05*	0.70	0.82	0.97
Modality AV x Stroke	0.83 (0.28)	< .01**	1.31	2.30	4.03
Modality AV x Apex	-0.83 (0.28)	< .01**	0.25	0.43	0.76
Modality AV x Hold	-0.12 (0.24)	.61	0.54	0.88	1.43
Modality AV x Recoil	0.02 (0.31)	.93	0.56	1.02	1.88

Table 20: Results of fixed effects in model M18 ($AIC = 6697.18$) predicting the marking of prominence from the variables modality, condition, fundamental frequency, duration, gesture-stroke, gesture-apex, gesture-hold, and gesture-recoil.

Experiment I



(a) $z-f_0$



(b) z -Duration

Experiment I

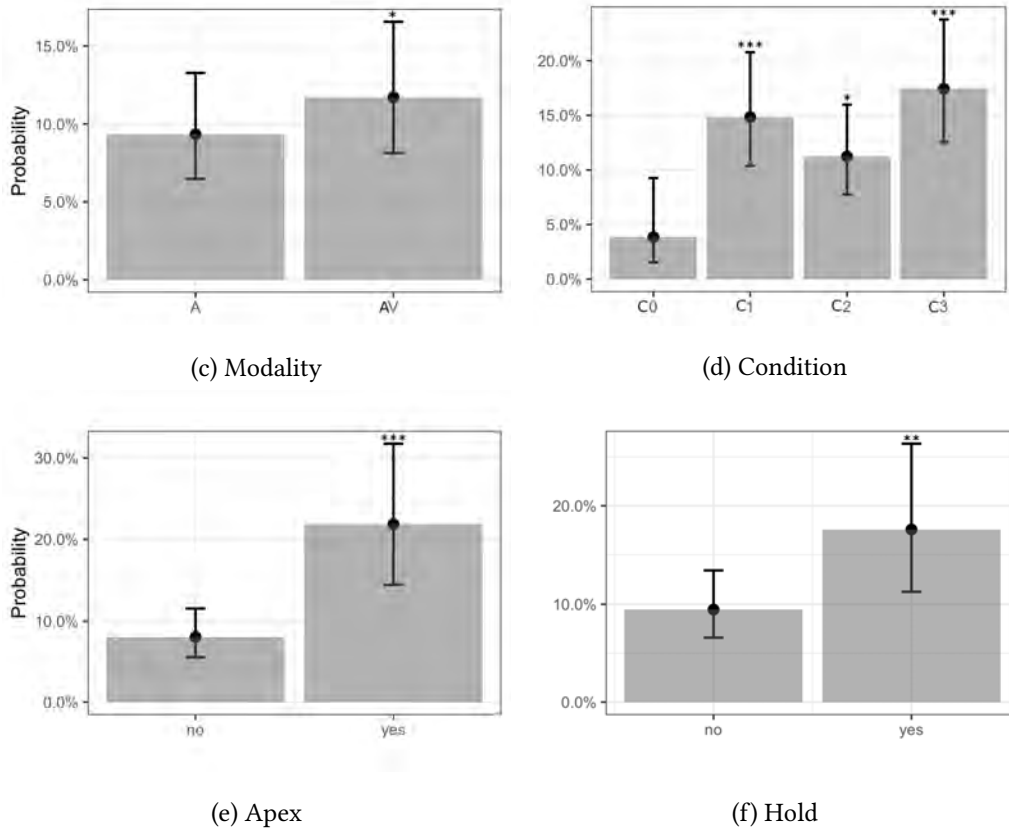


Figure 39: Main effects of M18 ($AIC = 6697.18$) for continuous variables: (a) fundamental frequency, and (b) duration; and for categorical variables: (c) condition and (d) modality. Shading and error bars are 95% CI.

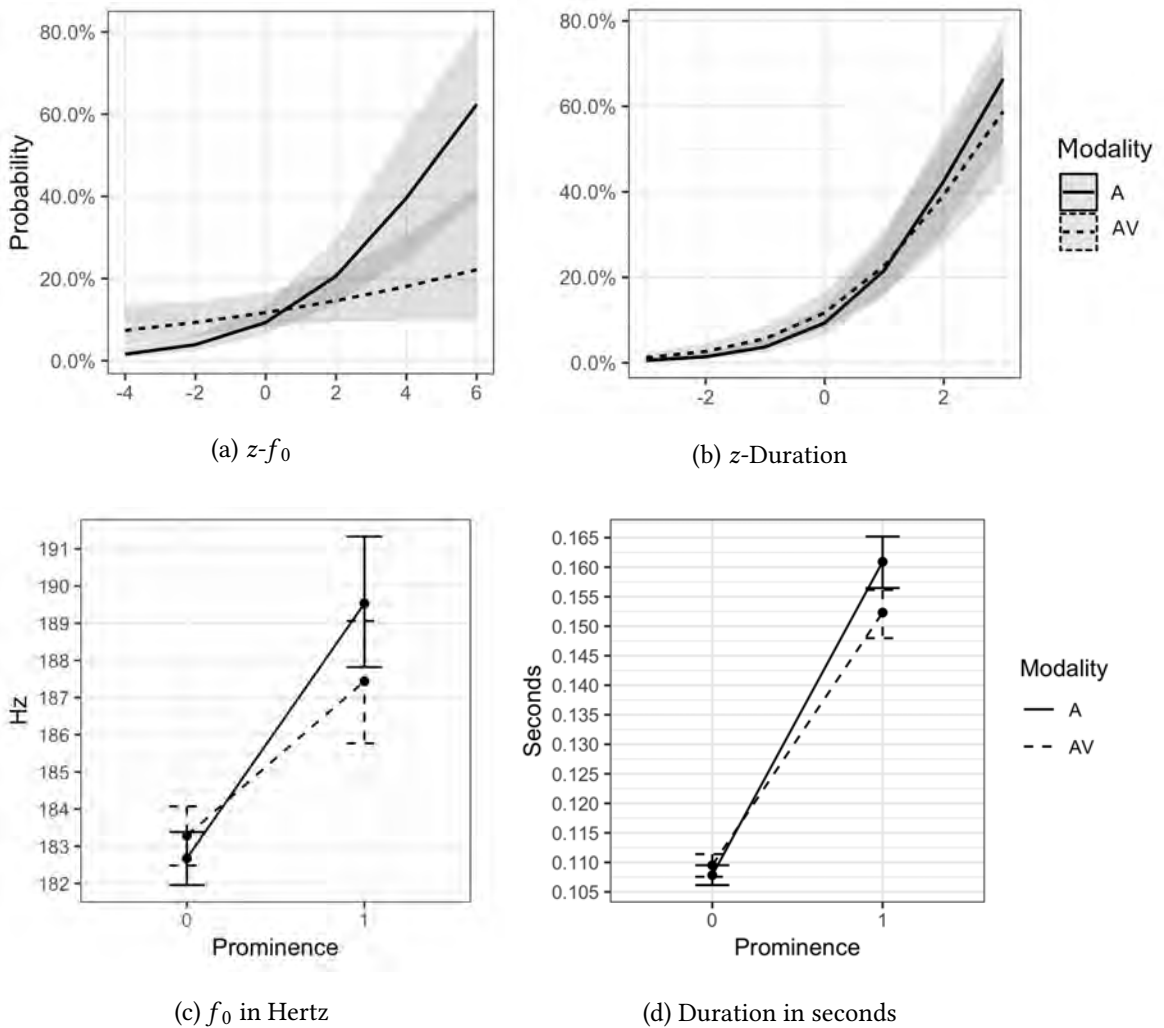
Interactions

This interaction between the audiovisual modality and the experimental conditions revealed that words were less likely to be considered prominent in the audiovisual modality by 2.08 times ($\beta = -0.60$, $SE = 0.23$, $z = -2.52$, $p < .05$) and 1.81 times ($\beta = -0.73$, $SE = 0.24$, $z = -3.02$, $p < .05$) in C1 and C3, respectively. No significant interaction, however, was found for C2.

Not only did the audiovisual modality interact with the experimental conditions C1 and C3, but it also made that the visual and the acoustic cues of prominence were perceived differently. On one hand, apexes coinciding with a word in

Experiment I

the audiovisual modality were 2.32 times less likely to be given a mark of prominence than in the audio-only modality ($\beta = -0.83$, $SE = 0.28$, $z = -2.93$, $p < .05$). Precisely the opposite, however, was found for strokes, which increased the odds of prominence for the words they co-occurred with by 2.30 ($\beta = -0.83$, $SE = 0.28$, $z = 2.91$, $p < .05$, $OR = 2.30$). On the other hand, words considered prominent in the audiovisual modality were 1.40 times less likely to be predicted by the effect of f_0 ($\beta = -0.33$, $SE = 0.08$, $z = -3.87$, $p < .001$), and 1.21 times less likely by the effect of duration ($\beta = -0.19$, $SE = 0.08$, $z = -2.29$, $p < .05$) when compared to the audio-only modality (Figure 40).



Experiment I

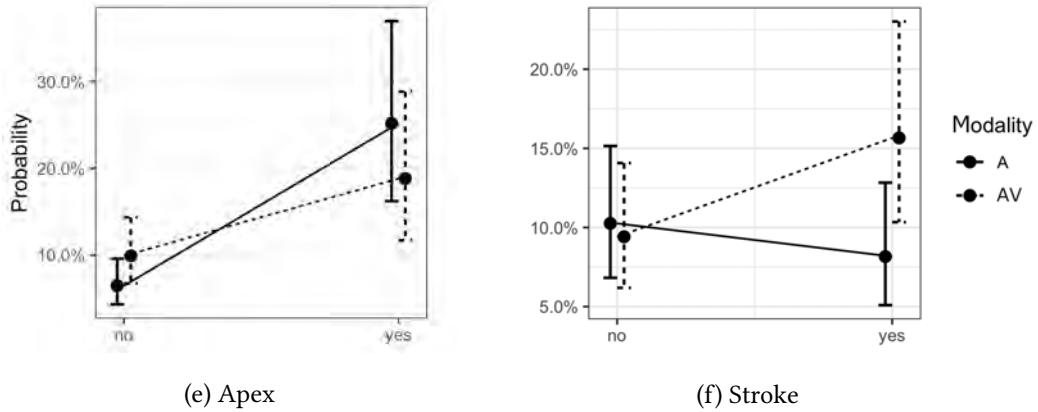


Figure 40: Interactions between modality and predictors in model M18. Graphs (b) and (c) show interactions corresponding to the acoustic values of f_0 and duration, respectively, as a function of participants' responses (non-prominent vs. prominent).

Model M19, in which the non-significant predictor *gesture-recoil* was removed, showed β -values that hardly varied for the same significant predictors as M18 (Table 21).

Predictor	β (SE)	p	95% CI for odds ratio		
			Lower limit	Odds ratio	Upper limit
Intercept	-3.91 (0.50)	< .001***	0.01	0.02	0.05
Modality (0=A, 1=AV)	0.55 (0.23)	< .05*	1.11	1.73	2.71
Condition 1	1.84 (0.53)	< .001***	2.27	6.39	17.99
Condition 2	1.31 (0.51)	< .05*	1.37	3.71	10.10
Condition 3	1.98 (0.51)	< .001***	2.63	7.24	19.97
z-F0 (standardized f_0)	0.46 (0.07)	< .001***	1.37	1.59	1.85
z-Duration (standardized duration)	0.98 (0.09)	< .001***	2.24	2.67	3.19
Stroke (0=no, 1=yes)	-0.34 (0.25)	.33	0.43	0.71	1.17
Apex (0=no, 1=yes)	1.56 (0.28)	< .001***	2.77	4.80	8.31
Hold (0=no, 1=yes)	0.68 (0.23)	< .01**	1.25	1.98	3.16
Modality AV x Condition 1	-0.73 (0.24)	< .01**	0.30	0.48	0.77
Modality AV x Condition 2	-0.28 (0.22)	.19	0.49	0.75	1.15
Modality AV x Condition 3	-0.60 (0.24)	< .05*	0.34	0.55	0.87
Modality AV x z-F0	-0.33 (0.08)	< .001***	0.60	0.72	0.85
Modality AV x z-Duration	-0.19 (0.08)	< .05*	0.70	0.82	0.97
Modality AV x Stroke	0.82 (0.27)	< .01**	1.32	2.28	3.94
Modality AV x Apex	-0.83 (0.28)	< .01**	0.25	0.43	0.76
Modality AV x Hold	-0.13 (0.24)	.61	0.55	0.87	1.40

Table 21: Results of fixed effects in model M19 ($AIC = 6697.83$), in which the non-significant predictor *gesture-recoil* was removed.

4.4 Discussion

This experiment aimed at contributing with a new methodological proposal to the study of the multimodal perception of prominence as well as at offering new insights into Castilian Spanish as language of study. By doing so, it sought to overcome some of the limitations of previous methodologies having used either animated agents or elicited gestures with controlled speech stimuli in experimental settings. The obtained results showed that the methodology proved successful and that participants tended to mark more words in the experimental conditions—where they relied only on certain acoustic cues—than the two trained listeners who marked words in the control condition; in this sense, participants also seemed to respond differently than the control in the audiovisual modality as a result of the perceptual effect of the visual cues of prominence.

Stimuli creation

On the one hand, studies on prominence perception with animated agents in the form of talking heads have limited themselves to reproduce eyebrow and head movements but have excluded hand movements (e.g. House et al., 2001; Prieto et al., 2011). On the other hand, stimuli used in experimental settings have mostly been created through audiovisual recordings in a controlled environment (e.g. Dohen & Lœvenbruck, 2009; Krahmer & Swerts, 2007). For example, the participants of Krahmer and Swerts' study were instructed to utter a short sentence and produce concomitantly a quick visual gesture with either hand, eyebrows, or head on a specific target word. They were also allowed to train until they felt they could not improve the realisation of the gesture accompanying the uttered sentence. These audiovisual recordings were later used as stimuli in an experiment on the effect of gestures to prominence perception.

The shortcomings of previous methods make it difficult to address how the different acoustic correlates of prominence relate to one another and also to gestures. In this study, the speech signal was manipulated to neutralise the prominence-lending properties of f_0 and intensity in three different experimental conditions. Whereas the manipulation of intensity did not present much of a problem, manipulation of f_0 easily resulted in the complete flattening of f_0 contours, thus eliminating the naturalness of the speech stimulus. As a result, intonation peaks were reduced while respecting minimal f_0 contours within a narrow range of 20 Hz for all sentences, corresponding to 2.22 ST ($SD = 0.68$ ST). In this case, a 20-Hz difference between the highest and lowest peak was used, which gave an average of 2.22 ST for all sentences, thus falling within the perceptual threshold proposed by 't Hart (1981), although just noticeable differences (JND 's) for perceiving f_0 changes in intonation differ among researchers ('t Hart, 1981; Klatt, 1973; Rietveld & Gussenhoven, 1985). Nevertheless, a semitone scale was used in Experiment II, as explained above (§ 3.3.3). The manipulation of stimuli was limited in that it was not possible to conceive of a suitable way to also manipulate duration in the audiovisual modality without creating a mismatch between the neutralised duration of syllables and the articulatory gestures of speakers.

Hypothesis 1: methodology

Initially, a correlation test was conducted, and participants' agreement through Cohen's kappa was calculated in order to assess the feasibility of the experimental task despite the variability inherent to the spontaneous speech stimuli and the manipulations on the acoustic correlates of prominence. The correlation test showed consistency between the prominence marks given by participants and sentence length, especially in longer sentences. Agreement among

participants' marks of prominence was calculated through mean Cohen's kappa, which yielded a value of $\kappa = .39$. This low agreement reflects the disparity of judgements across the three experimental conditions in both modalities, which included very different acoustic and visual cues of prominence, especially when compared to the inter-rater agreement obtained by two trained listeners ($\kappa = .83$), who counted on all acoustic cues of prominence in their marking. When participants' prominence marks were compared pairwise to those of the two trained listeners, mean Cohen's kappa decreased slightly ($\kappa = .38$), thus revealing a deviance from this 'gold standard'. Nevertheless, previous studies using stimuli lacking any acoustic manipulations, and presented in the audio-only modality, reached just slightly higher values. For example, Streefkerk (2002) obtained a mean Cohen's kappa of $\kappa = .56$ for prominence marking at a word level, and of $\kappa = .48$ for prominence marking at a syllable level, while Mo et al. (2008), using Fleiss's kappa for agreement among multiple raters obtained a value of $\kappa = .39$.

In this experiment, prominence marking was conducted at a word level because the length of most sentences would have made it difficult for participants to label syllables or syllabic groups. Besides, it has been observed that words are more meaningful units than syllables for prominence marking by naive listeners (e.g. Rosenberg & Hirschberg, 2009; Streefkerk, 2002), as previously discussed (§ 3.1.1). Similarly, it was opted for binary marking (prominent vs. non-prominent) in order to simplify the experimental task. In other studies on the perception of prominence, binary prominence marking was conducted, for example, for two target words (House et al., 2001), short sentences with several target words (Krahmer & Swerts, 2007), or read-aloud sentences from a telephone corpus (Streefkerk et al., 1997) (see Table 8 for a summary). In the procedure applied here, similar to Streefkerk et al.'s (1997) study, all words in the sentence were made available for marking, which could also demand a great effort from the participants depend-

ing on sentence length. However, although this was compensated by offering the possibility to participants to play back each stimulus several times until completion of the marking, it was possible that participants relied excessively on reasoning rather than on perceptual processes. Therefore, a single playback of the stimuli in the marking task might be a better option in a subsequent experiment, together with the use of short sentences as stimuli. By doing so, participants can rely more on short-term memory in their judgements (see Quak et al., 2015, for details). Besides, it was observed that the experimental task, consisting on binary prominence marking, should also include filler sentences that require constant attention from participants on the screen, especially during the audiovisual modality. It was the case that raters, knowing that they were to rate for prominence during the experimental task, often closed their eyes in advance to concentrate on the speech input and disregarded the visual cues of prominence at their disposal.

In order to understand which variables determined prominence marking in this pilot study, different generalised linear mixed models (GLMMs) were built, and they were later compared by means of the Akaike Information Criterion *AIC* (Akaike, 1973) (see Table 19). Despite the fact that model building and selection based on *AIC* are still relatively uncommon in the field of linguistics (e.g. Adamou et al., 2018), this procedure has been gaining popularity over the last decade in other fields such as biology, ecology, or evolution, where it is now often applied (see e.g. Arnold, T. W., 2010; Grueber et al., 2011, for details). In this study, the estimation and comparison of different models showed that in the minimal adequate model, *M18*, prominence marking was determined by the predictors *modality*, *condition*, *fundamental frequency*, *duration*, *gesture-stroke*, *gesture-apex*, *gesture-hold*, together with the interactions of *modality* with: *condition*, *fundamental frequency*, *duration*, *gesture-stroke*, and *gesture-apex*. Neither

intensity nor *gesture-retraction* were included in the final model, and although *gesture-recoil* was, it proved non-significant.

Although random slopes have been suggested to be included in mixed models for all significant predictors to assess the significance of fixed effects (e.g. Baayen et al., 2008; Barr et al., 2013), it becomes increasingly difficult to specify a maximal random-effects structure as suggested by Barr et al. (2013). For example, differences in the random effects between model *M18* and model *M19*—the latter included random slopes for *modality* and *condition* in by-subject random effects—did not result in a lower *AIC* value. Additionally, when trying to declare random effects in this *M18* including both *modality* and *condition*, the model, *M18'*, failed to converge and was dropped out. This problem has previously been pointed out by Bates et al. (2015a), since attempting to incorporate main effects and interactions for subjects and items often requires the estimation of a high number of variance-covariance parameters (see § 3.2.1).

Hypothesis 2: acoustic correlates of prominence

As predicted in hypothesis 2.1, the differences in prominence marks of all experimental conditions were significant when compared to the control condition. The highest number of marks was found in C3, when listeners relied on minimal acoustic cues of prominence in the speech signal. These differences proved also significant in the statistical model *M18*, the minimal adequate model, which confirmed that words were more likely to be marked as prominent in all three experimental conditions.

As for hypothesis 2.2, it is beyond the scope of this pilot study to pinpoint the variables responsible for the marks of prominence given in each condition. In addition, the similar pattern observed in C1 and C3 might be due to the fact that participants lacked the cues of prominence provided by f_0 . Fundamental frequency

has traditionally been considered a reliable cross-linguistic cue of prominence and has also often been held as perceptually more important than duration for prominence perception (e.g. Beckman & Edwards, 1994; Kohler, 2008; Pierrehumbert, 1980; Terken, 1996). Nevertheless, in the responses given by participants, duration had a stronger effect than f_0 , maybe due to the fact that it was the only correlate available across conditions. The crucial role of duration in the perception of phrasal prominence is also supported by previous studies (e.g. Kochanski et al., 2005; Ortega-Llebaria, 2006; Mo, 2008a; Silipo & Greenberg, 2000). Interestingly enough, both correlates were found to have an effect on the probability of words being given a mark of prominence through the predictor *modality*, and they seemed to play a less important role in prominence perception in the audio-visual modality. Seen from a different perspective, the stressed syllable of words needed to be higher in pitch and longer in duration in the audio-only than in the audiovisual modality to be considered prominent (Figure 40c and Figure 40d).

Hypothesis 3: gestural correlates of prominence

In the case of hypothesis 3.1 regarding the role of the audiovisual modality, it was observed that the number of marks given by participants when the visual cues of prominence were present increased in all conditions respect to the audio-only modality, although none of these differences was significant (Figure 36a). In this sense, all conditions proved significant when compared to the control in the audio only modality (Figure 36b), but in the audiovisual modality only C3 did (Figure 36c), i.e. the marks given in the audiovisual modality did not increase proportionately in all conditions respect to the audio-only modality, as happened in the control condition, but achieved similar values.

Furthermore, as seen in the statistical model *M18*, generally words were more likely to be considered prominent when the visual cues of prominence were

present. However, in the interaction between modality and condition, this effect was reversed and listeners were less likely to mark words in C1 and C3 in the audiovisual modality than in C0. This was not the case for C2, which did not prove significantly different from the control condition when both modalities were compared. Possibly, the uncertainty caused by the absence of clear acoustic cues resulted in more random marking of prominence, and the visual information might help participants concentrate their prominence marks around the more clearly perceived visual cues of prominence.

Since most gestures were performed with more than one articulator, the analysis did not focus on differences among gesture articulators, but rather on gesture phases. The marks of prominence given to each gesture phase were observed to vary among conditions and modalities. Hypothesis 3.2 was not supported by the results, which showed that stroke seemed to be the only phase that significantly received more marks of prominence in the audiovisual modality (Table 17 and Figure 37b). This was confirmed in model M18 by an interaction between *stroke* and *modality* revealing that participants were more likely to mark words coinciding with this gesture phase than with apexes in the audiovisual modality. Conversely, apexes made words much less likely to be considered prominent in the audiovisual modality. The time-alignment between apexes and pitch accents has been consistently reported (e.g. Esteve-Gibert & Prieto, 2013; Jannedy & Mendoza-Denton, 2005; Leonard & Cummins, 2010; Loehr, 2004), so it is possible that the manipulations conducted in the speech signal might be responsible for this difference. To a lesser extent, participants also considered prominent words that coincided with the gesture phase of hold, probably because the still movement of articulators—mainly hands—afforded a perceptually clear cue.

Additionally, this Experiment I has confirmed Ambrazaitis and House's (2017) results for the prevalence of gestures simultaneously performed with more than

one articulator. In their study, Ambrazaitis and House analysed use of eyebrow and head movements to convey information structure using recordings of TV newsreaders and found that eyebrow movements were rarely produced in isolation but occurred much more frequently together with a head movement. Similarly, in this study using spontaneous speech material, it was observed that gestures were mostly produced by combining different body parts, especially hands and head (51.1% of the occurrences); and, to a much lesser extent, hands, eyebrows, and head (11.1%). In the case of gestures that were produced with a single body part, it was mostly the hands that performed the gesture (23.5%). Interestingly enough, and differently from Ambrazaitis and House, very few occurrences of gestures produced with eyebrows and head (1.1%) were found, which might be due to the limited expressivity of newsreaders compared to more expressive samples from the speech material used in this study, i.e. spontaneous speech from television talent shows.

All in all, this pilot study indicates that the methodology here proposed can be successfully applied to the study of the multimodal perception of prominence. Spontaneous speech extracted from television talent shows was used to create stimuli applying the neutralization of both f_0 and intensity as acoustic cues of prominence. This and the use of a binary marking at a word level—provided variability is controlled for—have proved reliable in a prominence-marking task. Thus, a large-scale study is envisaged in Experiment II to better understand the multimodal perception of prominence in each experimental condition, which is to confirm the results obtained here, which have provisionally confirmed some of the hypotheses that were initially formulated. On the other hand, research on how acoustic and visual cues of prominence interact presents some limitations

Experiment I

both as to how naturally elicit gestures in controlled settings and as to what extent multimodal perception studies with animated agents can be generalised. The approach offered here can contribute in this sense by adding ecological validity to previous studies and can serve as a useful method in the more general research of multimodal interactions.

Experiment II

5.1 Introduction

The aim of this second experiment was to build on the methodology laid out in the previous pilot study in order to analyse how the acoustic cues of prominence are used by listeners with and without visual cues to detect phrasal prominence in Castilian Spanish.

The acoustic correlates of prominence present cross-linguistic differences. In the case of Spanish, lexical prominence is rendered by a flat pitch contour, together with longer duration and stronger intensity for unaccented stressed syllables; while phrasal prominence, i.e. accented stressed syllables, is cued by longer duration, higher f_0 , larger f_0 excursions, and increased overall intensity (Ortega-Llebaria, 2006). Furthermore, the preponderant role of duration, both as a correlate of lexical stress and phrasal stress, was confirmed by Vogel et al. (2016), although they also observed that f_0 contributed to cue both stressed and unstressed syllables in the absence of pitch accents and that intensity helped duration in cueing accented stressed syllables, but not accented unstressed ones.

On the other hand, it has been observed that the visual correlates of prominence (e.g. gestures performed with hands, eyebrows, or head) interact with verbal prosody (e.g. Al Moubayed et al., 2010; Granström et al., 1999; Kim et al.,

2014; Krahmer & Swerts, 2007; Prieto et al., 2011; Scarborough et al., 2009). In the case of prominence perception, visual cues result in stronger production and perception of verbal prominence (Krahmer & Swerts, 2007); and in the case of facial gesturing, for example, it has been found to systematically influence the perception of verbal prominence (Dohen & Løevenbruck, 2009; House et al., 2001; Swerts & Krahmer, 2008).

Most studies having addressed the interaction of visual and verbal prominence have so far made use of both lip-synchronised animated agents (e.g. Al Moubayed & Beskow, 2009; Granström et al., 1999; House et al., 2001; Krahmer et al., 2002a,b; Prieto et al., 2011) and experimental settings in which gestures are elicited with controlled speech stimuli (e.g. Dohen & Løevenbruck, 2009; Foxton et al., 2010; Krahmer & Swerts, 2007). In both cases, the visual cues of prominence—limited to beats produced by eyebrow raises and head nods, and occasionally also by hand gestures (Krahmer & Swerts, 2007)—have been observed to enhance verbal prominence perception.

Certain studies have pointed out differences between men and women in the audiovisual perception of speech, which has been supported by neuroanatomical differences, with a stronger activation in bilateral brain areas causing a more efficient audiovisual language processing in women (Dancer et al., 1994; Öhrström & Traunmüller, 2004; Ruytjens et al., 2006, 2007; Watson et al., 1996).

Additionally, different gesture phases align with different intonational units, and a strong temporal connection has been reported between apexes and pitch accents (Esteve-Gibert & Prieto, 2013; Kendon, 1972; Jannedy & Mendoza-Denton, 2005; Loehr, 2012), giving support to the phonological synchrony rule (PSR) posited by McNeill (1992).

This Experiment II is based on the pilot study conducted on Experiment I, however in this case some methodological differences exist from Experiment

I. First of all, the study was conducted online, and stimuli were presented to participants in two modalities: audio-only and audiovisual in four independent experiments corresponding to four experimental conditions, which involved the neutralization of the acoustic correlates of prominence—similar as in Experiment I (§ 3.3.3). The neutralization was conducted on (a) f_0 in Exp1, (b) intensity in Exp2, and (c) both f_0 and intensity in Exp3; no neutralisation in the control condition, Exp0. Similar to Experiment I the prominence marks provided by two trained listeners served as ‘gold standard’ for agreement comparisons.

Since obtaining details of the variables predicting prominence perception in each experimental condition was not possible in Experiment I, the between-subjects design of this Experiment II allowed to conduct analyses separately for each experimental condition in order to better understand the factors involved in the multimodal perception of acoustic prominence. Due to the provisional nature of the results in Experiment I, all initial hypothesis are maintained. The aspects common to both experiments have been previously discussed: speech material (§ 3.3.2), participants (§ 3.3.4), gesture annotation (§ 3.3.5), and data analysis (§ 3.3.6); however, details of the current experiment differing from Experiment I are reported in each specific section.

5.2 Methodology

5.2.1 Participants

Participants were mainly recruited through social media to take part in an online study that was advertised as an experiment on memory and perception. For their participation they were not financially compensated.

Several criteria were used to assure the reliability of the collected data: participants had to have Castilian Spanish as their mother tongue; be settled in Spain

Experiment II

at the moment of participation (according to their IP-addresses); and have taken at least 6 minutes to complete the experimental task, but no more than 13 minutes (mean time for completion was 9' 17", $SD = 2' 32''$), since the intention was to prevent them from overly relying on logical inferences in their marking for prominence. After applying these criteria the answers provided by 240 naïve listeners (68 men and 172 women) were used—30 per modality and condition—, adding up to total of 12960 rated words. The declared age of participants ranged between 18 and 66 years ($M = 36.98$, $SD = 10.55$; $M_{\text{men}} = 39.80$, $SD = 10.75$; $M_{\text{women}} = 35.86$, $SD = 10.26$), with a predominance of participants under 50 years of age (Figure 41, see Appendix B1a for details).

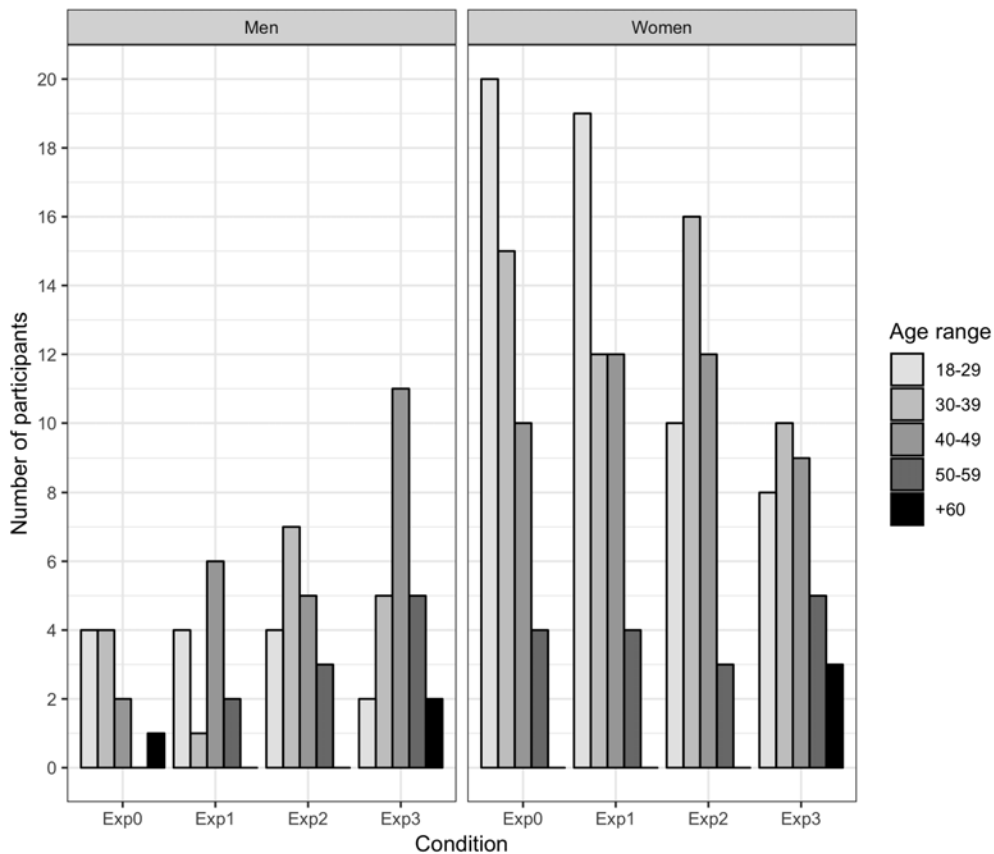


Figure 41: Details of participants' age per condition.

Experiment II

At the end of the experiment, information about participants' mother tongue and musical training was collected. For the reasons previously explained (§ 3.3.4), it was decided to control for the effect of participants' ability to perceive small changes of frequency in the prominence-marking task, especially given the nature of the stimuli, in which f_0 was neutralised in two of the four conditions by keeping the intonation curve within a short range of 2 ST. Thus, participants were grouped according to their level of musical training into those with no musical training (*none*), up to 5 years of formal musical training (*little*), and over 5 years of formal musical training (*much*) (Figure 42, see Appendix B1b for details).

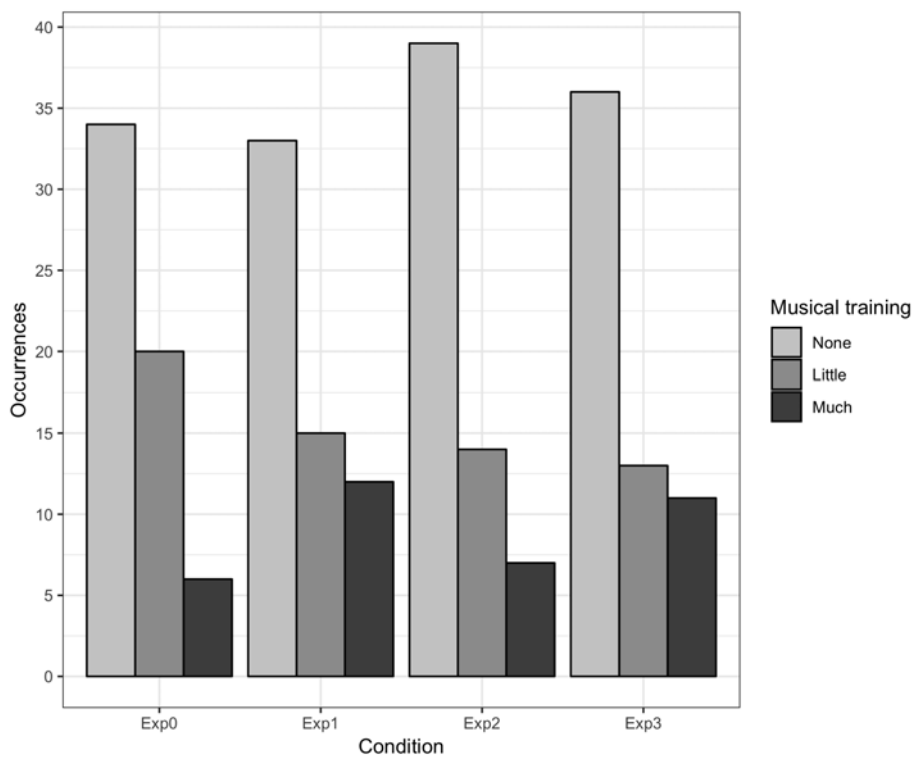


Figure 42: Details of participants' level of musical training per condition.

5.2.2 Stimuli

Four target sentences were manipulated to neutralize f_0 and intensity as acoustic correlates of prominence and were administered in three different experimental conditions, as described above (§ 3.3.3). The only difference respect to Experiment I is that neutralised f_0 was manually smoothed within a 2-semitone-range between the highest and lowest values of the resulting intonation curve. A fourth, condition was included with a control group relying on all acoustic cues of prominence.

In experimental condition Exp1, where f_0 was neutralised, the values for intensity ranged between 62-80 dB ($M = 73.11$, $SD = 3.54$) for the stressed vowel and for duration, 0.027-0.379 seconds for the syllable ($M = 0.110$, $SD = 0.069$). In Exp2, the values for intensity were averaged at 69 dB, while f_0 ranged between 140-307 Hz ($M = 199.31$, $SD = 40.04$) for maximum f_0 in the stressed vowel (or vowel in adjacent syllable), with the same values as in Exp1 for duration. Finally, in Exp3, both f_0 and intensity were neutralised as in the previous two conditions, and duration served as the only acoustic cue of prominence, with the same values as in Exp1 and Exp2. In the control condition, Exp0, participants relied on all non-manipulated acoustic cues of prominence. Additionally, as in Experiment I, two trained listeners also annotated the non-manipulated stimuli both in the audio-only and in the audiovisual modality (§ 4.2.2).

5.2.3 Experiment design

This Experiment II was designed to be conducted via the Internet (§ 5.2.1) as four independent experiments, one for each of the conditions to be tested. Due to the time limitations that carrying out an experiment on the Internet may involve, the experiment was designed in such a way that 30 participants only rated once

the same stimulus in one of the modalities and in one of the conditions. This also avoided unwanted learning effects in participants. The resulting between-subjects design had 60 participants taking part in each experimental condition: half of them rated the stimuli in only one of the two modalities, while the other half did so in the remaining modality. Adding up the four conditions, a total of 240 participants rated the same stimuli.

Stimuli were randomised for each condition, and they included 4 manipulated target sentences, 2 non-manipulated sentences, and 3 filler sentences asking participants to report on either a visual element in the images they had just seen, or a word from the sentence they had just heard (see Appendix B for details). Filler sentences were intended to make participants pay close attention to the images displayed on the screen. It had been observed in the previous experiment, similar to Krahmer and Swert's (2007) experience, that participants tended to close their eyes and concentrate on the acoustic cues of prominence once they had got used to the experimental task, thus neglecting the images on the screen that were at their disposal in the audiovisual modality.

5.2.4 Hypotheses

For this Experiment II, all three hypotheses 1, 2, and 3 previously formulated for Experiment I are retained, even if some were not fully supported in the results presented there. Due to the provisional nature of the results obtained in Experiment I, it is intended to offer more solid arguments in this Experiment II in order to finally refute or confirm the hypotheses initially formulated. Apart from them, two more hypotheses are now added for individual variables that were not possible to analyse due to the small sample size:

4. Musical training

Participants having had formal musical training will perform significantly different than those with little or no musical training when prominence marking involves pitch discrimination skills, especially when f_0 is available as an acoustic cue.

5. Participant gender

Women will perform significantly differently from men at the experimental task in the audiovisual modality.

5.2.5 Procedure

The experiment was conducted using the online survey software *SmartSurvey*¹. At the recruitment stage, participants were asked to take around 10 minutes to conduct the experiment on a computer in a quiet environment through headphones. In each experimental condition, they were firstly presented with a brief set of instructions explaining that the experiment involved two different tasks: on the one hand, binary prominence marking, in which they had to mark any words they considered prominent in the sentence; and on the other hand, a memory task, in which they had to report on certain visual elements that had been displayed on the screen, or certain words in the sentence they had just heard (see Appendix B1 for details). After instructions, participants had the opportunity to get acquainted with the experiment in a series of trials.

As for both experimental tasks, they were conducted in two stages: firstly, participants were presented with a sentence on the screen (Figure 43a); then, a second screen revealed the task they were required to complete, whether a prominence task (Figure 43b) or a memory task, i.e. filler sentences (Figure 43c). In the case of prominence marking, participants were allowed to play back the

¹ <https://smartsurvey.co.uk>.

Experiment II

clip just once more in order to decide which words were prominent and click on the check-boxes displayed under the words uttered by the speaker. In the case of the memory task, they had to answer a multiple choice question.

Importantly, participants were not allowed to go back to a previous screen through the 'back-arrow' icon on their web browsers. If they did, the experiment was immediately invalidated. Participants were not allowed either to skip a question without giving a response, as the red asterisk at the end of the question displayed on the screen indicated (Figure 43).

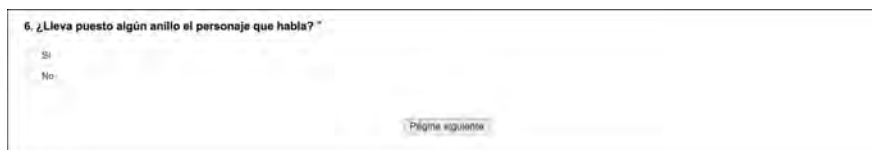
Experiment II



(a)



(b)



(c)

Figure 43: Three sample screens corresponding to the two experimental tasks of the experiment in the audiovisual modality. Screens (a) and (b) correspond to the same stimulus for prominence marking: (a) shows a videoclip for participants to watch just once. After clicking on *Página siguiente* ('Next page') at the bottom, screen showed in (b) is displayed to mark prominence, as indicated by checkboxes. Screen (c) shows an example of the memory task, with a question about any visual element displayed on the screen. Screen (c), same as (b), is displayed after having watched a videoclip just once like that in (a).

5.3 Results

5.3.1 Descriptive statistics

5.3.1.1 Prominence marks

The four target sentences included 54 words available for marking to each participant and added up to 12960 words, out of which 3202 (24.7%) received a mark of prominence. Prominence marks per sentence ranged between 2.76 and 3.97 ($M = 3.33$, $SD = 1.98$). The two speakers received, respectively, 1619 marks of prominence out of 7440 words (21.7%), and 1583 out of 5520 (28.6%) (see Appendix B2 for details). In addition, *Negation* was the word class that ranked highest, receiving a mark of prominence in 65.8% of the cases (see Appendix B3 for details).

5.3.1.2 Inter-rater agreement

Inter-rater agreement was calculated as in Experiment I (§ 4.3.1.2, Table 15). Mean Cohen's kappa (1960) was computed for all pairs of participants ($N = 240$) according to their marks per experimental condition and modality ($n = 30$) (Figure 22).

The highest agreement was reached by the 30 participants that rated the audio-only modality in Exp0 ($\kappa = .42$), which contrasted with the poorer agreement of the 30 participants that rated the same condition in the audiovisual modality ($\kappa = .34$). In addition, the agreement calculated independently for two trained listeners relying on all acoustic cues of prominence in both the audio-only ($\kappa = 1.0$) and the audiovisual modality ($\kappa = .89$) was very high.

Experiment II

Condition	Modality			
	Audio-only (A)	A-TL	Audiovisual (AV)	AV-TL
Exp0	0.422	0.430	0.343	0.287
Exp1	0.290	0.297	0.269	0.222
Exp2	0.283	0.291	0.380	0.386
Exp3	0.350	0.366	0.356	0.368
Trained listeners	1.00		0.89	

Table 22: Details of inter-rater agreement expressed as mean Cohen’s kappa for all pairs of participants per condition and modality: audio-only (A) and audiovisual (AV). Additionally, mean values for pairwise comparisons of all participants’ marks including those of two trained listeners in both modalities are given as A-TL and AV-TL.

With the neutralization of the acoustic cues of prominence, agreement of participants was generally poorer in the three experimental conditions. In the case of Exp1, a similar decrease as in Exp0 was also found in the audiovisual modality ($\kappa = .27$), which represents the lowest agreement overall. Conversely, this was reversed in Exp2, where the visual cues made participants increase their agreement not only over the audio-only modality ($\kappa = .38$), but also over the agreement of the control condition in the audiovisual modality. In Exp3, agreement values were almost identical in the audio-only ($\kappa = .350$) and in the audiovisual modality ($\kappa = .356$).

Interestingly, when the marks of all participants were compared pairwise to those of two trained listeners, which served as a ‘gold standard’, mean Cohen’s kappa increased slightly for Exp0 in the audio-only modality ($\kappa = .43$), but decreased considerably in the audiovisual modality ($\kappa = .28$). A similar pattern was observed for the other three conditions in the audio-only modality. This was not the case, however, in the audiovisual modality, where this increase was minimal for Exp2 and Exp3, while it decreased in Exp1. This reveals that agreement of participants’ responses was higher when compared one-to-one to the ‘gold standard’ in the audio-only than in the audiovisual modality.

5.3.1.3 Prominence and acoustic cues

In order to observe in more detail the distribution of marks, the total words available for marking was broken down per modality and condition. As it can be observed in Table 23, words received fewer prominence marks in the audiovisual modality across conditions. This was not the case for the control condition, Exp0, which received the highest number of marks in the audiovisual modality (26.7%) as well as the lowest number of marks in the audio-only modality (21.0%). When compared to control, all experimental conditions received fewer marks, with a maximum difference of 5.5% observed in Exp2.

Modality	Words for marking per condition and modality	Prominent words (%)			
		Exp0	Exp1	Exp2	Exp3
	3240	774 (23.8)	820 (25.3)	829 (25.5)	779 (24.0)
A	1620	341 (21.0)	415 (25.6)	459 (28.3)	403 (24.8)
AV	1620	433 (26.7)	405 (25.0)	370 (22.8)	376 (23.2)
AV ($\pm\%$)		+5.7	-0.6	-5.5	-1.6

Table 23: Distribution of marks of prominence per experimental condition for both modalities combined ($n = 3240$) and for each modality separately ($n = 1620$): audio-only (A) and audiovisual (AV). Participants relied on intensity and duration in Exp1, on f_0 and duration in Exp2, and on duration in Exp3. In Exp0, participants relied on all acoustic cues of prominence. Total words for marking added up to $N = 12960$.

5.3.1.4 Prominence and visual cues

Prominence was also analysed per gesture phase for all conditions. In this case, the audio-only modality, where participants did not rely on the visual cues of prominence served as the baseline to compare the marks given to words coinciding with the different gesture phase in the audiovisual modality. Due to the small sample size of stimuli in this Experiment II, gestures were not analysed ac-

Experiment II

ording the articulator with which they were performed. Overall, it was observed that fewer words coinciding with each different gesture phases were considered prominent in the audiovisual condition (Table 24).

Gesture phase	Words for marking per condition and modality	Prominent words (%)					
		C0			C1		
		A	AV	AV ($\pm\%$)	A	AV	AV ($\pm\%$)
Preparation	270	25 (9.2)	28 (10.3)	+2.9	35 (12.9)	32 (11.8)	+5.5
Stroke	420	19 (4.5)	27 (6.4)	+1.9	28 (6.6)	43 (10.2)	+3.6
Apex	540	213 (39.4)	262 (48.5)	+9.1	232 (42.9)	225 (41.6)	-1.3
Hold	210	55 (26.2)	64 (30.4)	+4.2	74 (35.2)	58 (27.6)	-7.6
Recoil	180	21 (11.6)	19 (10.5)	-1.1	31 (17.2)	28 (15.5)	-1.7
Total	1620	341 (21.0)	433 (26.7)	+ 5.7	415 (25.6)	405 (25.0)	-0.6

(a) Prominent words according to gesture phase for Exp0 and Exp1

Gesture phase	Words for marking per condition and modality	Prominent words (%)					
		C2			C3		
		A	AV	AV ($\pm\%$)	A	AV	AV ($\pm\%$)
Preparation	258	42 (15.5)	28 (10.3)	-3.4	18 (13.4)	27 (18.3)	+4.9
Stroke	174	22 (5.2)	67(15.9)	+10.7	26 (6.1)	42 (10.0)	+3.9
Apex	270	246 (45.5)	238 (44.0)	-1.5	232 (42.9)	217 (40.1)	-2.8
Hold	164	79 (37.6)	47 (22.3)	-15.3	74 (35.2)	66 (31.4)	-3.8
Recoil	78	25 (13.8)	22 (12.2)	-1.6	37 (20.5)	26 (14.4)	-6.1
Total	1062	459 (28.3)	370 (22.8)	-5.5	403 (24.8)	376 (23.2)	-1.6

(b) Prominent words according to gesture phase (cont.) for Exp2 and Exp3

Table 24: Different gesture phases coinciding with words given a mark of prominence per condition and modality ($n = 1620$; $N = 12960$). Values for the audio-only modality, where no visual information was available, served to compare the marks given by participants in the audiovisual modality. This difference is expressed as AV ($\pm\%$). Acoustic cues were intensity and duration (Exp1), f_0 and duration (Exp2), and only duration (Exp3). All three acoustic cues of prominence were available in the control condition (Exp0).

This difference was highest in Exp2 for the hold phase, where marks decreased by 15.3% between the audio-only modality (79 occurrences, 37.6%) and

the audiovisual modality (43 occurrences, 22.3%). Interestingly, preparation, apex, and hold consistently decreased in all conditions in the audiovisual modality, while strokes is the only phase that increased in all experimental conditions.

The gesture phase with the highest marks of prominence was apex, which received 262 marks out of 540 occurrences (48.5%) in the audiovisual modality in the control condition (Exp0). A minimum was observed for strokes also for Exp0 in the audiovisual modality, with 6 out of 174 occurrences (3.4%).

5.3.1.5 Prominence per sentence: P-score

One of the advantages of presenting four target sentences to participants was the possibility of observing in detail how the marks of prominence were distributed for all sentences across conditions and modalities. Thus, prominence marks were pooled to achieve a more fine-grained scale of prominence. The proportion of participants who considered a word as prominent was expressed as a prominence score (P-score) ranging between 0 and 1 (e.g. Cole et al., 2010; Mo, 2008a; Swerts, 1997). This was calculated as the number of marks given to a certain word within a sentence for each condition and modality divided by the number of possible marks for that word ($n = 30$) (Table 25).

Experiment II

Word	Total marks	Marks per condition and modality								P-score (e.g. Exp3-AV)
		Ex0		Exp1		Exp2		Exp3		
		A	AV	A	AV	A	AV	A	AV	
y	32	4	5	3	6	10	4	0	0	.00
más	208	28	30	24	25	27	29	22	23	.76
si	5	1	1	0	0	0	1	0	2	.06
lo	4	0	0	2	0	0	0	0	0	.00
has	16	1	4	3	1	1	2	3	1	.03
hecho	40	4	6	5	5	7	4	5	4	.13
con	23	1	3	3	3	3	5	3	2	.06
conciencia	159	16	19	22	22	16	20	22	22	.73
de	5	1	0	2	1	1	0	0	0	.00
lo	4	1	1	2	1	0	1	0	0	.00
que	9	1	1	3	1	0	2	1	0	.00
estabas	65	5	9	11	7	10	9	9	5	.16
haciendo	94	9	9	11	10	12	11	17	15	.50

Table 25: Example of cumulative marks of prominence given by 240 participants (60 per condition, 30 per condition and modality) to a target sentence that illustrate how P-scores are calculated (e.g. condition Exp3 in the audiovisual modality).

These values were then compared with the P-scores calculated for two trained listeners as a ‘gold standard’ for all target sentences (Figure 44). Generally, the P-scores of participants seemed to coincide with those of trained listeners, especially for clearly prominent words, most of which co-occurred with the apex of a gesture. This can be seen, for example, in words of: sentence 1 /'nunka/; sentence 2 /'unos/ and /'otros/; sentence 3 /'sako/; and sentence 4 /mas/. This is not the case, however, for all prominent words coinciding with an apex, as in /'ido/ in sentence 3 and /kon'θienθia/ and /mal/ in sentence 4, where the P-scores of participants are consistently lower than those of trained listeners.

Differently, words with low or none prominence in the ‘gold standard’ often reached higher P-scores when marked by participants, as the last words of sen-

Experiment II

tence 1 /'una 'mala pa'labra/; both the initial sequence in sentence 2, /te'nemos mas afini'da/, and the ending words prior to the last word coinciding with an apex, /ke kon/; the beginning of sentence 3, /i lo pri'mero/; and the end of sentence 4, /es'tabas a'θiendo/.

Experiment II

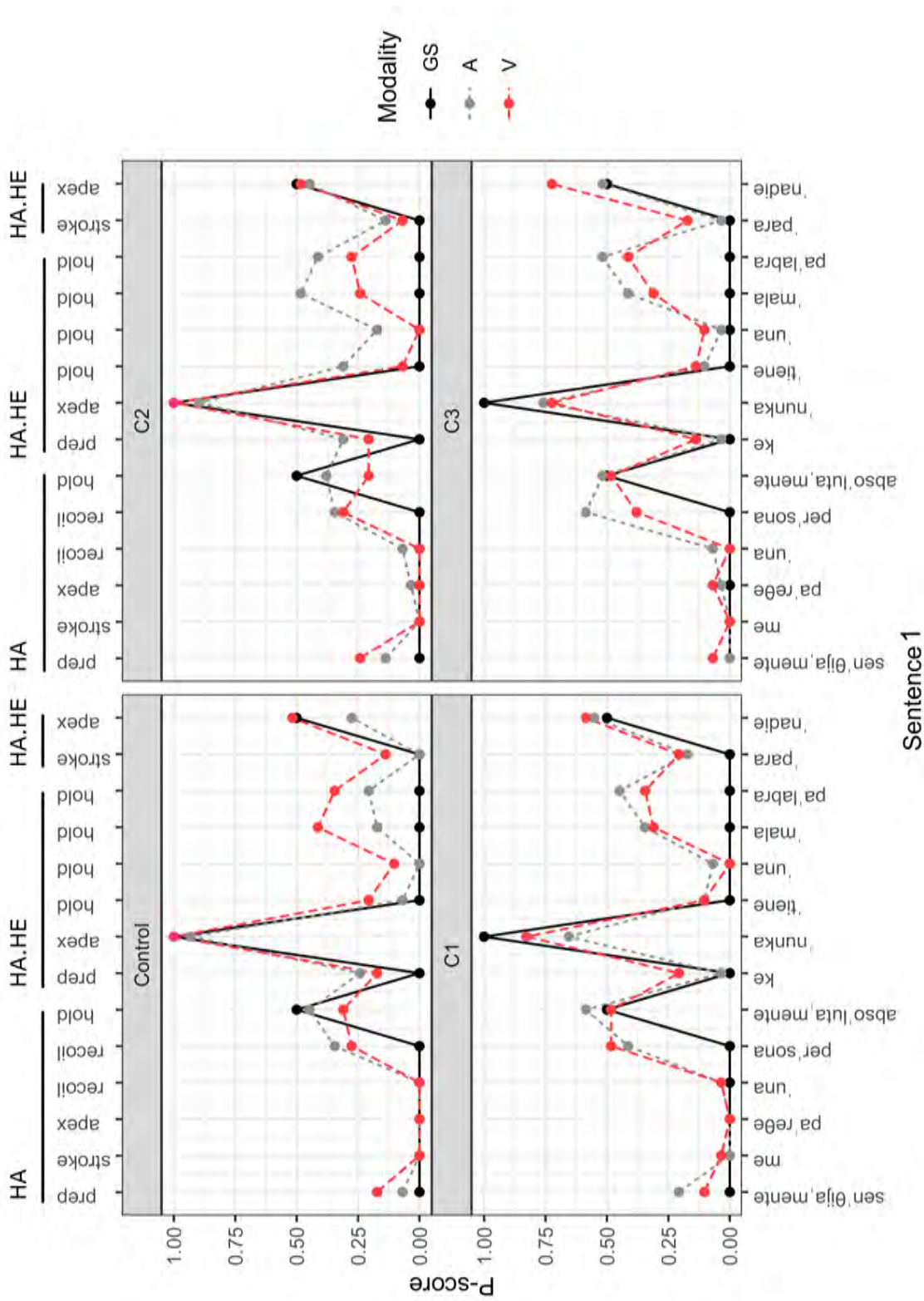
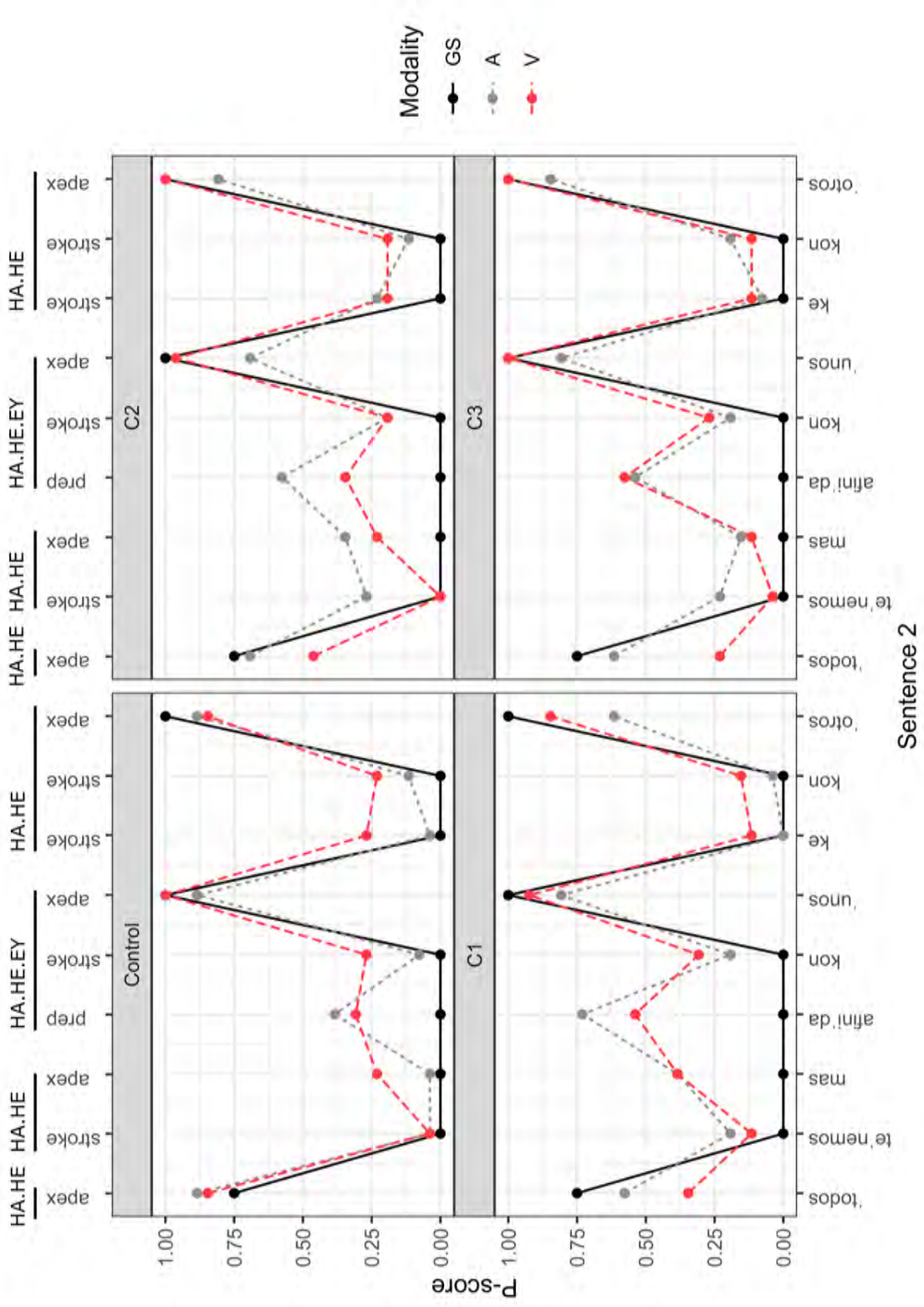
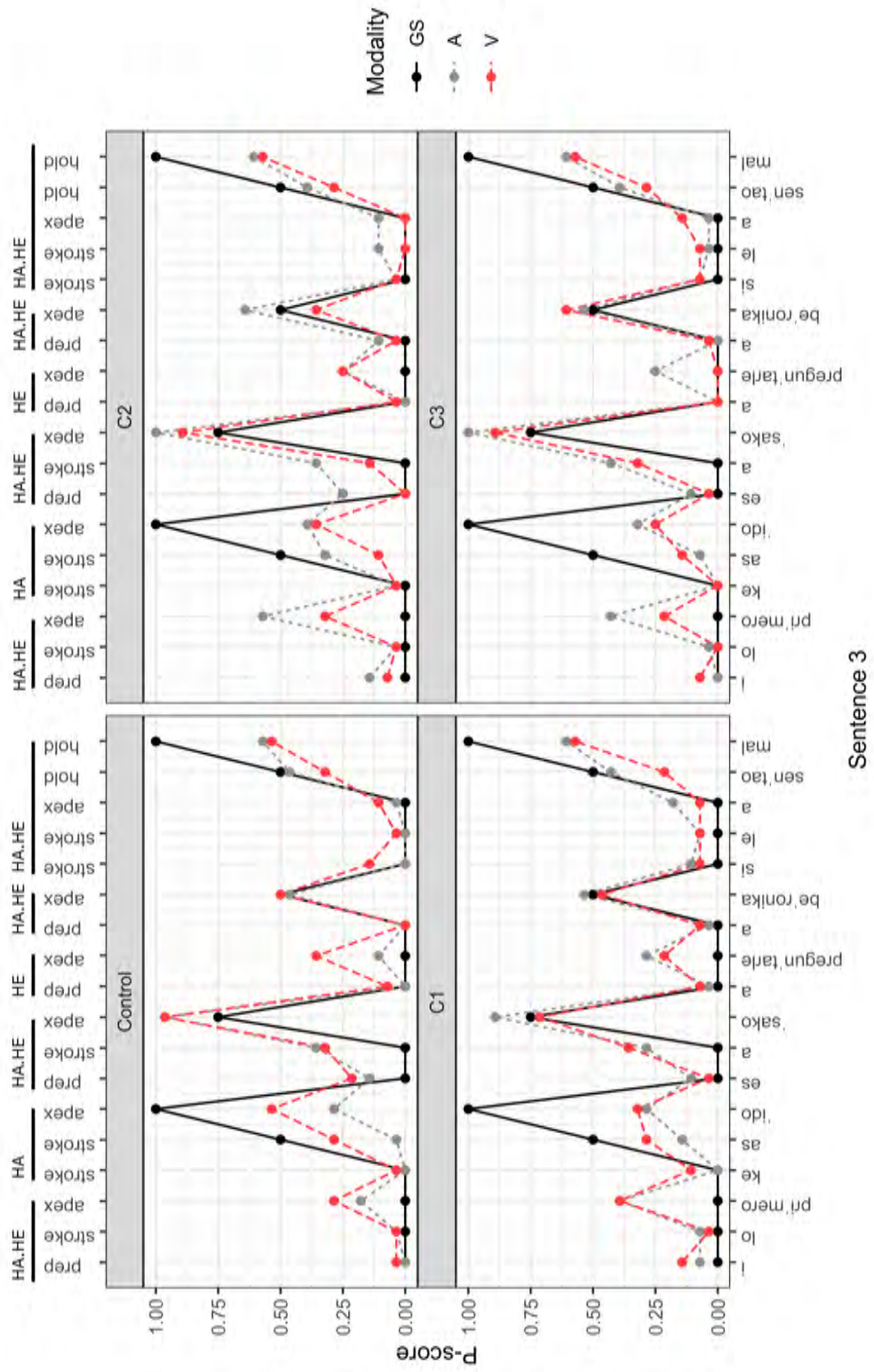


Figure 44: Graphs with P-scores for target sentences resulting from the cumulative marks of 240 participants (30 per condition and modality). Solid black lines shows P-scores of the 'gold standard' (GS) provided by two trained listeners. Dotted grey lines correspond to scores in the audio-only (A) modality, while dotted red lines represent scores in the audiovisual (AV) modality. The lower X-axis shows phonetic transcription, and the upper x-axis, the gesture articulator and the phase coinciding with each word (HA: hand; HA.HE: hand and head; HA.EY.HE: hand, eyebrows, and head). In control condition (Exp0), all acoustic cues were available for participants. They relied on intensity and duration in Exp1, on f_0 and duration in Exp2, and on duration in Exp3.

Experiment II



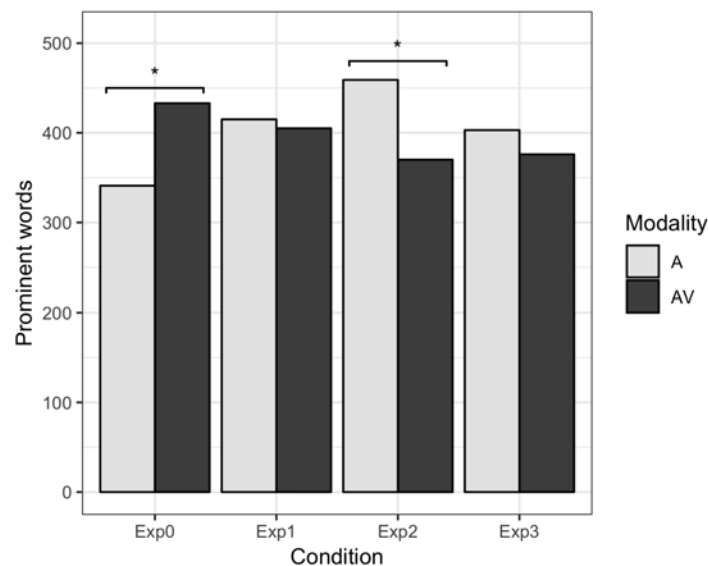
Experiment II



5.3.2 Inferential statistics

5.3.2.1 Number of prominence marks

The difference found between the marks given to each of the two speakers (6.9%) was not significant, $\chi^2(1) = 0.45$, $p > .05$. However, when prominence marks given by participants were compared between the audio-only and the audiovisual modality, a significant difference was found in Exp0, $\chi^2(1) = 14.05$, $p < .05$, and Exp2, $\chi^2(1) = 12.55$, $p < .05$. No overall differences were found between the control condition and each experimental condition nor among experimental conditions (Figure 45a). Furthermore, when the audio-only modality was independently analysed, significant differences were found between the control condition Exp0 and Exp1, $\chi^2(1) = 9.19$, $p < .01$; between Exp0 and Exp2, $\chi^2(1) = 22.72$, $p < .001$; and between Exp0 and Exp3, $\chi^2(1) = 6.49$, $p < .01$; as well as between Exp2 and Exp3, $\chi^2(1) = 4.78$, $p < .01$ (Figure 45b). In the audiovisual modality, significant differences were only found between Exp0 and Exp2, $\chi^2(1) = 6.36$, $p < .05$; and between Exp0 and Exp3, $\chi^2(1) = 5.16$, $p < .05$ (Figure 45c).



(a) All conditions in both modalities

Experiment II

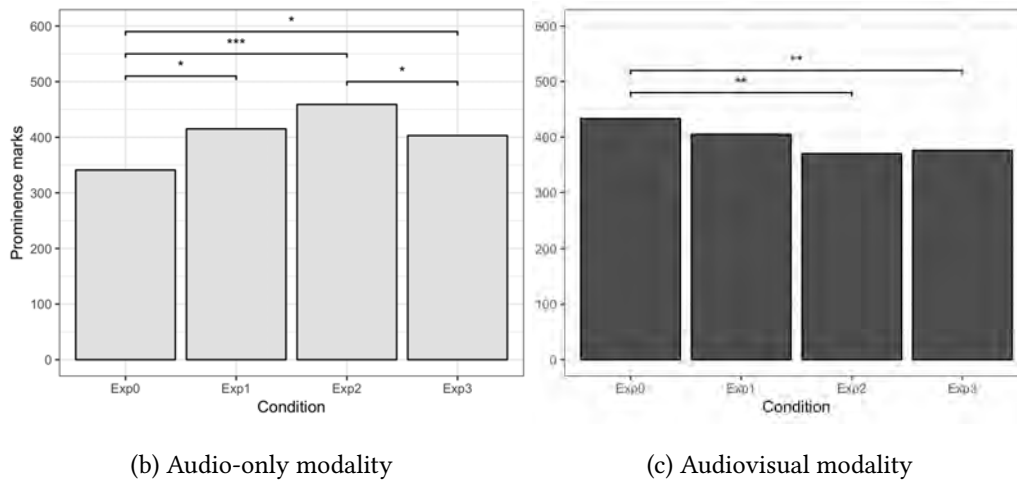


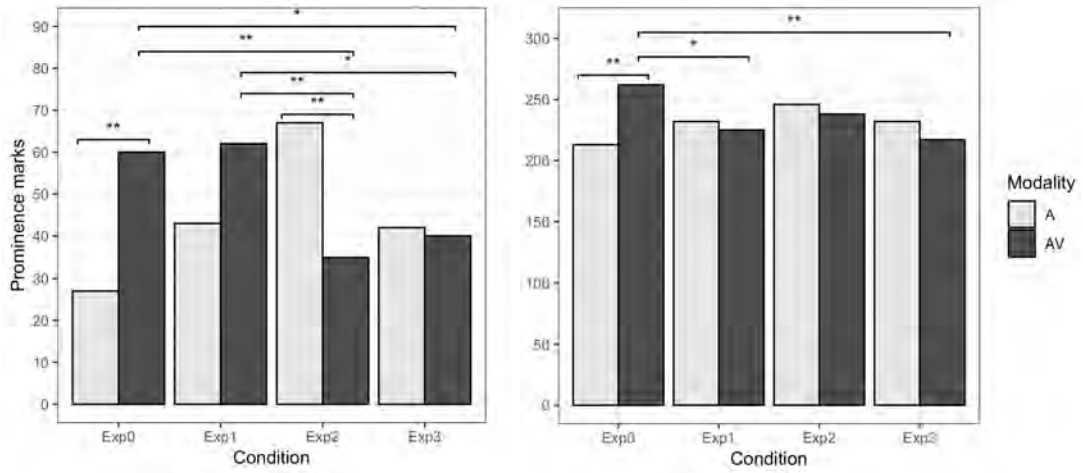
Figure 45: Significant differences in the number of prominence marks per modality and condition. Participants relied on intensity and duration in Exp1, on f_0 and duration in Exp2, and on duration in Exp3. In Exp0, participants relied on all acoustic cues of prominence.

Similarly, marks given to words coinciding with each gesture phase were tested both between modalities in each condition and among conditions only in the audiovisual modality. Significant differences between modalities were found for words coinciding with strokes in Exp0, $\chi^2(1) = 13.13$, $p < .01$; and Exp2, $\chi^2(1) = 10.72$, $p < .01$ (Figure 46a). Also apexes showed significant differences between modalities in Exp0, $\chi^2(1) = 8.65$, $p < .01$ (Figure 46b). As for holds, differences were found in Exp2, $\chi^2(1) = 10.89$, $p < .01$ (Figure 46c). No differences between modalities were found for preparation nor recoil in any condition.

Significant differences in the audiovisual modality for strokes were found between Exp0 and Exp2, $\chi^2(1) = 6.83$, $p < .01$; between Exp0 and Exp3, $\chi^2(1) = 4.09$, $p < .05$; as well as between Exp1 and Exp2, $\chi^2(1) = 7.87$, $p < .01$; Exp1 and Exp3, $\chi^2(1) = 4.92$, $p < .05$ (Figure 46a). In the case of the apex phase of gestures, significant differences in the audiovisual modality among conditions were found between Exp0 and Exp1, $\chi^2(1) = 4.84$, $p < .05$; and between Exp0 and Exp3, $\chi^2(1) = 7.26$, $p < .01$ (Figure 46b). The only difference for holds was found

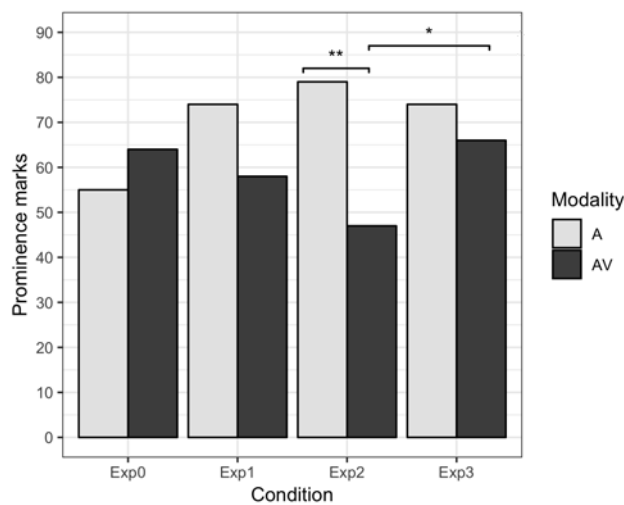
Experiment II

between Exp2 and Exp3, $\chi^2(1) = 3.92, p < .05$ (Figure 46c). No differences in the audiovisual modality among conditions were found for preparation nor recoil.



(a) Stroke

(b) Apex



(c) Hold

Figure 46: Graphs showing significant differences per condition and modality in the number of prominence marks coinciding with different gesture phases. Participants relied on intensity and duration in Exp1, on f_0 and duration in Exp2, and on duration in Exp3. In Exp0, participants relied on all acoustic cues of prominence.

5.3.2.2 Model building and model selection procedure

Similar to Experiment I, several GLMMs with a logit link function were estimated for a binomial distribution. GLMMs are an extension of linear mixed models (LMMs) for non-normal distributions, and both have been previously applied in linguistic research (e.g. Adamou et al., 2018; Gries, 2015; Masson-Carro et al., 2017; Quené, 2008). Mixed-models are more robust statistical tools than conventional analysis of variance (ANOVAs) (Jaeger, 2008), as previously discussed (§ 3.2.1 and § 3.2.1).

The between-subjects experiment design allowed not only to conduct partial comparisons of each experimental condition with the control condition, as done in Experiment I, but also to analyse each condition separately: control condition (§ 5.3.2.4), Exp1 (intensity and duration as acoustic cues § 5.3.2.5), Exp2 (f_0 and duration § 5.3.2.6), and Exp3 (only duration § 5.3.2.7). All models were built through the *lme4* package (Bates et al., 2015b) in R (2018), and all were estimated using maximum likelihood (Laplace Approximation) were optimised with *BOBYQA* (Powell, 2009) to increase iterations and avoid convergence errors.

5.3.2.3 Analysis A: Control vs. experimental conditions

Initially, a first set of models was built aiming at comparing all conditions to the control, Exp0, using the same procedure as in Experiment I (§ 4.3.2.3). The dependent variable *prominence* was modelled using the same variables of: *speaker* ($n = 2$), *sentence* ($n = 4$), *word* ($n = 54$), *modality* ($n = 2$), and *condition* ($n = 5$). Unlike Experiment I, the latter variable included a control in which 60 participants—30 per modality—relied on all acoustic cues of prominence. Additionally, the ‘gold standard’ provided by the marks of two trained listeners was also included in the comparison, so as to have a reference for the marks given by participants in the control condition, Exp0.

Together with these variables, each gesture phase was introduced separately into the model in order to indicate their co-occurrence with the words receiving a mark of prominence: *preparation* ($n = 2$), *stroke* ($n = 2$), *apex* ($n = 2$), *hold* ($n = 2$), *recoil* ($n = 2$). The gesture phase *retraction* was not included because no occurrences of this phase were found in the four target sentences.

Finally, as independent continuous variables, *fundamental frequency*, *intensity*, *spectral balance*, and *duration* were fed into the model. These continuous variables included the acoustic values of the 54 words available for marking (see § 3.3.6 for details on the conducted measurements). The continuous variables were standardised as z -scores per sentence, since the acoustic values made participants give a mark of prominence relative to the phrasal environment in which words were uttered. The variable *fundamental frequency* was additionally standardised per speaker to avoid bias in pitch due to gender differences (see Appendix B2, B3, and B4 for their respective distributions).

Fixed effects were declared in model *G1* with *modality* predicting *prominence* through an interaction with the gestures phases *preparation*, *stroke*, *apex*, *hold*, *recoil*, and the standardised acoustic values of *ff* (*fundamental frequency*), *intensity*, *spectral balance* and *duration*. The decision to declare this as the initial model was motivated by the results obtained in Experiment I (§ 4.3.2) and the research questions. The predictors *participant gender* and *musical training* were also added in this first model.

Random effects were declared with intercepts through the notation ($1 | \dots$) for both participants ($1 | \textit{participant}$) and words ($1 | \textit{speaker/sentence/word}$). By-item random effects were declared as the nested variables of *word* within *sentence* within *speaker*, as previously discussed (§ 3.2.1, Figure 26c). In the formula notation of the *lme4* package developed for R (2018) by Bates et al. (2015b), model *G1* was declared as:

```
prominence ~ modality * (condition + preparation + stroke + apex + hold + recoil +  
+ z_intensity + z_H1H2 + z_ff + z_duration + musical_training + partic_gender) +  
+ (1|participant) + (1|speaker/sentence/word)
```

The set of resulting models was compared using the Akaike Information Criterion *AIC* in R (2018) through the package *AICcmodavg* (Mazerolle, 2017). The Akaike Information Criterion serves as an ordinal score to compare the quality of the model. The *AIC* computes a correction based on the number of estimable parameters, a so-called ‘penalisation for complexity’, and the *AIC* value has no meaning of its own except as a way to rank models (Akaike, 1973; Burnham & Anderson, 2002) (§ 3.2.4). Model building proceeded from models *G1* to *G13* by removing progressively different non-significant predictors, and the *AIC* value of models decreased accordingly (Table 28). From these, *G11* yielded the lowest value (*AIC* = 9940.89), followed by *G10* (*AIC* = 9941.55) and *G12* (*AIC* = 9941.78), both within 2 *AIC* points from the top-ranked model (Table 28).

In the subsequent models, from *G13* to *G21* different random structures were declared, in which the upper level *speaker* of the nested by-item random effect was removed, as in model *G13*. Thus models *G14*, *G15*, and *G16* were extensions of models *G10*, *G11*, and *G12*, respectively, in which slopes for the effects of modality were introduced in by-items random effects. Then, models *G17* and *G15*, their only difference being the declaration of random slopes in *G18* for the effect of the variable *condition*, with an increase of the *AIC* value. In model *G18*, the effect of the predictor *gesture-preparation* was removed, and in the last two models, the predictor *z-duration* was removed from the interaction with *modality*. Finally the minimal adequate model from the set of estimated models was *G17*, with an *AIC* value of 9929.41 (w_i 0.73) and random slopes for the effect of *modality* on by-item random effects.

Model	Prominence as a function of (-) fixed effects, with by-subject and by-item random effects	LogLikelihood	K	AIC	Δ_i	w_i
G1	modality x (condition + prep + stroke + apex + hold + recoil + retraction + z-ff + z-intensity + z-H1H2 + z-duration) + (1 participant) + (1 speaker/sentence/word)	-4942.24	32	9948.49	19.08	0.00
G2	modality x (condition + prep + stroke + apex + hold + recoil + z-ff + z-intensity + z-H1H2 + z-duration) + recoil + (1 participant) + (1 speaker/sentence/word)	-4944.39	31	9950.78	21.37	0.00
G3	modality x (condition + prep + stroke + apex + hold + z-ff + z-intensity + z-duration) + recoil + z-H1H2 + (1 participant) + (1 speaker/sentence/word)	-4944.43	30	9948.86	19.45	0.00
G4	modality x (condition + prep + stroke + apex + hold + z-ff + z-H1H2 + z-duration) + recoil + z-intensity + (1 participant) + (1 speaker/sentence/word)	-4944.56	30	9949.13	19.72	0.00
G5	modality x (condition + prep + stroke + apex + hold + z-ff + z-duration) + recoil + z-H1H2 + z-intensity + recoil + (1 participant) + (1 speaker/sentence/word)	-4944.60	29	9947.21	17.80	0.00
G6	modality x (condition + stroke + apex + hold + z-duration) + prep + recoil + z-H1H2 + z-intensity + (1 participant) + (1 speaker/sentence/word)	-4944.94	28	9945.87	16.46	0.00
G7	modality x (condition + stroke + apex + hold + z-duration) + prep + recoil + z-H1H2 + z-intensity + z-ff + (1 participant) + (1 speaker/sentence/word)	-4945.28	27	9944.55	16.46	0.00
G8	modality x (condition + stroke + hold + z-duration) + prep + apex + recoil + z-H1H2 + z-intensity + z-ff + (1 participant) + (1 speaker/sentence/word)	-4945.91	26	9943.83	14.41	0.00
G9	modality x (condition + stroke + z-duration) + prep + apex + hold + recoil + z-H1H2 + z-intensity + z-ff + (1 participant) + (1 speaker/sentence/word)	-4946.70	25	9943.39	13.98	0.00
G10	modality x (condition + stroke + z-duration) + prep + apex + hold + z-H1H2 + z-intensity + z-ff + (1 participant) + (1 speaker/sentence/word)	-4947.78	24	9941.55	12.14	0.00
G11	modality x (condition + stroke + z-duration) + prep + apex + hold + z-H1H2 + z-ff + (1 participant) + (1 speaker/sentence/word)	-4947.44	23	9940.89	11.48	0.00
G12	modality x (condition + stroke + z-duration) + apex + hold + z-H1H2 + z-ff + (1 participant) + (1 speaker/sentence/word)	-4948.89	22	9941.78	12.37	0.00

(a) Summary of AIC results for the GLMM random-intercept models.

Model	Prominence as a function of (-) fixed effects, with by-subject and by-item random effects	LogLikelihood	K	AIC	Δ_i	w_i
G13	modality x (condition + stroke + z-duration) + apex + hold + z-H1H2 + z-ff + (1 participant) + (1 sentence/word)	-4948.89	21	9939.79	10.37	0.00
G14	modality x (condition + stroke + z-duration) + prep + apex + hold + z-H1H2 + z-intensity + z-ff + (1 participant) + (modality sentence/word)	-4940.91	27	9935.94	12.14	0.00
G15	modality x (condition + stroke + z-duration) + prep + apex + hold + z-H1H2 + z-ff + (1 participant) + (modality sentence/word)	-4941.66	26	9935.52	6.11	0.03
G16	modality x (condition + stroke + z-duration) + apex + hold + z-H1H2 + z-ff + (1 participant) + (modality sentence/word)	-4942.84	25	9935.68	12.37	0.00
G17	modality x (condition + stroke) + prep + apex + hold + z-H1H2 + z-ff + z-duration + (1 participant) + (modality sentence/word)	-4915.71	49	9929.41	0.00	0.73
G18	modality x (condition + stroke) + prep + apex + hold + z-H1H2 + z-ff + z-duration + (1 participant) + (condition + modality sentence/word)	-4905.41	61	9932.82	3.41	0.13
G19	modality x (condition + stroke) + apex + hold + z-H1H2 + z-ff + z-duration + (1 participant) + (modality sentence/word)	-4942.84	25	9935.68	12.37	0.00
G20	modality x (condition) + prep + stroke + apex + hold + z-H1H2 + z-ff + z-duration + (1 participant) + (modality sentence/word)	-4943.24	24	9934.47	5.06	0.06
G21*	modality x (condition) + prep + stroke + apex + hold + textit-z-H1H2 + z-ff + z-duration + (1 participant) + (condition + modality sentence/word)	-4904.14	82	9972.29	42.88	0.00

(b) Summary of AIC results for the GLMM random-slope models.

Table 26: Summary of AIC results for the GLMMs relating the marking of prominence by participants as a function of predictors declared as fixed effects.

All models include by-subject and by-item random effects. Table (a) shows random-intercept models; Table (b) shows different structures in random effects. K indicates the estimated parameters in the model. The statistics associated to each model are Akaike Information Criterion (AIC), increase (Δ_i) of each model respect to the minimum AIC value, the log-likelihood of the model, and the Akaike weight (w_i) for each candidate model to be the minimal adequate model. Models labelled with an asterisk failed to converge.

Details of minimal adequate model G17

Model G17 was initially checked for overdispersion with a negative result ($\Phi_{\text{Pearson}} = 0.83, p > .05$; see Appendix B5a for details). The estimates for the predictors of G17 revealed that marking words as prominent was determined by the acoustic correlates of f_0 and duration, so that one standard deviation of f_0 ($SD = 29.19$ Hz) and of duration ($SD = 0.063$ s) increased by 1.29 ($\beta = 0.25, SE = 0.08, z = 3.05, p < .01$) and 2.02 ($\beta = 0.70, SE = 0.14, z = 4.92, p < .001$) the odds of words to be considered prominent, respectively. Intensity and spectral balance did not show to contribute significantly in the markings of prominence made by participants (Figure 48, Table 27).

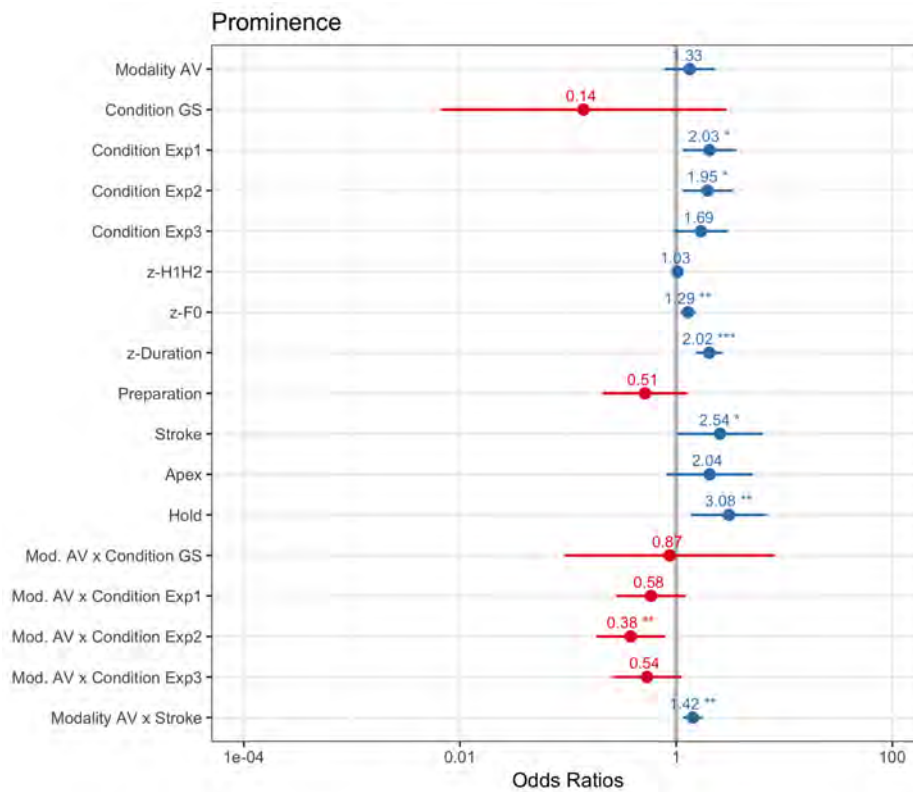


Figure 47: Forest plot showing odds ratios for main effects and interactions predicting prominence in G17 (AIC = 9929.41). For OR < 1, effect size equals 1/OR. Error bars are 95% CI.

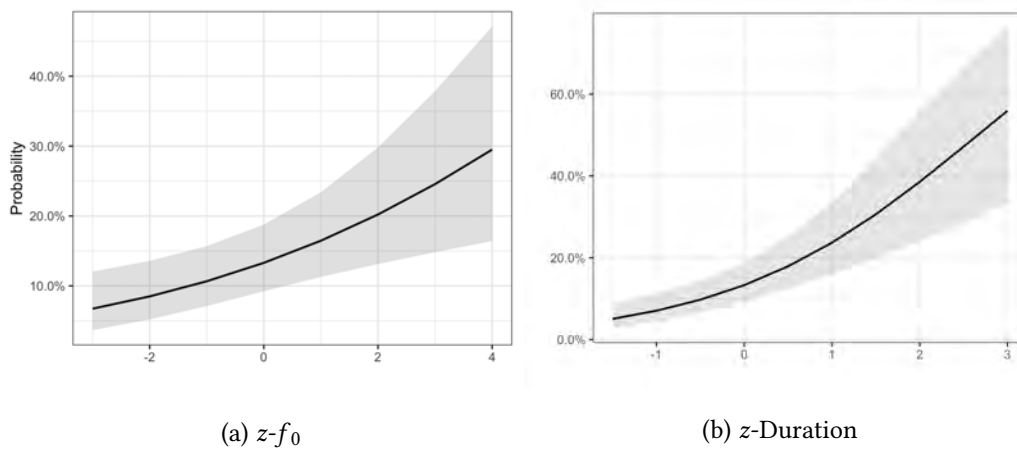
The control condition served as the baseline against which to compare all experimental conditions. Additionally, the marks given by two trained listeners, which were used as ‘gold standard’, were additionally used to assess the markings of the 60 participants belonging to the control group. Firstly, no significant differences were found between the control and the ‘gold standard’ ($\beta = -1.97$, $SE = 1.55$, $z = -1.27$, $p > .05$). Secondly, in both Exp1 and Exp2 conditions—which had intensity and duration, and f_0 and duration as their respective acoustic cues—words were 2.03 times ($\beta = 0.70$, $SE = 0.29$, $z = 2.43$, $p < .05$) and 1.95 times ($\beta = 0.66$, $SE = 0.27$, $z = 2.47$, $p < .05$) more likely to be marked as prominent respect to the control condition. This was not the case, however for Exp3, where only duration was available among the acoustic cues, although the results showed that it fell short of significance ($\beta = 0.52$, $SE = 0.29$, $z = 2.43$, $p = .07$).

Prominence marks tended significantly to coincide with the gesture phases of *stroke* and *hold*, so that words coinciding with a stroke were 2.54 more likely to be marked as prominent ($\beta = 0.93$, $SE = 0.46$, $z = 2.00$, $p < .05$), while the odds for those words coinciding with a hold increased up to 3.08 ($\beta = 1.12$, $SE = 0.40$, $z = 2.74$, $p < .01$). No other gesture phase—not even apexes—revealed any significant differences.

Experiment II

Predictor	β (SE)	p	95% CI for odds ratio		
			Lower limit	Odds ratio	Upper limit
Intercept	-3.13 (0.44)	< .001***	0.02	0.04	0.10
Modality (0=A, 1=AV)	0.28 (0.27)	.29	0.78	1.33	2.28
Condition GS	-1.97 (1.55)	.20	0.01	0.14	10.61
Condition Exp1	0.70 (0.29)	< .05*	1.15	2.03	3.59
Condition Exp2	0.66 (0.27)	< .05*	1.15	1.95	3.31
Condition Exp3	0.52 (0.29)	.07	0.95	1.69	2.99
z -H1H1 (standardized f_0)	0.03 (0.05)	.61	0.92	1.03	1.16
z -F0 (standardized f_0)	0.25 (0.08)	< .01**	1.09	1.29	1.51
z -Duration (standardized duration)	0.70 (0.14)	< .001***	1.53	2.02	2.68
Preparation (0=no, 1=yes)	-0.66 (0.46)	.14	0.21	0.51	1.27
Stroke (0=no, 1=yes)	0.93 (0.46)	< .05*	1.02	2.54	6.31
Apex (0=no, 1=yes)	0.71 (0.46)	.12	0.81	2.04	5.09
Hold (0=no, 1=yes)	1.12 (0.40)	< .01**	1.38	3.08	6.86
Modality AV x Condition GS	-0.14 (1.14)	.90	0.09	0.87	8.16
Modality AV x Condition Exp1	-0.53 (0.37)	.15	0.58	0.28	1.21
Modality AV x Condition Exp2	-0.96 (0.37)	< .05*	0.18	0.38	0.79
Modality AV x Condition Exp3	-0.62 (0.37)	.09	0.26	0.54	1.14
Modality AV x Stroke	0.35 (0.10)	< .01**	1.15	1.42	1.76

Table 27: Results of fixed effects in model G17 ($AIC = 9929.41$) predicting the marking of prominence from the variables modality, condition, fundamental frequency, duration, gesture-stroke, gesture-apex, gesture-hold, and gesture-recoil.



Experiment II

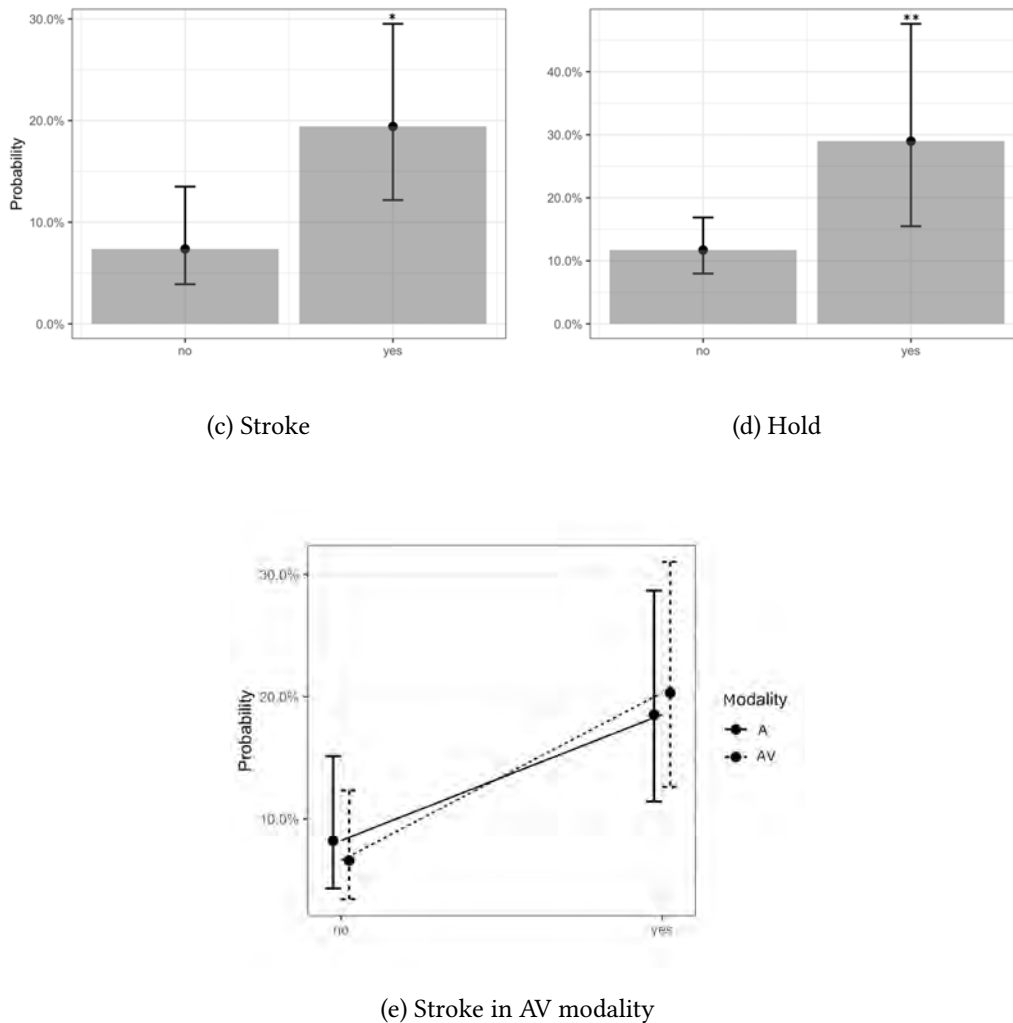


Figure 48: Main effects of minimal adequate model G17 ($AIC = 9929.41$) are displayed in graphs (a), (b), (c), and (d). Graph (e) shows an interaction between modality and stroke.

As for the overall role of the audiovisual modality, no differences were found in the marks given by participants compared to the audio-only modality. This was also the case when experimental conditions were compared to the control condition except in Exp2, where marking for prominence was 0.38 times less likely to occur than in the audiovisual modality ($\beta = -0.96$, $SE = 0.37$, $z = -2.59$, $p < .01$). This difference had already proved significant in the chi-square test (see

Figure 45a). Finally, words that coincided with a stroke were 1.42 more likely to be perceived as prominent in the audiovisual modality ($\beta = 0.35$, $SE = 0.11$, $z = 3.21$, $p < .01$), but no other interactions were found between the audiovisual modality and the visual cues of prominence.

Summary

In this first analysis, a set of models was built with the aim of comparing the markings provided by participants across conditions and modalities in a similar way as done in Experiment I (§ 4.3.2.3). The minimal adequate model, *G17*, was chosen from a set of 21 models, and it revealed that participants tended to mark more words as prominent when they lacked some of the acoustic cues of prominence. Additionally, the control group, relying on all acoustic cues, did not perform differently from two trained listeners whose marks served as a ‘gold standard’ and against which their performance was compared.

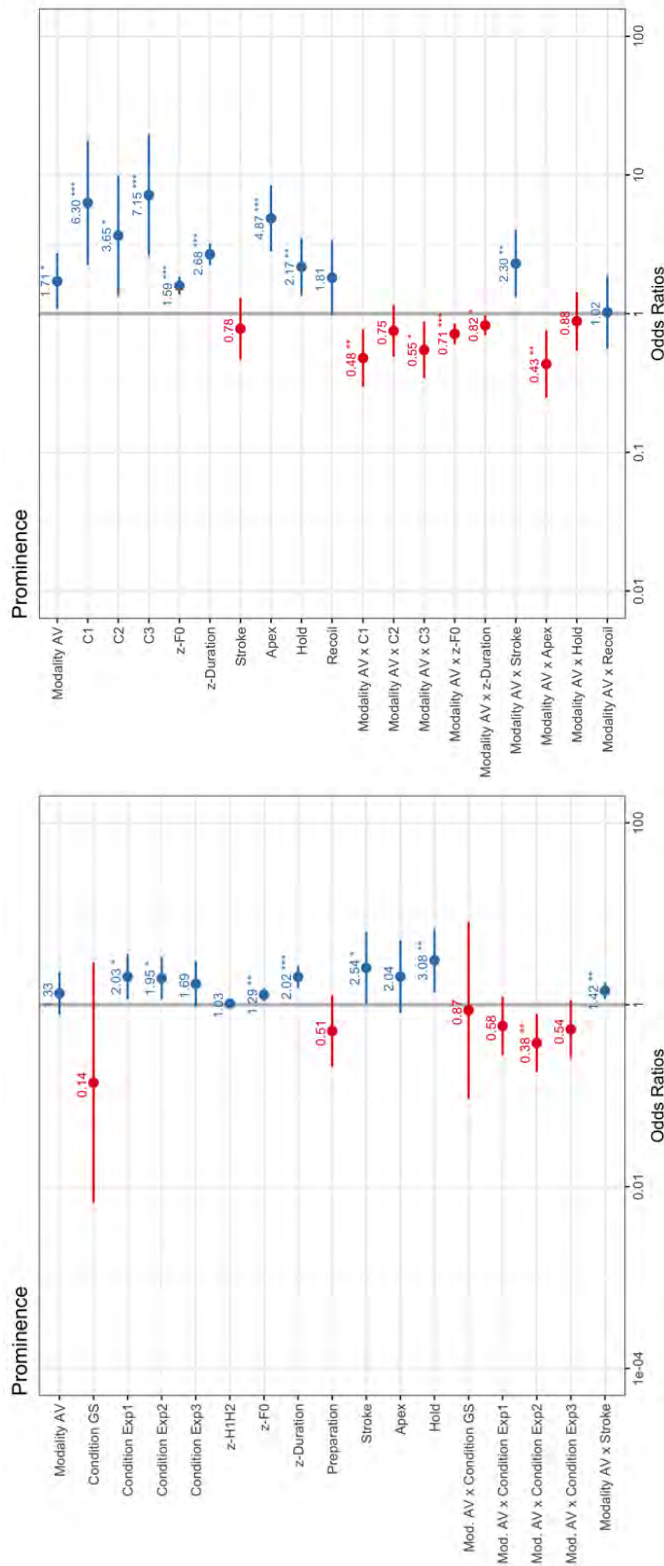
Although the chi-square test did not show any significant difference in the number of marks given to all three conditions when compared to the control, model *G17* made evident that a trend exists for all conditions to increase the probability of words to be marked as prominent. This trend proved significant in Exp1 and Exp2, but not in Exp3, which fell short of significance (Table 27). This reflects a similar pattern as that observed in Experiment I (§ 4.3.2.4), where participants tended to ‘overmark’ words in all experimental conditions. Similarly, the audiovisual modality showed an opposite trend when some acoustic cues of prominence were absent in the experimental conditions, although model *G17* only showed significant differences for Exp2.

As for the acoustic correlates of prominence, model *G17* shows that both f_0 and duration played a major role in the marks given by participants. However, none of these two acoustic correlates showed any significant difference in the au-

audiovisual modality, contrary to what was observed in Experiment I. Additionally, model *G17* shows that, overall, words given a mark of prominence significantly coincided with the gesture phases of stroke and hold, but not apex. In Experiment I, conversely, participants had marked words coinciding with apex and holds but not strokes (see Figure 49 for a comparison). It is possible that the hold phase of gestures offer a reliable cue of prominence under adverse acoustic conditions, although the different procedure in each experiment, i.e. the number of times participants were allowed to receive the stimulus and the subsequent processes of logical inference, might also have had an influence on the differences for strokes and apexes observed in both experiments. However, in both experiments strokes seemed to significantly increase the number of marks of prominence given to the words they co-occur with in the audiovisual modality. This is not the case for the apex phase of gestures, which seemed to reduce this probability in Experiment I, although no such difference was observed in Experiment II.

Unlike Experiment I, where it was not possible to obtain details of the variables predicting prominence perception in each experimental condition, the between-subjects design of Experiment II allowed to conduct analyses separately for each of them in order to better understand the role of the factors involved in the multimodal perception of acoustic prominence.

Experiment II



(a) G17 in Experiment II

(b) M18 in Experiment II

Figure 49: Comparison of odds ratios for variables predicting prominence in the top-ranked models of (a) Experiment II and (b) Experiment I.

5.3.2.4 Analysis B: Control condition ‘Exp0’

In this analysis, a set of models was built with the responses given by the 60 participants that, taking part in the experiment, provided marks of prominence for either the audio-only or the audiovisual modality serving as control condition. They relied on all acoustic cues of prominence. The dependent variable *prominence* was modelled as done previously (§ 4.3.2.3 and § 5.3.2.3). An interaction was declared between modality and the remaining variables: *condition*, the gesture phases, and the standardized acoustic values of f_0 , *intensity*, *H1H2*, and *duration*. Additionally, this time also the variables *musical-training* and *participant gender* were initially included in the global model:

```
prominence ~ modality * (condition + preparation + stroke + apex + hold + recoil +  
+ z_intensity + z_H1H2 + z_ff + z_duration + musical_training + partic_gender) +  
+ (1|participant) + (1|speaker/sentence/word)
```

Model building proceeded initially from model g_01 to model g_014 by removing non-significant predictors. Models g_012 , g_013 , and g_014 achieved a very similar *AIC* value. These three models were extended, and random slopes were declared for each of them from g_015 to g_021 . The minimal adequate model was model g_020 ($AIC = 2334.54$), the extension of model g_014 , in which slopes for the effect of *modality* on by-item random effects had been declared. However, when slopes were estimated for the effect of modality on participants, the *AIC* value increased. Finally, a competitive model, g_016 reached a close value ($AIC = 2235.76$), and although two more models g_013 and g_014 also showed close values, they were not further considered since they did not include slopes for by-item random effects (Table 28)

Model	Prominence as a function of (-) fixed effects, with by-subject and by-item random effects	LogLikelihood	K	AIC	Δ_i	w_i
$g_{0,1}$	modality x (prep + stroke + apex + hold + recoil + z-ff + z-intensity + z-H1H2 + z-duration + musical-training + participant-gender) + (1 participant) + (1 speaker/sentence/word)	-1150.42	30	2360.85	26.31	0.00
$g_{0,2}$	modality x (prep + stroke + apex + recoil + z-ff + z-intensity + z-H1H2 + z-duration + musical-training) + hold + participant-gender + (1 participant) + (1 speaker/sentence/word)	-1150.44	28	2356.88	22.34	0.00
$g_{0,3}$	modality x (prep + stroke + apex + recoil + z-ff + z-intensity + z-H1H2 + z-duration) + hold + musical-training + participant-gender + (1 participant) + (1 speaker/sentence/word)	-1151.89	26	2355.78	21.24	0.00
$g_{0,4}$	modality x (prep + stroke + apex + recoil) + hold + z-ff + z-intensity + z-H1H2 + z-duration + musical-training + participant-gender + (1 participant) + (1 speaker/sentence/word)	-1153.18	22	2350.35	15.82	0.00
$g_{0,5}$	modality x (prep + stroke + apex + recoil) + hold + z-ff + z-intensity + z-H1H2 + z-duration + (1 participant) + (1 speaker/sentence/word)	-1153.93	19	2345.85	11.32	0.00
$g_{0,6}$	modality x (stroke + apex + recoil) + prep + hold + z-ff + z-intensity + z-H1H2 + z-duration + (1 participant) + (1 speaker/sentence/word)	-1154.24	18	2344.47	9.93	0.00
$g_{0,7}$	modality x (stroke + apex + recoil) + prep + hold + z-ff + z-H1H2 + z-duration + (1 participant) + (1 speaker/sentence/word)	-1154.24	17	2342.47	7.93	0.01
$g_{0,8}$	modality x (stroke + apex + recoil) + hold + z-ff + z-H1H2 + z-duration + (1 participant) + (1 speaker/sentence/word)	-1154.29	16	2340.59	6.05	0.02
$g_{0,9}$	modality x (stroke + apex) + hold + recoil + z-ff + z-H1H2 + z-duration + (1 participant) + (1 speaker/sentence/word)	-1154.53	15	2339.07	4.53	0.03
$g_{0,10}$	modality x (stroke + apex) + hold + recoil + z-ff + z-duration + (1 participant) + (1 speaker/sentence/word)	-1155.67	14	2339.34	4.80	0.03
$g_{0,11}$	modality x (stroke + apex) + hold + z-ff + z-duration + (1 participant) + (1 speaker/sentence/word)	-1156.80	13	2339.59	5.05	0.03
$g_{0,12}$	modality x (stroke) + apex + hold + z-ff + z-duration + (1 participant) + (1 speaker/sentence/word)	-1157.03	12	2338.05	3.52	0.05
$g_{0,13}$	modality x (stroke) + apex + z-ff + z-H1H2 + z-duration + (1 participant) + (1 speaker/sentence/word)	-1155.21	13	2336.42	1.89	0.12
$g_{0,14}$	modality x (stroke) + apex + hold + z-ff + z-H1H2 + z-duration + (1 participant) + (1 speaker/sentence/word)	-1155.21	13	2336.42	1.89	0.12

(a) Summary of AIC results for the GLMM random-intercept models.

Model	Prominence as a function of (-) fixed effects, with by-subject and by-item random effects	LogLikelihood	K	AIC	Δ_i	w_i
$g_0 15$	modality x (stroke) + apex + hold + z-ff + z-duration + (1 participant) + (modality speaker/sentence/word)	-1152.54	18	2341.08	6.55	0.05
$g_0 16$	modality x (stroke) + apex + hold + z-ff + z-duration + (1 participant) + (modality sentence/word)	-1152.88	15	2335.76	1.23	0.17
$g_0 17$	modality x (stroke) + apex + z-ff + z-H1H2 + z-duration + (1 participant) + (modality speaker/sentence/word)	-1157.83	13	2351.66	21.24	0.00
$g_0 18^*$	modality x (stroke) + apex + z-ff + z-H1H2 + z-duration + (1 participant) + (modality sentence/word)	-1158.45	15	2346.9	12.36	0.08
$g_0 19$	modality x (stroke) + apex + hold + z-ff + z-H1H2 + z-duration + (1 participant) + (modality speaker/sentence/word)	-1150.98	19	2339.97	5.43	0.02
$g_0 20$	modality x (stroke) + apex + hold + z-ff + z-H1H2 + z-duration + (1 participant) + (modality sentence/word)	-1151.27	16	2334.54	0.00	0.31
$g_0 21$	modality x (stroke) + apex + hold + z-ff + z-H1H2 + z-duration + (modality participant) + (modality sentence/word)	-1155.21	13	2337.439	2.85	0.08

(b) Summary of AIC results for the GLMM random-slope models.

Table 28: Summary of AIC results for the GLMMs in condition Exp0. The marking of prominence by participants is modelled as a function of predictors declared as fixed effects. All models include by-subject and by-item random effects. Table (a) shows random-intercept models; Table (b) shows different structures in random effects. K indicates the estimated parameters in the model. The statistics associated to each model are Akaike Information Criterion (AIC), increase (Δ_i) of each model respect to the minimum AIC value, the log-likelihood of the model, and the Akaike weight (w_i) for each candidate model to be the minimal adequate model. Models labelled with an asterisk failed to converge.

Details of minimal adequate model g_020

Model g_020 did not show overdispersion ($\Phi_{\text{Pearson}} = 0.80, p > .05$; see Appendix B5b for details). It is apparent from this model that the prominence marks given by participants were strongly determined by the acoustic correlates of f_0 ($\beta = 0.45, SE = 0.13, z = 3.42, p < .001$) and duration ($\beta = 0.91, SE = 0.16, z = 5.46, p < .001$). This means that an increase of 1 standard deviation in f_0 ($SD = 36.84$ Hz) and duration ($SD = 0.064$ s) increased the odds of words to be perceived as prominent by 1.57 and 2.50 times, respectively. The acoustic cue of intensity did not prove significant, but spectral balance—measured as the difference in amplitude between H1 and H2—was close to significance ($\beta = 0.25, SE = 0.14, z = 1.79, p = .07$) (Table 29).

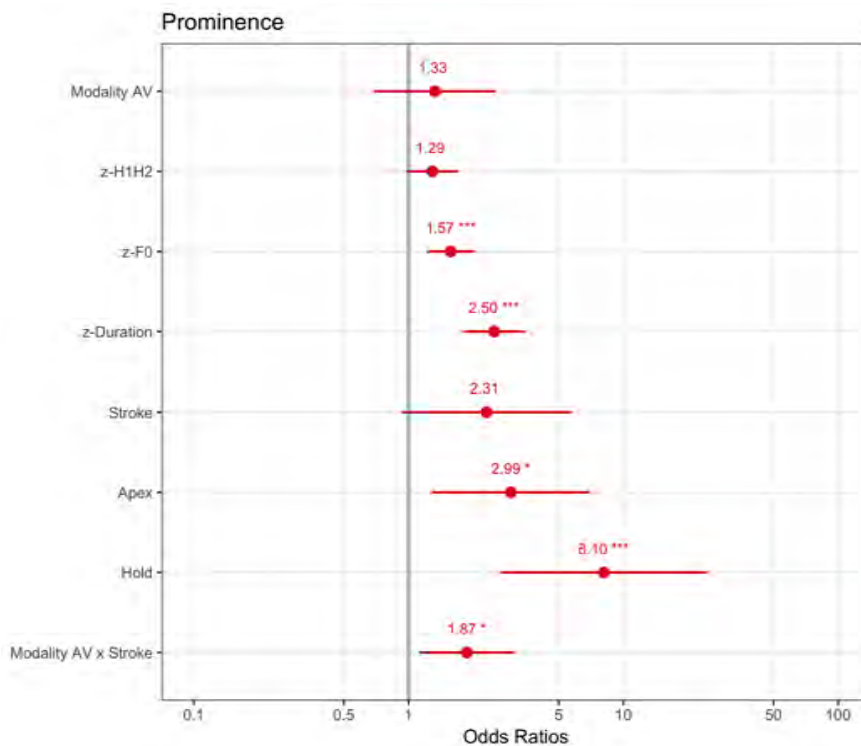


Figure 50: Forest plot showing odds ratios for main effects and interactions predicting prominence in model g_020 ($AIC = 2334.54$). Error bars are 95% CI.

Experiment II

Predictor	β (SE)	p	95% CI for odds ratio		
			Lower limit	Odds ratio	Upper limit
Intercept	-3.70 (0.44)	< .001***	0.01	0.02	0.06
Modality (0=A, 1=AV)	0.28 (0.33)	.39	0.69	1.33	2.54
z-H1H2 (standardized spectral balance)	0.25 (0.14)	.07	0.98	1.29	1.70
z-F0 (standardized f_0)	0.45 (0.13)	< .001***	1.21	1.57	2.04
z-Duration (standardized duration)	0.91 (0.16)	< .001***	1.80	2.50	3.48
Stroke (0=no, 1=yes)	0.83 (0.46)	.07	0.93	2.31	5.72
Apex (0=no, 1=yes)	1.09 (0.43)	< .05*	1.28	2.99	6.98
Hold (0=no, 1=yes)	2.09 (0.56)	< .001***	2.68	8.10	24.51
Modality AV x Stroke	0.62 (0.26)	< .05*	1.12	1.87	3.12

Table 29: Results of fixed effects in model g_{020} ($AIC = 2334.54$) predicting the marking of prominence from the variables modality, spectral balance, fundamental frequency, duration, gesture-stroke, gesture-apex, gesture-hold.

Participants marked words that significantly coincided with the apex phase ($\beta = 1.09$, $SE = 0.43$, $z = 2.53$, $p < .05$) and the hold phase of gestures ($\beta = 2.09$, $SE = 0.56$, $z = 3.70$, $p < .001$), so that words coinciding with both phases were 2.99 and 8.10, respectively, more likely to receive a mark of prominence than words not coinciding with any of these gesture phases.

No significant main effect of the predictor *modality* was observed, although an interaction of the audiovisual modality and the stroke of gestures proved significant and the odds for words to be perceived as prominent increased by 1.87. No other gesture phases were found to significantly determine prominence perception in the audiovisual modality.

Experiment II

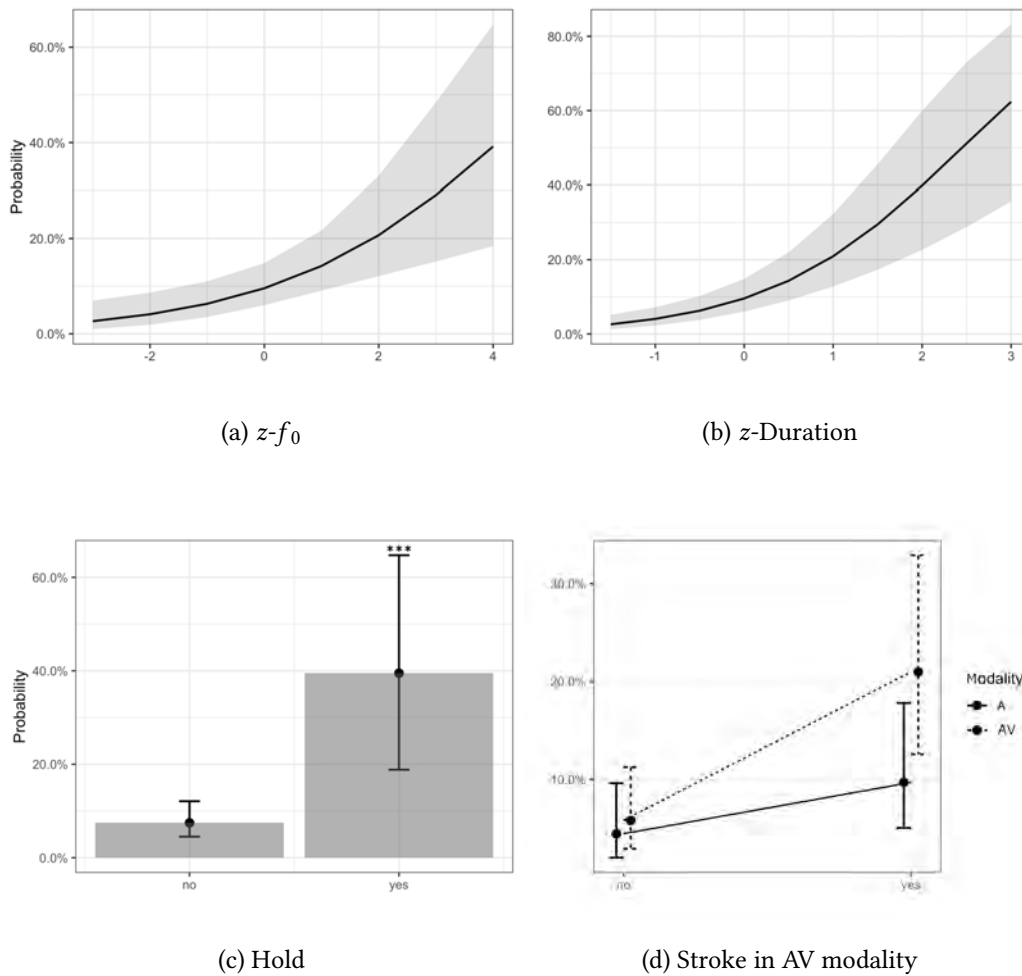


Figure 51: Graphs (a), (b), and (c) show main effects in model g_020 ($AIC = 2334.54$). Graph (d) shows an interaction between modality and stroke.

A competitive model, g_016 ($AIC = 2335.76$), reached an AIC value 1.23 Δ -points worse than g_020 . This model omitted the non-significant predictor $z-H1H2$, and as a result the estimated parameter *stroke* reached significance ($\beta = 0.93$, $SE = 0.46$, $z = 2.02$, $p < .05$) (Table 30).

Experiment II

Predictor	β (SE)	p	95% CI for odds ratio		
			Lower limit	Odds ratio	Upper limit
Intercept	-3.65 (0.44)	< .001***	0.01	0.03	0.06
Modality (0=A, 1=AV)	0.31 (0.33)	.35	0.71	1.36	2.62
z-F0 (standardized f_0)	0.48 (0.13)	< .001***	1.26	1.63	2.11
z-Duration (standardized duration)	0.95 (0.16)	< .001***	1.87	2.59	3.59
Stroke (0=no, 1=yes)	0.93 (0.46)	< .05*	1.03	2.54	6.29
Apex (0=no, 1=yes)	1.05 (0.43)	< .05*	1.24	2.87	6.67
Hold (0=no, 1=yes)	1.84 (0.54)	< .001***	2.18	6.32	18.28
Modality AV x Stroke	0.58 (0.26)	< .05*	1.06	1.79	3.02

Table 30: Results of fixed effects in model g_{016} (AIC = 2335.76), which predicts the marking of prominence from the same variables as the minimal adequate model g_{020} , but it omits z-H1H2, causing the predictor stroke to reach significance.

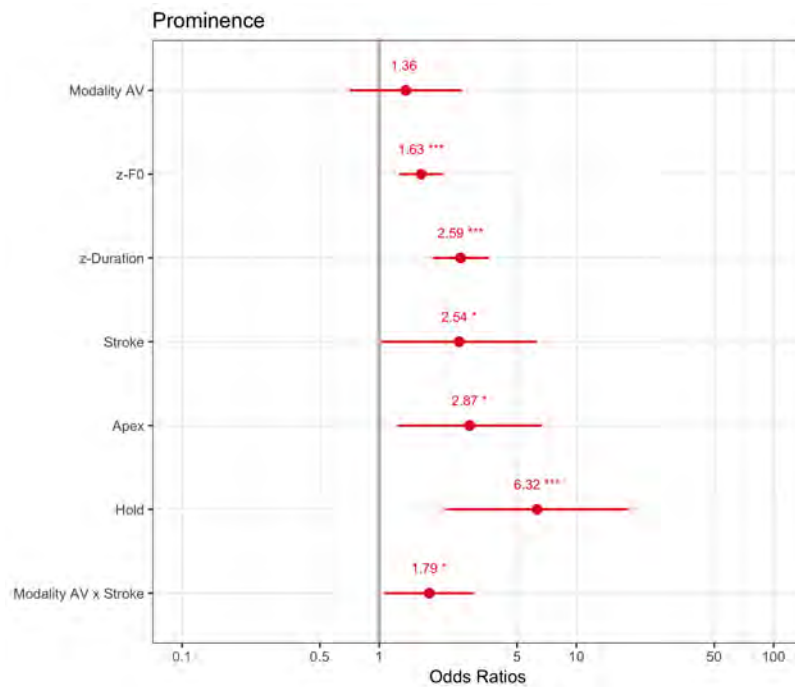


Figure 52: Odds ratios for main effects and interactions predicting prominence in $Exp0$ for competitive model g_{016} . For OR < 1, effect size equals 1/OR. Error bars are 95% CI.

Summary

The set of models built to analyse the potential of variables for predicting prominence marking provided a minimal adequate model g_{020} , which was fol-

lowed by model $g_0 16$ within 2 Δ -points. The minimal adequate model differed from its competitor in the non-significance of its predictors $z\text{-}H1H2$ and *stroke*. Nevertheless, *stroke* proved significant in the second-best model $g_0 16$ when $z\text{-}H1H2$ was removed. In the results section both models were reported, as suggested by T. W. Arnold (2010), but only the minimal adequate model was discussed, since it achieved a best *AIC* value despite including more predictors (see § 3.2.4).

Model $g_0 20$ revealed that participants relying on all acoustic cues of prominence used f_0 and duration to identify prominent words, but not intensity. Rather, they relied more on spectral balance, although this fell short of significance ($p = .07$). The marks of prominence provided by participants in this control condition significantly coincided with both the gesture phase of apex and, especially, with the gesture phase of hold. Interestingly, also strokes were found significant in the second-best model $g_0 16$ (Table 30).

Additionally, differences in agreement among participants showed that the audiovisual modality was more challenging than the audio-only modality to judge prominence. In this sense, participants got closer in their ratings to the ‘gold standard’ when they had to mark prominent words auditorily rather than visually (Figure 22). Nevertheless, although no differences were observed for the overall effect of the audiovisual modality on the marks of prominence, participants gave more marks to words coinciding with the stroke phase of gestures when the visual cues of prominence were available to them. This difference is in line with the significant difference between modalities that the chi-square test revealed for strokes (Figure 46a). Interestingly, even if the chi-square test conducted for apexes also showed a similar difference between modalities (Figure 46b), the minimal adequate model did not revealed any significant difference between modalities for apexes.

5.3.2.5 Analysis C: First condition ‘Exp1’ (intensity and duration)

In the first experimental condition, Exp1, the acoustic cue of f_0 had been neutralised. As a result, the intonation curve of the manipulated stimuli in this condition was kept within a 2-semitone range between the lowest and the highest f_0 values in each sentence. A total of 60 participants (30 per modality) marked words for prominence.

In the analysis of the collected data, the inclusion of variables predicting the dependent variable *prominence* was initially motivated by previous analyses (§ 4.3.2.3, § 5.3.2.3, and § 5.3.2.4). An interaction of *modality* was declared with all gesture phases and the acoustic cues of prominence. The variables *musical-training* and *participant gender* were also included as predictors in the global model:

```
prominence ~ modality * (condition + preparation + stroke + apex + hold + recoil +  
+ z_intensity + z_H1H2 + z_ff + z_duration + musical_training + partic_gender) +  
+ (1|participant) + (1|speaker/sentence/word)
```

The set of models were initially built by removing non-significant predictors from the interaction. The random-intercept models $g_1 9$ and $g_1 10$ reached the lowest *AIC* value, differing in only one parameter: *z-H1H2*. Later, model $g_1 11$ did not yield a lower value when the predictor *apex* was removed from the interaction.

In the subsequent models, from $g_1 12$ to $g_1 17$, different random structures were explored for the two models with the lowest *AIC* value. As a result, removal of the upper level *speaker* from the nested by-item random effect reduced by 2 Δ -points both models $g_1 9$ and $g_1 10$. The resulting $g_1 12$ and $g_1 13$ were extended by including random slopes for the effect of *modality* in models from $g_1 14$ to $g_1 17$. However, none of them achieved a lower *AIC* value than that of the minimal adequate model $g_1 13$ (*AIC* = 2719.73) (Table 31).

Model	Prominence as a function of (-) fixed effects, by-subject and with by-item random effects	LogLikelihood	K	AIC	Δ_i	w_i
$g_{1,1}$	modality x (prep + stroke + apex + hold + recoil + z-ff + z-intensity + z-H1H2 + z-duration + musical-training + participant-gender) + (1 participant) + (1 speaker/sentence/word)	-1338.53	30	2737.06	17.33	0.00
$g_{1,2}$	modality x (prep + stroke + apex + hold + recoil + z-ff + z-intensity + z-H1H2 + z-duration + musical-training) + participant-gender + (1 participant) + (1 speaker/sentence/word)	-1338.80	29	2735.60	15.87	0.00
$g_{1,3}$	modality x (prep + stroke + apex + hold + recoil + z-intensity + z-ff + z-duration + musical-training) + z-H1H2 + participant-gender + (1 participant) + (1 speaker/sentence/word)	-1338.87	28	2733.73	14.00	0.00
$g_{1,4}$	modality x (prep + stroke + apex + hold + recoil + z-ff + z-duration + musical-training) + z-intensity + z-H1H2 + participant-gender + (1 participant) + (1 speaker/sentence/word)	-1339.14	27	2732.29	12.56	0.00
$g_{1,5}$	modality x (prep + stroke + apex + hold + z-duration + musical-training) + recoil + z-ff + z-intensity + z-H1H2 + participant-gender + (1 participant) + (1 speaker/sentence/word)	-1339.21	25	2728.42	8.69	0.00
$g_{1,6}$	modality x (prep + stroke + apex + z-ff + z-duration + musical-training) + hold + recoil + z-intensity + z-H1H2 + participant-gender + (1 participant) + (1 speaker/sentence/word)	-1339.21	26	2730.77	11.04	0.00
$g_{1,7}$	modality x (stroke + apex + hold + z-duration + musical-training) + prep + recoil + z-intensity + z-H1H2 + z-ff + participant-gender + (1 participant) + (1 speaker/sentence/word)	-1339.48	26	2730.96	11.23	0.00
$g_{1,8}$	modality x (stroke + apex + musical-training) + prep + hold + recoil + z-intensity + z-H1H2 + z-ff + z-duration + participant-gender + (1 participant) + (1 speaker/sentence/word)	-1341.21	22	2726.43	6.70	0.01
$g_{1,9}$	modality x (stroke + apex + musical-training) + prep + z-intensity + z-H1H2 + z-ff + z-duration + (1 participant) + (1 speaker/sentence/word)	-1342.09	19	2722.19	2.46	0.11
$g_{1,10}$	modality x (stroke + apex + musical-training) + prep + z-intensity + z-ff + z-duration + (1 participant) + (1 speaker/sentence/word)	-1342.86	18	2721.73	2.00	0.14
$g_{1,11}$	modality x (stroke + musical-training) + prep + apex + z-intensity + z-ff + z-duration + (1 participant) + (1 speaker/sentence/word)	-1345.92	17	2725.84	6.11	0.02
$g_{1,12}$	modality x (stroke + apex + musical-training) + prep + z-intensity + z-H1H2 + z-ff + z-duration + (1 participant) + (1 sentence/word)	-1342.09	18	2720.19	0.46	0.31
$g_{1,13}$	modality x (stroke + apex + musical-training) + prep + z-intensity + z-ff + z-duration + (1 participant) + (1 sentence/word)	-1342.86	17	2719.73	0.00	0.38

(a) Summary of AIC results for the GLMM random-intercept models.

Model	Prominence as a function of (-) fixed effects, with by-subject and by-item random effects	LogLikelihood	K	AIC	Δ_i	w_i
$g_{1,14}$	modality x (stroke + apex + musical-training) + prep + z-intensity + z-H1H2 + z-ff + z-duration + + (1 participant) + (modality speaker/sentence/word)	-1341.95	25	2733.89	14.16	0.00
$g_{1,15}$	modality x (stroke + apex + musical-training) + prep + z-intensity + z-H1H2 + z-ff + z-duration + + (1 participant) + (modality sentence/word)	-1342.03	22	2728.06	8.33	0.01
$g_{1,16}^*$	modality x (stroke + apex + musical-training) + prep + z-intensity + z-ff + z-duration + + (1 participant) + (modality speaker/sentence/word)	-1342.73	24	2733.46	13.73	0.00
$g_{1,17}$	modality x (stroke + apex + musical-training) + prep + z-intensity + z-ff + z-duration + + (1 participant) + (modality sentence/word)	-1342.81	21	2727.61	7.88	0.01

(b) Summary of AIC results for the GLMM random-slope models.

Table 31: Summary of AIC results for the GLMMs in condition Exp1. The marking of prominence by participants is modelled as a function of predictors declared as fixed effects. All models include by-subject and by-item random effects. Table (a) shows random-intercept models; Table (b) shows different structures in random effects. K indicates the estimated parameters in the model. The statistics associated to each model are Akaike Information Criterion (AIC), increase (Δ_i) of each model respect to the minimum AIC value, the log-likelihood of the model, and the Akaike weight (w_i) for each candidate model to be the minimal adequate model. Models labelled with an asterisk failed to converge.

Details of minimal adequate model $g_1 13$

The minimal adequate model $g_1 13$ was checked for overdispersion ($\Phi_{\text{Pearson}} = 0.79, p > .05$; see Appendix B5c for details) with similar negative results as the tests conducted for previous models. In this case, in which f_0 had been neutralised, the marking of prominence was determined by the sole acoustic correlate of duration ($\beta = 0.74, SE = 0.15, z = 4.82, p < .001$), so that an increase of 1 standard deviation ($SD = 0.064$ s) raised the odds by 2.11 for words to be considered prominent. Spectral balance did not seem to be relied on by participants, while maximum intensity of vowels contributed positively together with duration to perceive prominence, although it fell short of significance ($\beta = 0.22, SE = 0.11, z = 1.90, p = .05$) (Figure 53, Table 32).

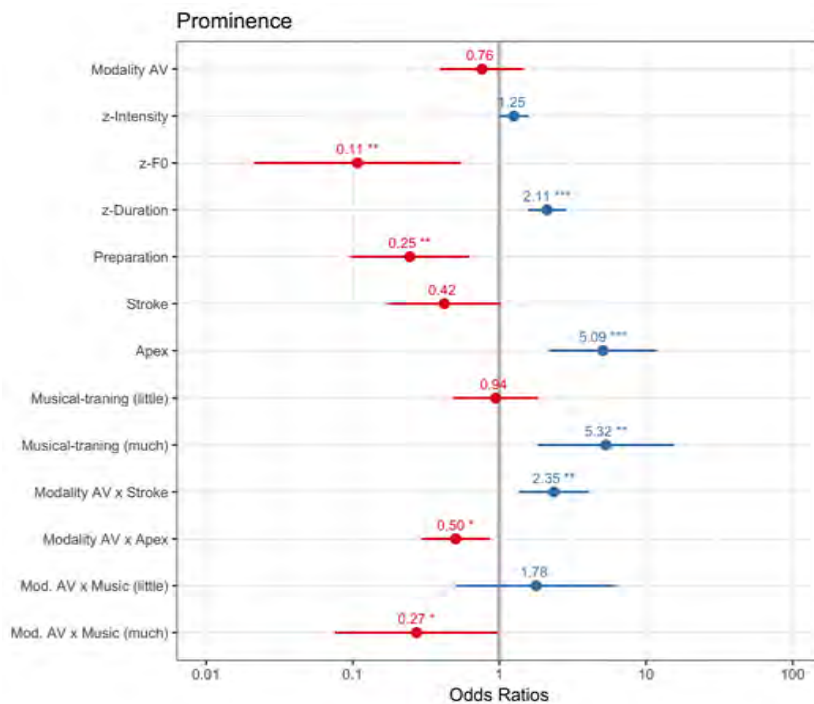


Figure 53: Forest plot showing odds ratios for main effects and interactions predicting prominence in $g_1 13$ ($AIC = 2719.73$). For $OR < 1$, effect size equals $1/OR$. Error bars are 95% CI.

Experiment II

Predictor	β (SE)	p	95% CI for odds ratio		
			Lower limit	Odds ratio	Upper limit
Intercept	-2.65 (0.50)	< .001***	0.03	0.07	0.019
Modality (0=A, 1=AV)	-0.27 (0.33)	.41	0.39	0.76	1.47
z -Intensity (standardized intensity)	0.22 (0.11)	.05	0.99	1.25	1.58
z -F0 (standardized f_0)	-2.22 (0.82)	< .01**	0.02	0.11	0.54
z -Duration (standardized duration)	0.74 (0.15)	< .001***	1.56	2.11	2.87
Preparation (0=no, 1=yes)	-1.40 (0.47)	< .01**	0.10	0.25	0.63
Stroke (0=no, 1=yes)	-0.86 (0.45)	.05	0.17	0.42	1.02
Apex (0=no, 1=yes)	1.62 (0.43)	< .001***	2.17	5.09	11.94
Musical training (little)	-0.05 (0.34)	.86	0.48	0.94	1.84
Musical training (much)	1.67 (0.54)	< .01**	1.83	5.32	15.52
Modality AV x Stroke	0.85 (0.28)	< .01**	1.36	2.35	4.07
Modality AV x Apex	-0.68 (0.27)	< .05*	0.29	0.50	0.86
Modality AV x Musical training (little)	0.57 (0.64)	.37	0.51	1.78	6.22
Modality AV x Musical training (much)	-1.30 (0.65)	< .05*	0.08	0.27	0.98

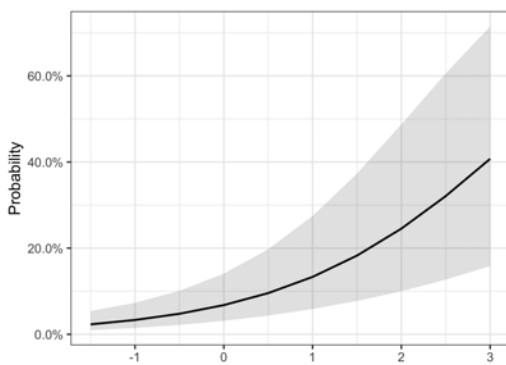
Table 32: Results of fixed effects in model g_1 13 ($AIC = 2719.73$) predicting the marking of prominence from the variables modality, intensity, fundamental frequency, duration, gesture-preparation, gesture-stroke, gesture-apex, and musical-training.

The model also showed a strong negative effect of f_0 ($\beta = -2.22$, $SE = 0.82$, $z = -2.96$, $p < .01$), so that the minimal increases in pitch perceived by participants ($SD = 0.11$ Hz) actually reduced the odds 9.09 times of marking a word as prominent. Nevertheless, participants with more than 5 years of formal musical training were 5.32 more likely to mark words as prominent than those with no musical training at all ($\beta = 1.67$, $SE = 0.54$, $z = 3.06$, $p < .01$); while no such difference was found for those with fewer than 5 years of musical training. Interestingly, in the audiovisual modality, another group of musically trained participants were 3.70 times less likely to mark words as prominent as in the audio-only modality ($\beta = -1.30$, $SE = 0.65$, $z = -1.99$, $p < .05$).

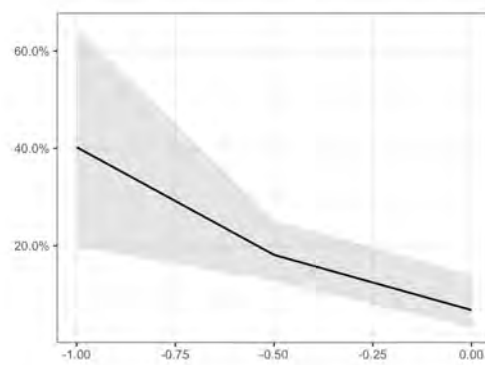
Even if participants lacked any visual cues of prominence in the audio-only modality, words coinciding with apexes were 5.09 times more likely to be given a mark of prominence ($\beta = 1.62$, $SE = 0.43$, $z = 3.74$, $p < .001$). Conversely, the odds

Experiment II

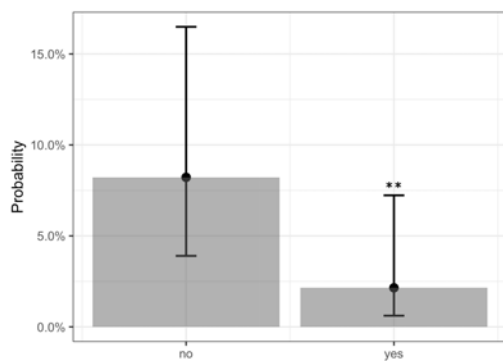
for words to be considered prominent when they coincided with the preparation phase of gestures decreased by 4. Also, a negative effect for words being marked as prominent was observed for the stroke phase of gestures, which fell short of significance ($\beta = -0.86$, $SE = 0.45$, $z = -1.91$, $p = .05$). However, when participants could rely on the visual cues of prominence available in the audiovisual modality, words coinciding with apices were 2 times less likely to be given a mark of prominence ($\beta = -0.69$, $SE = 0.27$, $z = -2.50$, $p < .05$), while strokes actually made participants increase their prominence marks 2.35 times ($\beta = 0.85$, $SE = 0.28$, $z = 3.06$, $p < .01$).



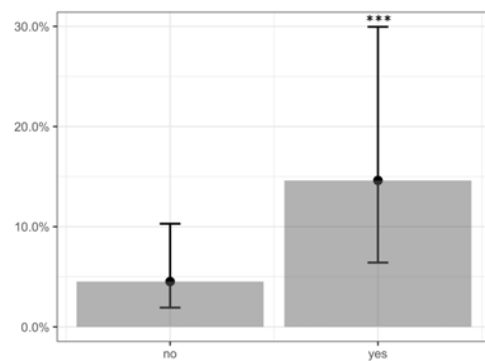
(a) z-Duration



(b) Neutralised z-f₀

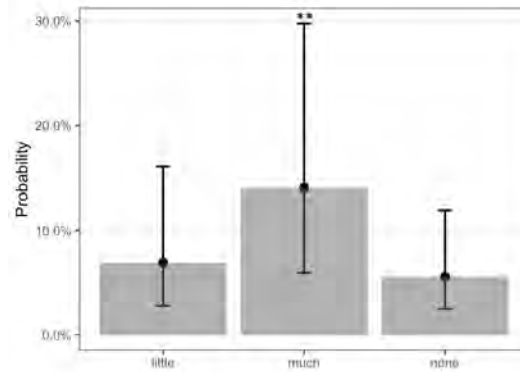


(c) Preparation

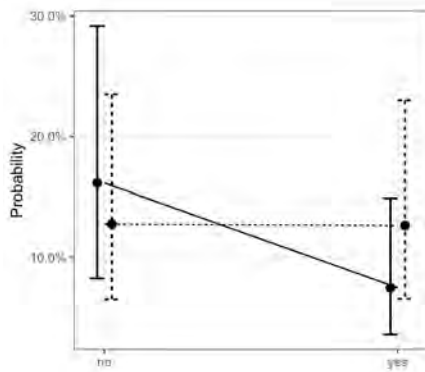


(d) Apex

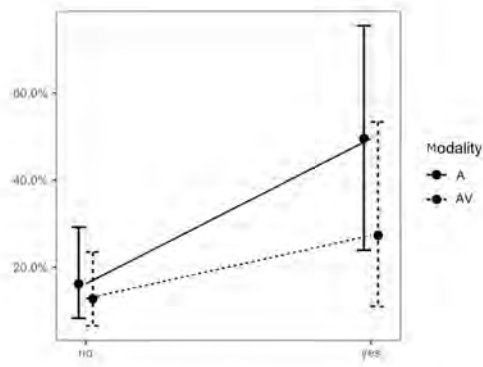
Experiment II



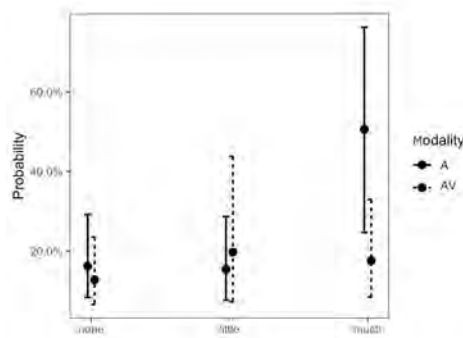
(e) Musical training



(f) Stroke in AV modality



(g) Apex in the AV modality



(h) Musical training in AV modality

Figure 54: Main effects and interactions in model g_{13} ($AIC = 2719.73$).

Experiment II

Next to the top-ranked model just described, the competitive model, g_{12} ($AIC = 2720.19$), was within 2 Δ -points and included the predictor $z\text{-H1H2}$. The coefficient estimates of this model were very similar to those of the minimal adequate model, with the exception that in this second-best model the predictor *stroke* reached significance ($\beta = -1.02$, $SE = 0.46$, $z = -2.20$, $p < .05$) (Table 33).

Predictor	β (SE)	p	95% CI for odds ratio		
			Lower limit	Odds ratio	Upper limit
Intercept	-2.57 (0.50)	< .001***	0.03	0.08	0.020
Modality (0=A, 1=AV)	-0.27 (0.33)	.76	0.39	0.76	1.47
z -Intensity (standardized intensity)	0.24 (0.11)	< .05*	1.01	1.28	1.61
z -H1H2 (standardized spectral balance)	0.22 (0.17)	.20	0.88	1.25	1.75
z -F0 (standardized f_0)	-2.38 (0.83)	< .01**	0.02	0.09	0.47
z -Duration (standardized duration)	0.74 (0.15)	< .001***	1.56	2.11	2.85
Preparation (0=no, 1=yes)	-1.48 (0.48)	< .01**	0.09	0.23	0.58
Stroke (0=no, 1=yes)	-1.02 (0.46)	< .05*	0.14	0.36	0.89
Apex (0=no, 1=yes)	1.68 (0.43)	< .001***	2.30	5.37	12.55
Musical training (little)	-0.05 (0.34)	.86	0.48	0.94	1.85
Musical training (much)	1.67 (0.54)	< .01**	1.83	5.33	15.54
Modality AV x Stroke	0.85 (0.28)	< .01**	1.36	2.36	4.07
Modality AV x Apex	-0.68 (0.27)	< .05*	0.29	0.50	0.86
Modality AV x Musical training (little)	0.57 (0.64)	.37	0.51	1.78	6.23
Modality AV x Musical training (much)	-1.30 (0.65)	< .05*	0.08	0.27	0.98

Table 33: Results of fixed effects in model g_{12} ($AIC = 2720.19$) predicting the marking of prominence from the variables modality, intensity, spectral balance, fundamental frequency, duration, gesture-preparation, gesture-stroke, gesture-apex, and musical-training.

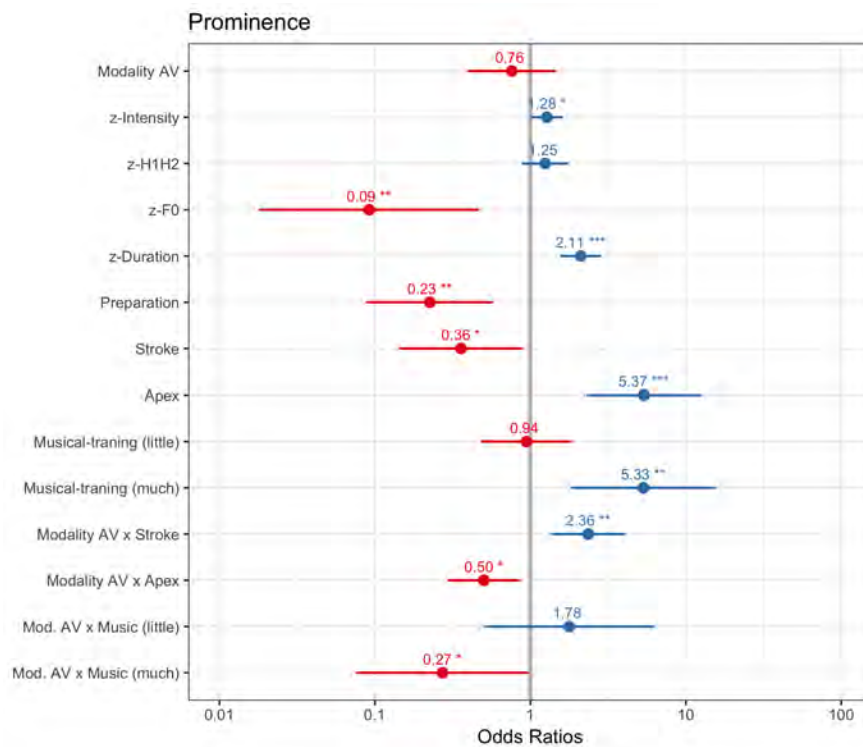


Figure 55: Odds ratios for main effects and interactions predicting prominence in Exp1 for competitive model g_{12} . For $OR < 1$, effect size equals $1/OR$. Error bars are 95% CI.

Summary

In this first experimental condition Exp1, in which f_0 had been neutralised, model building proceeded by including all variables as predictors, also *musical training* and *participant gender*. The minimal adequate model from the set of built models was model g_{13} ($AIC = 2719.73$), followed by model g_{12} ($AIC = 2720.19$) within less than 2 Δ -points. The minimal adequate model was more parsimonious than its competitor and did not include the variable $z-H1H2$ (Figure 55).

In model g_{13} , it was observed that, in the absence of f_0 as acoustic cue of prominence, syllable duration was heavily relied on by participants to give marks of prominence. Although intensity fell short of significance, it may have also

played an important role in perceiving prominence in the absence of f_0 , especially since it proved significant in model $g_1 12$ (Table 33).

Participants marked words that consistently coincided with the apex phase of gestures. However, those words coinciding with the preparation phase were much less likely to be considered prominent. Despite the fact that modality proved non-significant, the visual cues of prominence were found to influence the marks provided by participants. Surprisingly, words coincidental with apexes were considerably less likely to be perceived as prominent in the audiovisual modality. In their turn, strokes of gestures were slightly more relied on in the audiovisual modality to mark prominent words, while their overall negative effect for the words they coincided with fell short of significance ($p = .05$). Interestingly, this pattern in which strokes seem to have the lead when participants marked words in the audiovisual modality is in line with the results obtained in the previous analyses, as in Experiment I (§ 4.3.2.4, Figures 40e and 40f) and Exp0 (§ 5.3.2.4, Figure 51d), while apexes seem to have had a minor role. Similar as in Exp0, participants found more challenging to agree on the marks they gave to words in the audiovisual modality, and overall they agreed much less than in the control condition for both modalities (Figure 22).

Furthermore, maybe the most surprising result was the significance of the predictor *musical training*. In the audio-only modality participants who declared having received formal music training for more than 5 years were much more likely to give marks of prominence than participants with no musical training. However, when the visual cues of prominence were available, the performance of musically trained participants was different, and they were considerably less likely to give marks of prominence than in the audio-only modality.

5.3.2.6 Analysis D: Second condition ‘Exp2’ (f_0 and duration)

The condition in which intensity had been neutralised as an acoustic cue of prominence was carried out as an independent between-subjects experiment, as were the other experimental conditions in this Experiment II.

Similarly as in Exp1, the declaration of the global model included an interaction between the predictor *modality* and all gesture phases and the acoustic cues of prominence. The variables *musical training* and *participant gender* were also included as predictors:

```
prominence ~ modality * (condition + preparation + stroke + apex + hold + recoil +  
+ z_intensity + z_H1H2 + z_ff + z_duration + musical_training + partic_gender) +  
+ (1|participant) + (1|speaker/sentence/word)
```

In models from g_21 to g_214 non-significant predictors were progressively removed in order to reach a more parsimonious model. Initially, models g_29 ($AIC = 2235.70$) and g_210 ($AIC = 2536.10$) reached the lowest AIC value. The only difference between each other was the removal of the predictor *hold* from the interaction with *modality*. Although these two models were closely followed by models g_28 and g_214 , in which both more and fewer parameters were estimated, different random structures were explored for g_29 and g_210 in the subsequent models, from model g_215 to model g_220 . However, the minimal adequate model was g_215 ($AIC = 2533.70$), in which the upper level, *speaker*, of the nested by-item random effects was removed. It was followed by the competitive model g_216 ($AIC = 2234.10$), in which the absence of *hold* from the interaction with *modality* was the only difference from the minimal adequate model (Table 34).

Experiment II

Model	Prominence as a function of (-) fixed effects, with by-item and by-subject random effects	LogLikelihood	K	AIC	Δ_i	w_i
g_2^1	modality x (prep + stroke + apex + hold + recoil + z-intensity + z-HIH2 + z-duration + musical-training + participant-gender) + (1 speaker/sentence/word) + (1 participant)	-1243.56	30	2547.11	13.42	0.00
g_2^2	modality x (prep + stroke + apex + hold + recoil + z-intensity + z-HIH2 + z-duration + musical-training) + participant-gender + (1 speaker/sentence/word) + (1 participant)	-1243.56	29	2545.12	11.43	0.00
g_2^3	modality x (prep + stroke + apex + hold + z-intensity + z-duration + musical-training) + recoil + z-ff + z-HIH2 + participant-gender + (1 speaker/sentence/word) + (1 participant)	-1244.69	26	2541.37	7.68	0.01
g_2^4	modality x (prep + stroke + apex + hold + z-intensity + z-duration + musical-training) + z-ff + z-HIH2 + participant-gender + (1 speaker/sentence/word) + (1 participant)	-1244.80	25	2539.60	5.91	0.01
g_2^5	modality x (stroke + apex + hold + z-intensity + z-duration + musical-training) + prep + z-ff + z-HIH2 + participant-gender + (1 speaker/sentence/word) + (1 participant)	-1246.10	24	2540.20	6.51	0.01
g_2^6	modality x (prep + apex + hold + z-intensity + z-duration + musical-training) + stroke + z-ff + participant-gender + (1 speaker/sentence/word) + (1 participant)	-1246.63	23	2539.26	5.56	0.01
g_2^7	modality x (prep + apex + hold + z-intensity + z-duration + musical-training) + z-ff + participant-gender + (1 speaker/sentence/word) + (1 participant)	-1246.80	22	2537.60	3.90	0.03
g_2^8	modality x (apex + hold + z-intensity + z-duration + musical-training) + prep + z-ff + participant-gender + (1 speaker/sentence/word) + (1 participant)	-1247.10	21	2536.19	2.50	0.06
g_2^9	modality x (apex + hold + z-intensity + z-duration) + prep + z-ff + participant-gender + (1 speaker/sentence/word) + (1 participant)	-1250.85	17	2535.70	2.00	0.07
g_2^{10}	modality x (apex + z-intensity + z-duration) + prep + hold + z-ff + participant-gender + (1 speaker/sentence/word) + (1 participant)	-1252.05	16	2536.10	2.41	0.06
g_2^{11}	modality x (apex + z-intensity + z-duration) + prep + hold + z-ff + musical-training + participant-gender + (1 speaker/sentence/word) + (1 participant)	-1250.85	18	2537.70	4.00	0.03
g_2^{12}	modality x (apex + z-duration) + prep + hold + z-ff + z-intensity + participant-gender + (1 speaker/sentence/word) + (1 participant)	-1253.69	15	2537.38	3.68	0.03
g_2^{13}	modality x (apex + z-intensity + z-duration) + hold + z-ff + participant-gender + (1 speaker/sentence/word) + (1 participant)	-1253.44	15	2536.88	3.19	0.04
g_2^{14}	modality x (apex + z-duration) + hold + z-ff + participant-gender + (1 speaker/sentence/word) + (1 participant)	-1255.28	13	2536.56	2.86	0.05
g_2^{15}	modality x (apex + hold + z-intensity + z-duration) + prep + z-ff + participant-gender + (1 sentence/word) + (1 participant)	-1250.85	16	2533.70	0.00	0.20
g_2^{16}	modality x (apex + z-intensity + z-duration) + prep + hold + z-ff + participant-gender + (1 sentence/word) + (1 participant)	-1252.05	15	2534.10	0.41	0.16

(a) Summary of AIC results for the GLMM random-intercept models.

Model	Prominence as a function of (-) fixed effects, with by-item and by-subject random effects	LogLikelihood	K	AIC	Δ_i	w_i
g_2^{17}	modality x (apex + hold + z-intensity + z-duration) + prep + hold + z-ff + participant-gender + (modality sentence/word) + (1 participant)	-1249.60	20	2539.20	5.50	0.01
g_2^{18}	modality x (apex + hold + z-intensity + z-duration) + prep + hold + z-ff + participant-gender + (modality speaker/sentence/word) + (1 participant)	-1249.60	23	2545.20	11.50	0.00
g_2^{19}	modality x (apex + z-intensity + z-duration) + prep + hold + z-ff + participant-gender + (modality sentence/word) + (1 participant)	-1250.25	19	2538.49	4.80	0.02
g_2^{20}	modality x (apex + z-intensity + z-duration) + prep + hold + z-ff + participant-gender + (modality speaker/sentence/word) + (1 participant)	-1250.31	22	2544.62	10.92	0.00

(b) Summary of AIC results for the GLMM random-slope models.

Table 34: Summary of AIC results for the GLMMs in condition Exp2. The marking of prominence by participants is modelled as a function of predictors declared as fixed effects. All models include by-subject and by-item random effects. Table (a) shows random-intercept models; Table (b) shows different structures in random effects. K indicates the estimated parameters in the model. The statistics associated to each model are Akaike Information Criterion (AIC), increase (Δ_i) of each model respect to the minimum AIC value, the log-likelihood of the model, and the Akaike weight (w_i) for each candidate model to be the minimal adequate model.

Details of minimal adequate model $g_2 15$

The minimal adequate model, $g_2 15$, did not show overdispersion ($\Phi_{\text{Pearson}} = 0.87, p > .05$; see Appendix B5d for details) and revealed a strong effect of f_0 ($\beta = 0.45, SE = 0.13, z = 3.44, p < .001$) and duration ($\beta = 0.74, SE = 0.17, z = 4.42, p < .001$) predicting prominence. Thus, as standardized variables, 1 standard deviation increase in the continuous variables of f_0 ($SD = 36.84$ Hz) and in duration ($SD = 0.064$ s) made words 1.57 and 2.11 times, respectively, more likely to be considered prominent (Figure 56 and Table 35).

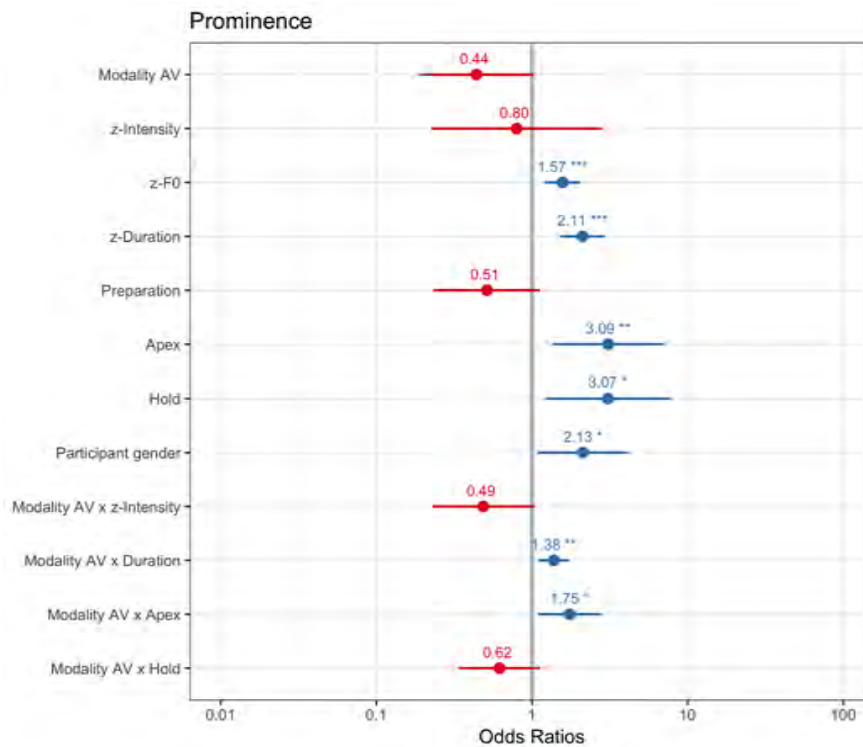


Figure 56: Forest plot showing odds ratios for main effects and interactions predicting prominence in $g_2 15$ ($AIC = 2533.70$). For $OR < 1$, effect size equals $1/OR$. Error bars are 95% CI.

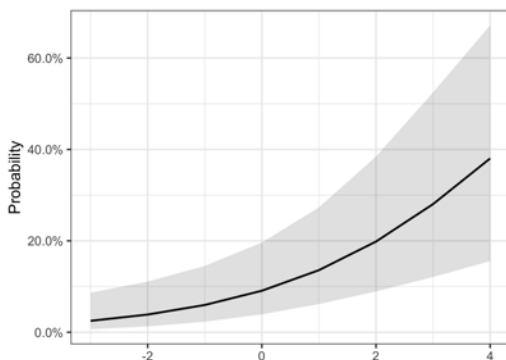
The marks of prominence given by participants significantly coincided with the gesture phases of apex ($\beta = 1.12, SE = 0.41, z = 2.70, p < .01$) and hold ($\beta =$

Experiment II

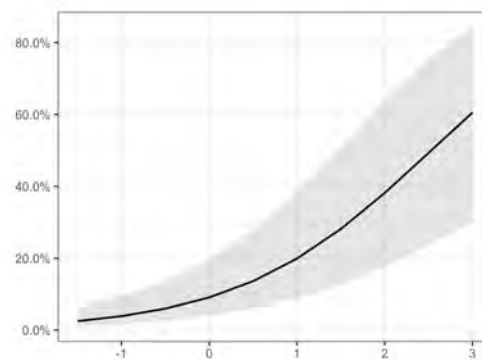
Predictor	β (SE)	p	95% CI for odds ratio		
			Lower limit	Odds ratio	Upper limit
Intercept	-2.91 (0.60)	< .001***	0.02	0.05	0.18
Modality (0=A, 1=AV)	-0.82 (0.42)	.05	0.19	0.44	1.02
z-Intensity (standardized intensity)	-0.22 (0.64)	.72	0.22	0.80	2.82
z-F0 (standardized f_0)	0.45 (0.13)	< .001**	1.22	1.57	2.04
z-Duration (standardized duration)	0.74 (0.17)	< .001***	1.52	2.11	2.95
Preparation (0=no, 1=yes)	-0.66 (0.40)	< .09	0.23	0.51	1.13
Apex (0=no, 1=yes)	1.12 (0.41)	< .01**	1.36	3.09	7.01
Hold (0=no, 1=yes)	1.12 (0.47)	< .05*	1.21	3.07	7.79
Participant gender	0.75 (0.34)	< .05*	1.09	2.13	4.16
Modality AV x Apex	0.55 (0.23)	< .05*	1.09	1.75	2.79
Modality AV x Hold	-0.48 (0.30)	.11	0.34	0.62	1.13
Modality AV x z-Intensity	-0.72 (0.38)	.06	0.23	0.49	1.04
Modality AV x z-Duration	0.32 (0.11)	< .01**	1.10	1.38	1.73

Table 35: Results of fixed effects in model g_2 15 (AIC = 2533.70) predicting the marking of prominence from the variables modality, intensity, fundamental frequency, duration, gesture-preparation, gesture-apex, gesture-hold, and participant gender.

1.12, $SE = 0.47$, $z = 2.36$, $p < .05$), so that words coincidental with apexes and holds were 3.09 and 3.07 times, respectively, more likely to be considered prominent. Additionally, the odds for words rated by women were 2.13 times more likely to be considered prominent than those words rated by men ($\beta = 0.75$, $SE = 0.34$, $z = 2.20$, $p < .05$)



(a) $z-f_0$



(b) z -Duration

Experiment II

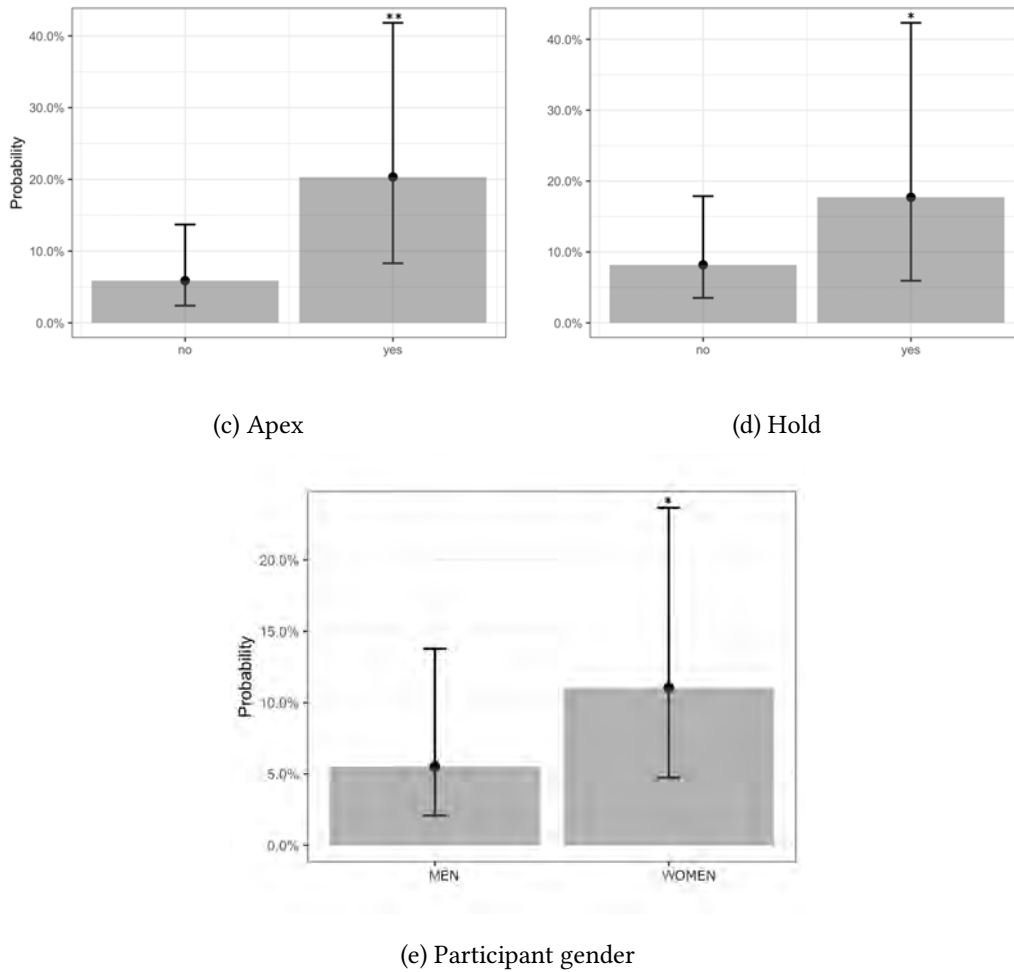


Figure 57: Main effects in model $g_2 15$ ($AIC = 2533.70$).

The audiovisual modality, where there was a slight tendency for words to be given fewer marks of prominence, fell short of significance ($\beta = -0.82$, $SE = 0.42$, $z = -1.90$, $p = .05$). However, words co-occurring together with an apex in this modality increased their odds of being perceived as prominent by 1.75 ($\beta = 0.55$, $SE = 0.23$, $z = 2.32$, $p < .05$). Similarly, the presence of the visual cues of prominence made that duration increased the odds of marking a words as prominent by 1.38 ($\beta = 0.32$, $SE = 0.11$, $z = 2.80$, $p < .01$). A contrary trend was observed for intensity, although this interaction did not reach significance ($\beta = -0.72$, $SE = 0.38$, $z = -1.86$, $p = .06$).

Experiment II

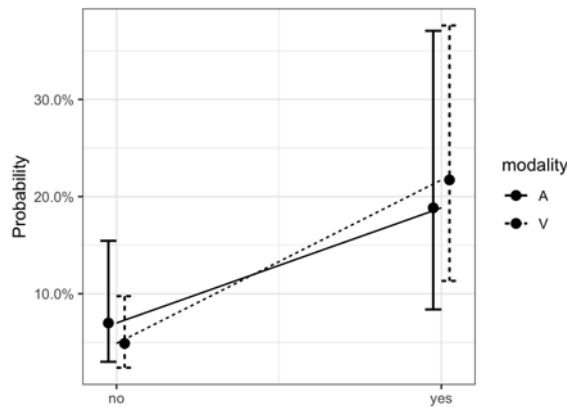
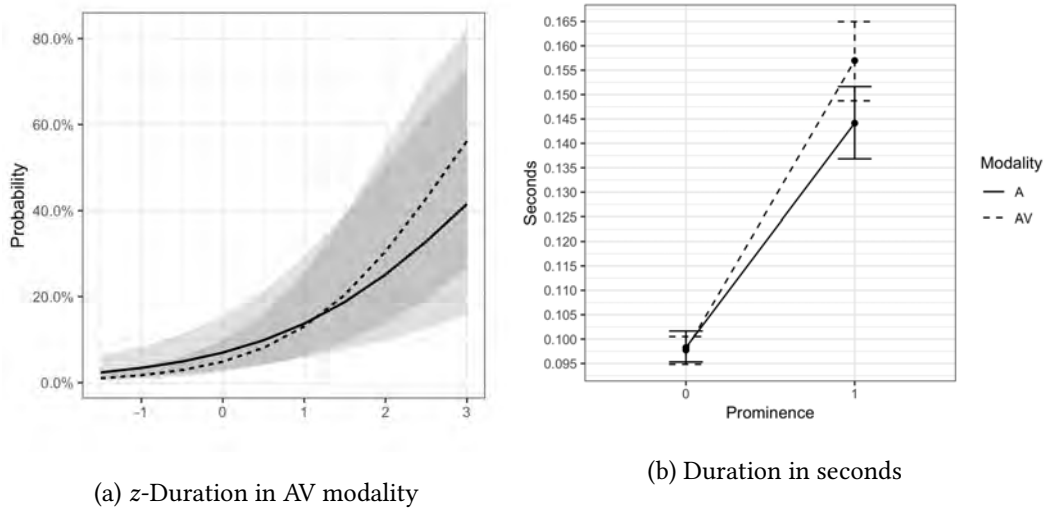


Figure 58: *Interactions between modality and duration, and modality and apex in model g_215 .*

The second-best model, g_216 ($AIC = 2534.10$), was within 2 Δ -points from the minimal adequate model, differing from it in the removal of the predictor *hold* from the interaction with *modality*, which made that the predictor *modality* reached significance (Figure 59 and Table 36). The goodness of fit of both models was not affected by the removal of the upper level *speaker* from the nested by-item random effects.

Experiment II

Predictor	β (SE)	p	95% CI for odds ratio		
			Lower limit	Odds ratio	Upper limit
Intercept	-2.87 (0.60)	< .001***	0.02	0.06	0.018
Modality (0=A, 1=AV)	-0.96 (0.42)	< .05*	0.38	0.17	0.88
z-Intensity (standardized intensity)	-0.24 (0.64)	.70	0.22	0.78	2.77
z-F0 (standardized f_0)	0.45 (0.13)	< .001**	1.22	1.57	2.04
z-Duration (standardized duration)	0.76 (0.17)	< .001***	1.54	2.15	2.99
Preparation (0=no, 1=yes)	-0.67 (0.40)	< .09	0.23	0.51	1.13
Apex (0=no, 1=yes)	1.07 (0.41)	< .05*	1.29	2.92	6.60
Hold (0=no, 1=yes)	0.93 (0.46)	< .05*	1.04	2.55	6.27
Participant gender	0.75 (0.34)	< .05*	1.09	2.13	4.17
Modality AV x Apex	0.71 (0.22)	< .001**	1.32	2.03	3.13
Modality AV x z-Intensity	-0.71 (0.38)	.06	0.23	0.49	1.05
Modality AV x z-Duration	0.29 (0.11)	< .01**	1.07	1.34	1.67

Table 36: Results of fixed effects in model g_216 ($AIC = 2534.10$) predicting the marking of prominence from the variables modality, intensity, fundamental frequency, duration, gesture-preparation, gesture-apex, gesture-hold, and participant gender.

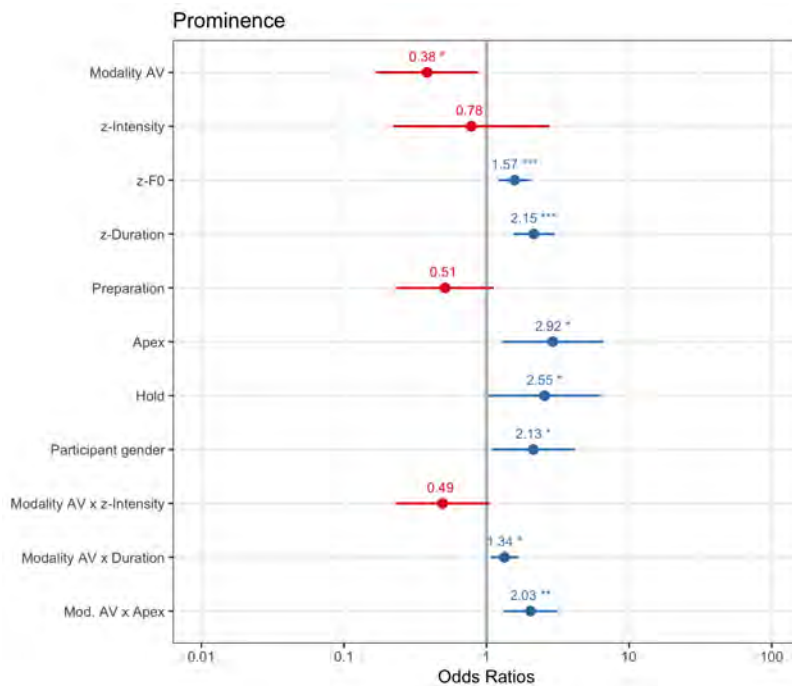


Figure 59: Odds ratios for main effects and interactions predicting prominence in Exp2 for competitive model g_216 . For $OR < 1$, effect size equals $1/OR$. Error bars are 95% CI.

Summary

The analysis of this second experimental condition Exp2, where intensity had been neutralised as an acoustic cue of prominence, started with the declaration of a global model. Initially, several models proved competitive and included both more and fewer predictors. From these, models $g_2 9$ ($AIC = 2535.70$) and $g_2 10$ ($AIC = 2536.10$) had the lowest AIC value. The removal of the upper level *speaker* from their nested by-item random effects resulted in the minimal adequate model $g_2 15$ and its competitor $g_2 16$. These two were reported, but only the minimal adequate model was discussed.

This minimal adequate model revealed that participants' marks of prominence were influenced by the acoustic correlates of f_0 and duration. Besides, their marks significantly coincided with words accompanied by the apex and hold phases of gestures. Unlike the results obtained so far, apexes were actually observed to make participants increase their marks of prominence in the audiovisual modality, while neither strokes nor holds played any role in the audiovisual modality in this second condition Exp2.

In this sense, the audiovisual modality showed a tendency for participants to reduce their marks of prominence when compared to the audio-only modality and although this effect fell short of significance ($p = .05$), it proved significant in the second-best model $g_2 16$ (Table 36) when the predictor *hold* was removed from the interaction with *modality* (Figure 59). Interestingly, and also differently from the analyses of the previous conditions, participants seemed to get closer to the 'gold standard' in both modalities, and they even reached a higher agreement on their marks of prominence in the audiovisual than in the audio-only modality.

Probably the most surprising result found in this condition Exp2 is the significant difference found between men and women, where the latter were much more likely to consider words as prominent. However, this did not seem to be

Experiment II

determined by the visual cues of prominence, since no interaction was found between the predictors *participant gender* and *modality*. Despite this, when the distribution of participants according to their gender was analysed in more detail, it was observed that a clear unbalance existed between both modalities (Figure 60). Therefore, it is possible that this fact could underlie the significant difference observed in the performance of male and female participants.

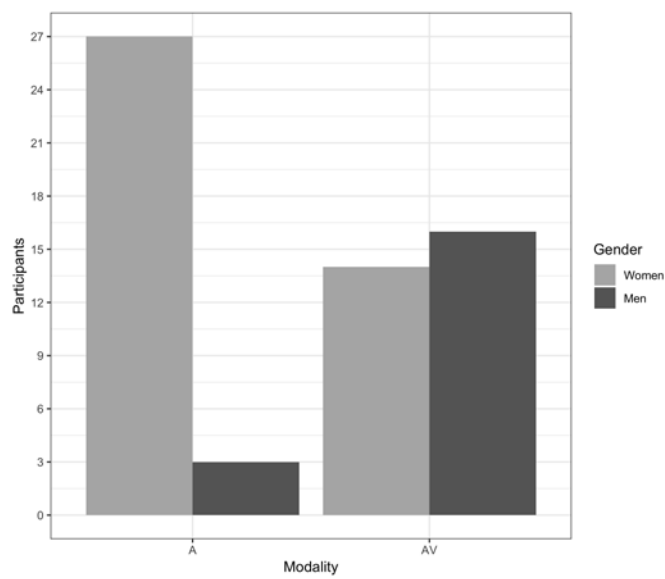


Figure 60: Number of participants in both modalities according to their gender in Exp2.

5.3.2.7 Analysis E: Third condition ‘Exp3’ (duration)

The third experimental condition, Exp3, for which a separate process of data collection was carried out with a group of 60 participants, involved the neutralisation of both f_0 and intensity. The modelling of participants’ responses started by declaring an initial model as done in the previous experimental conditions (§ 5.3.2.4, § 5.3.2.5, § 5.3.2.6):

```
prominence ~ modality * (condition + preparation + stroke + apex + hold + recoil +  
+ z_intensity + z_H1H2 + z_ff + z_duration + musical_training + partic_gender) +  
+ (1|participant) + (1|speaker/sentence/word)
```

Model selection proceeded progressively by removing non-significant predictors from model g_31 to model g_317 . From these, several models appeared within less than 2 Δ -points from the last model, g_317 ($AIC = 2408.90$), all having between one and three more predictors, i.e. g_316 , g_315 , g_314 , and g_313 .

In the subsequent models, from g_318 to g_323 , random effects structures for the model with the lowest AIC value, g_317 , and for its immediate competitor, were explored. Models g_318 and g_319 omitted the upper level *speaker* of by-item random effects. Then, random slopes were declared for the effect of the variable *modality* on by-item random effects in the remaining four models, from g_320 to g_323 . In these four models the upper level *speaker* that had just been removed from by-item random effects was alternatively reintroduced to test whether the AIC value decreased. However, the resulting AIC value did not improve that of the random-intercept models g_318 ($AIC = 2407.50$) and g_319 ($AIC = 2407.77$). From these two competitors, model g_318 was preferred for being the more parsimonious of the two as well as for having a lower AIC value (Table 37).

Experiment II

Model	Prominence as a function of (-) fixed effects, with by-subject and by-item random effects	Log-likelihood	K	AIC	Δ_i	w_i
$g_{3,1}$	modality x (prep + stroke + apex + hold + recoil + z-ff + z-intensity + z-H1H2 + z-duration + musical-training + participant-gender) + (1 participant) + (1 speaker/sentence/word)	-1185.27	30	2430.55	23.05	0.00
$g_{3,2}$	modality x (prep + stroke + apex + hold + recoil + z-ff + z-intensity + z-H1H2 + z-duration + musical-training) + participant-gender + (1 participant) + (1 speaker/sentence/word)	-1185.87	29	2429.74	22.24	0.00
$g_{3,3}$	modality x (prep + stroke + apex + hold + recoil + z-intensity + z-ff + z-duration + musical-training) + z-H1H2 + participant-gender + (1 participant) + (1 speaker/sentence/word)	-1185.88	28	2327.77	20.27	0.00
$g_{3,4}$	modality x (prep + stroke + apex + hold + recoil + z-intensity + z-ff + z-duration) + z-H1H2 + musical-training + participant-gender + (1 participant) + (1 speaker/sentence/word)	-1187.13	26	2426.27	18.77	0.00
$g_{3,5}$	modality x (stroke + apex + hold + recoil + z-ff + z-intensity + z-duration) + prep + z-H1H2 + musical-training + participant-gender + (1 participant) + (1 speaker/sentence/word)	-1187.17	25	2424.33	16.83	0.01
$g_{3,6}$	modality x (stroke + hold + recoil + z-ff + z-intensity + z-duration) + prep + apex + z-H1H2 + musical-training + participant-gender + (1 participant) + (1 speaker/sentence/word)	-1187.24	24	2422.47	14.97	0.01
$g_{3,7}$	modality x (stroke + hold + recoil + z-ff + z-duration) + prep + apex + z-intensity + z-H1H2 + musical-training + participant-gender + (1 participant) + (1 speaker/sentence/word)	-1187.27	23	2420.55	13.05	0.01
$g_{3,8}$	modality x (stroke + hold + recoil + z-ff) + prep + apex + z-intensity + z-H1H2 + z-duration + musical-training + participant-gender + (1 participant) + (1 speaker/sentence/word)	-1187.28	22	2418.56	11.06	0.00
$g_{3,9}$	modality x (stroke + hold + recoil + z-ff) + prep + apex + z-H1H2 + z-duration + musical-training + participant-gender + (1 participant) + (1 speaker/sentence/word)	-1187.29	21	2416.58	9.08	0.00
$g_{3,10}$	modality x (stroke + hold + recoil + z-ff) + prep + apex + z-H1H2 + z-duration + participant-gender + (1 participant) + (1 speaker/sentence/word)	-1188.00	19	2414.01	6.51	0.01
$g_{3,11}$	modality x (stroke + hold + recoil) + prep + apex + z-H1H2 + z-duration + participant-gender + (1 participant) + (1 speaker/sentence/word)	-1188.26	18	2412.53	5.03	0.02
$g_{3,12}$	modality x (stroke + hold + recoil) + apex + z-ff + z-H1H2 + z-duration + participant-gender + (1 participant) + (1 speaker/sentence/word)	-1188.66	17	2411.32	3.82	0.04
$g_{3,13}$	modality x (stroke + hold + recoil) + apex + z-ff + z-duration + participant-gender + (1 participant) + (1 speaker/sentence/word)	-1188.93	16	2409.85	2.35	0.08
$g_{3,14}$	modality x (stroke + hold + recoil) + apex + z-duration + participant-gender + (1 participant) + (1 speaker/sentence/word)	-1189.58	15	2409.16	1.66	0.11
$g_{3,15}$	modality x (stroke + recoil) + apex + hold + z-duration + participant-gender + (1 participant) + (1 speaker/sentence/word)	-1191.24	14	2410.48	2.98	0.06
$g_{3,16}$	modality x (hold + recoil) + stroke + apex + z-duration + participant-gender + (1 participant) + (1 speaker/sentence/word)	-1191.19	14	2410.38	2.88	0.06
$g_{3,17}$	modality x (recoil) + stroke + apex + hold + z-duration + participant-gender + (1 participant) + (1 speaker/sentence/word)	-1191.45	13	2408.90	1.40	0.13
$g_{3,18}$	modality x (recoil) + stroke + apex + hold + z-duration + participant-gender + (1 participant) + (1 sentence/word)	-1191.75	12	2407.50	0.00	0.25
$g_{3,19}$	modality x (stroke + hold + recoil) + apex + z-duration + participant-gender + (1 participant) + (1 sentence/word)	-1189.88	14	2407.77	0.27	0.22

(a) Summary of AIC results for the GLMM random-intercept models.

Model	Prominence as a function of (-) fixed effects, with by-subject and by-item random effects	LogLikelihood	K	AIC	Δ_i	w_i
g _{s,20}	modality x (recoil) + stroke + apex + hold + z-duration + participant-gender + + (1 participant) + (modality sentence/word)	-1190.91	16	2413.82	6.32	0.01
g _{s,21}	modality x (recoil) + stroke + apex + hold + z-duration + participant-gender + + (1 participant) + (modality speaker/sentence/word)	-1190.44	19	2418.87	11.37	0.00
g _{s,22}	modality x (stroke + hold + recoil) + apex + z-duration + participant-gender + + (1 participant) + (modality speaker/sentence/word)	-1198.66	18	2415.31	7.81	0.01
g _{s,23}	modality x (stroke + hold + recoil) + apex + z-duration + participant-gender + + (1 participant) + (modality sentence/word)	-1185.27	21	2420.12	12.62	0.00

(b) Summary of AIC results for the GLMM random-slope models.

Table 37: Summary of AIC results for the GLMMs in condition Exp3. The marking of prominence by participants is modelled as a function of predictors declared as fixed effects. All models include by-subject and by-item random effects. Table (a) shows random-intercept models; Table (b) shows different structures in random effects. K indicates the estimated parameters in the model. The statistics associated to each model are Akaike Information Criterion (AIC), increase (Δ_i) of each model respect to the minimum AIC value, the log-likelihood of the model, and the Akaike weight (w_i) for each candidate model to be the minimal adequate model.

Details of minimal adequate model $g_3 18$

The minimal adequate model $g_3 18$ ($AIC = 2407.50$) did not present overdispersion ($\Phi_{\text{Pearson}} = 0.77$, $p > .05$; see Appendix B5e for details). This model revealed that participants strongly relied on the only acoustic cue available to them in the task of binary prominence marking ($\beta = 0.91$, $SE = 0.19$, $z = 4.74$, $p < .001$). Thus, 1 standard deviation in syllable duration ($SD = 0.064$ s) increased 2.49 times the odds for marking words as prominent (Table 38).

Predictor	β (SE)	p	95% CI for odds ratio		
			Lower limit	Odds ratio	Upper limit
Intercept	-4.39 (0.56)	< .001***	0.00	0.01	0.04
Modality (0=A, 1=AV)	-0.11 (0.24)	.63	0.55	0.89	1.44
z-Duration (standardized duration)	0.91 (0.20)	< .001***	1.71	2.49	3.62
Stroke (0=no, 1=yes)	1.83 (0.50)	< .001***	2.36	6.25	16.55
Apex (0=no, 1=yes)	1.30 (0.54)	< .05*	1.27	3.67	10.62
Hold (0=no, 1=yes)	2.67 (0.71)	< .001***	3.61	14.56	58.70
Recoil (0=no, 1=yes)	1.56 (0.75)	< .05*	1.10	4.76	20.69
Participant gender	0.65 (0.24)	< .01**	1.19	2.49	3.62
Modality AV x Recoil	-0.70 (0.34)	< .05*	0.25	0.49	0.97

Table 38: Results of fixed effects in model $g_3 18$ ($AIC = 2407.50$) predicting the marking of prominence from the variables modality, duration, gesture-stroke, gesture-apex, gesture-hold, gesture-recoil and participant gender.

Participants' marks significantly coincided with all gesture phases except for the preparation phase of gestures. Those words coinciding with strokes were 6.25 times more likely to receive a mark of prominence ($\beta = 1.83$, $SE = 0.50$, $z = 3.68$, $p < .05$); odds raised up to 3.67 times for words coinciding with apexes ($\beta = 1.30$, $SE = 0.54$, $z = 2.40$, $p < .001$); for holds, odds increased the probability of words to be considered prominent by 14.56 times ($\beta = 2.67$, $SE = 0.71$, $z = 3.76$, $p < .05$); and finally, words coinciding with the recoil phase of gestures were 4.76

Experiment II

times more likely to be given a mark of prominence ($\beta = 1.56$, $SE = 0.75$, $z = 2.08$, $p < .001$).

In the audiovisual modality, however, participants did not perform differently from the audio-only modality, and only the recoil phase of gestures seemed to have an influence on the way participants marked prominence. In this case, words coincided with recoils in the audiovisual modality were 2.04 times less likely to receive a mark of prominence ($\beta = -0.70$, $SE = 0.34$, $z = -2.06$, $p < .05$).

Finally, a moderate effect was observed for the way women marked prominent words when compared to men. As it is apparent in this minimal adequate model, as well as in its competitor, $g_3 19$, women were 1.93 times more likely than men to give a word a mark of prominence.

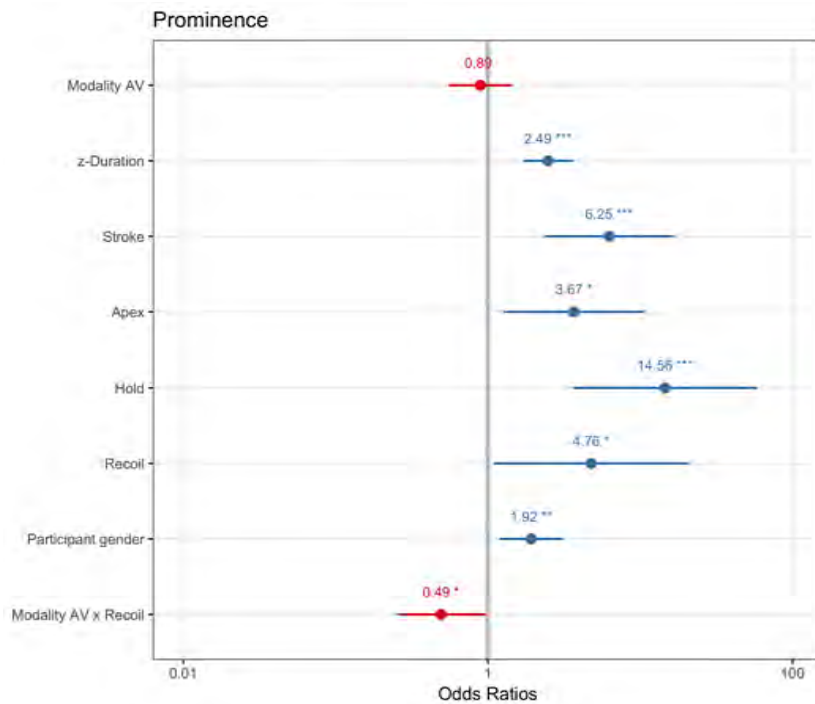
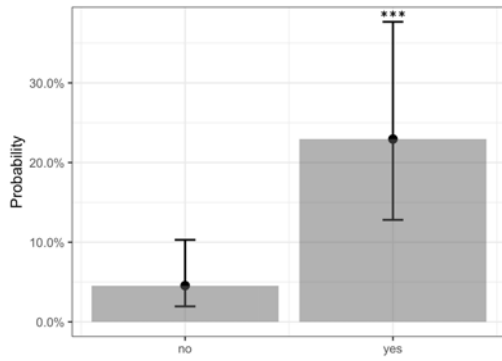
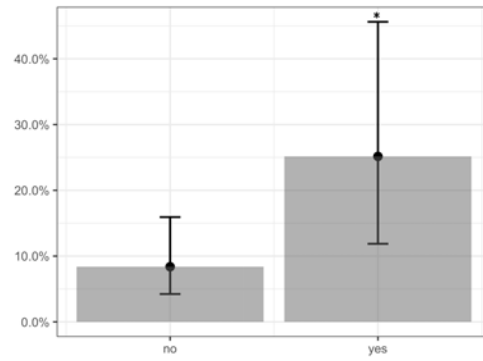


Figure 61: Forest plot showing odds ratios for main effects and interactions predicting prominence in $g_3 18$ ($AIC = 2407.50$). For $OR < 1$, effect size equals $1/OR$. Error bars are 95% CI.

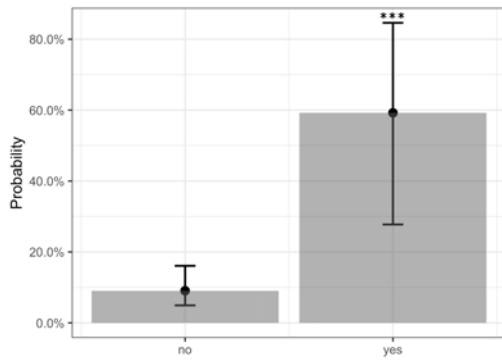
Experiment II



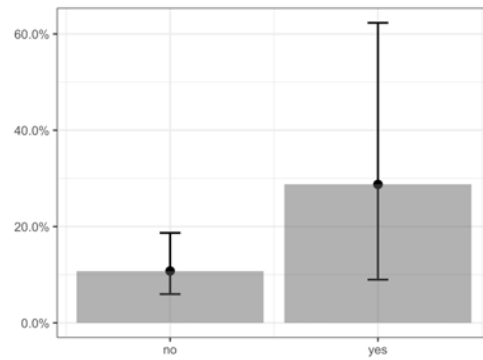
(a) Stroke



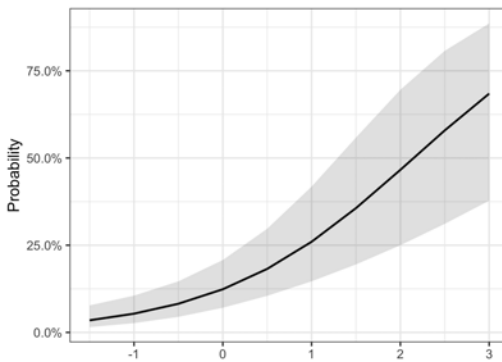
(b) Apex



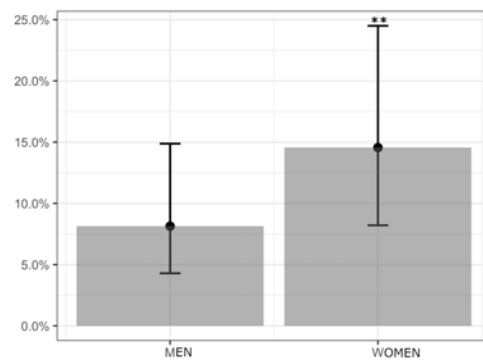
(c) Hold



(d) Recoil



(e) z-Duration



(f) Participant gender

Figure 62: Main effects in model $g_3 18$ (AIC = 2407.50).

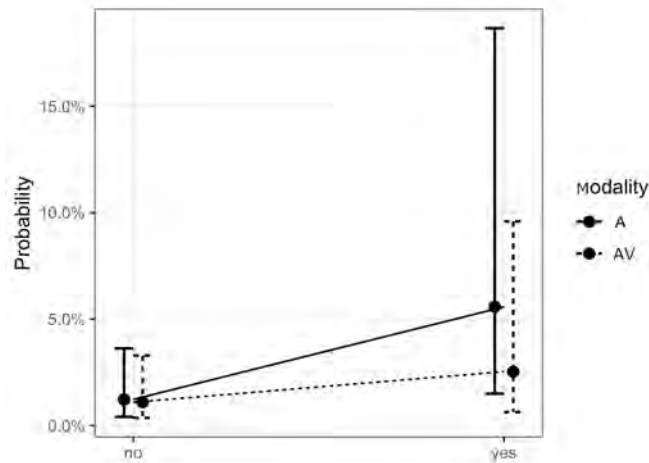


Figure 63: Interaction in model $g_3 18$ ($AIC = 2407.50$) between recoil and modality.

After the minimal adequate model, $g_3 18$, the next model within 2 Δ -points was model $g_3 19$ ($AIC = 2407.77$). This competitor included two more estimated parameters, which corresponded to the predictors *stroke* and *hold* in an interaction with *modality*. None of these two gesture phases proved significant through the effect of modality, and although both fell short of significance, they showed a tendency to make participants mark fewer prominent words coinciding with them (strokes, $\beta = -0.63$, $SE = 0.34$, $z = -1.83$, $p = .06$; and holds, $\beta = -1.16$, $SE = 0.41$, $z = -2.78$, $p = .06$). The sign and significance of the remaining odds ratios remained the same in this competitor model, even if the effect size for some of them varied, which especially affected the predictors *apex*, *hold*, and *recoil* (Table 39).

Experiment II

Predictor	β (SE)	<i>p</i>	95% CI for odds ratio		
			Lower limit	Odds ratio	Upper limit
Intercept	-4.71 (0.60)	< .001***	0.00	0.01	0.03
Modality (0=A, 1=AV)	-0.49 (0.40)	.21	0.75	1.64	3.55
z-Duration (standardized duration)	0.91 (0.19)	< .001***	1.71	2.49	3.63
Stroke (0=no, 1=yes)	2.16 (0.53)	< .001***	3.05	8.73	24.98
Apex (0=no, 1=yes)	1.30 (0.54)	< .05*	1.26	3.65	10.57
Hold (0=no, 1=yes)	3.06 (0.74)	< .001***	4.97	21.42	92.29
Recoil (0=no, 1=yes)	1.81 (0.76)	< .05*	1.37	6.12	27.44
Participant gender	0.65 (0.24)	< .01**	1.19	1.93	3.11
Modality AV x Stroke	-0.63 (0.34)	.06	0.27	0.53	1.05
Modality AV x Hold	-0.75 (0.40)	.06	0.21	0.47	1.04
Modality AV x Recoil	-0.70 (0.34)	< .05*	0.14	0.31	0.71

Table 39: Results of fixed effects in model $g_3 19$ (AIC = 2407.77) predicting the marking of prominence from the variables modality, duration, gesture-stroke, gesture-apex, gesture-hold, gesture-recoil and participant gender.

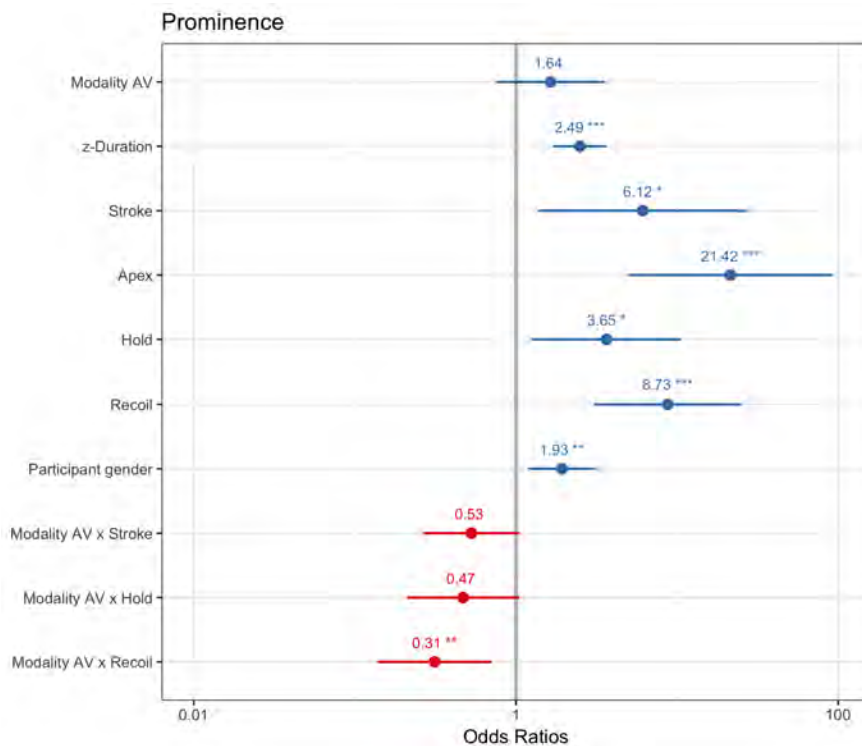


Figure 64: Odds ratios for main effects and interactions predicting prominence in Exp3 for competitive model $g_3 19$ (AIC = 2407.77). For OR < 1, effect size equals 1/OR. Error bars are 95% CI.

Summary

In this third experimental condition Exp3, both f_0 and intensity had been neutralised as acoustic cues of prominence. The initial model fitted to analyse the prediction of prominence with the variables of interest also included the predictors *musical training* and *participant gender*. Among the different models fitted in the model set, the AIC value of a few of them was within a narrow range from each other. From the subset of these competitive models, the two with the lowest AIC value, $g_3 14$ (AIC = 2409.16) and $g_3 17$ (AIC = 2408.90), and fewer predictors—15 and 13, respectively—, were subsequently declared with different by-item random effects structures. The minimal adequate model was the random-intercept model $g_3 18$ (AIC = 2407.50), whose by-item random effects did not include the upper level *speaker* in its nested structure.

The minimal adequate model revealed that syllable duration, as the only acoustic cue of prominence available to participants, had a strong effect on them for marking words as prominent. Furthermore, differently from Exp1, where also f_0 had been neutralised, in this Exp3 participants declaring having received more than 5 years of formal musical training did not seem to perform differently from other participants.

As for the gesture phases coincidental with words, almost all of them increased the probabilities of participants' giving those words a mark of prominence: strokes, apexes, holds, and recoils—only preparation did not show any effect. On the one hand, a previous comparison of this Exp3 against the control condition in model $G17$ (Figure 48) showed a tendency for participants to consider more words as prominent in all experimental conditions, although this time the difference in this Exp3 fell short of significance ($p = .07$) (Table 27). On the other hand, agreement in both the audio-only ($\kappa = .350$) and the audiovisual ($\kappa = .356$) modalities was very similar, with a very slight tendency in both mo-

dalities to improve when compared to the ‘gold standard’ ($\kappa = .366$ for the audio-only and $\kappa = .368$ for the audiovisual modality), revealing a closer agreement to the marks of two trained listeners than the agreement observed for participants in other conditions. In this Exp3 participants achieved a higher agreement than, for example, in Exp1, despite counting on fewer cues of prominence.

These results suggest that prominence marks in this Exp3 were differently distributed when compared to the previous conditions, possibly with a stronger reliance on the visual cues of prominence, although this did not result in a better agreement among participants in the audiovisual modality. Interestingly, among the gesture phases, only recoil showed differences between modalities. More precisely, recoils were observed to reduce the probability of words to be considered prominent when participants could count on the visual cues of prominence. Next to recoil, a similar trend was found in model g_319 for strokes and hold (Figure 64), although neither reached significance ($p = .06$ for both *stroke* and *hold*, Table 39).

This might be interpreted as evidence of the lesser influence of recoil on perceiving prominent words when the visual cues were available to participants. Following this line of reasoning, the gesture phases coincidental with words that might have randomly received marks of prominence in the audio-only modality—e.g. strokes, holds, recoils—, seem to have been less likely to drive participants’ prominence perception in the audiovisual modality. Therefore, a trend is apparent for words which were perceived as prominent in the audio-only modality to be considered non-prominent when participants could observe the visual cues they co-occurred with.

Finally, as observed in the second experimental condition Exp2, where intensity had been neutralised, women seemed much more likely than men to mark words as prominent, and this effect was not observed to interact with the vari-

Experiment II

able *modality*. In order to better understand this difference, the distribution of participants in each of these two experimental conditions was analysed (Figure 65), as in Exp2. However, while in Exp2, this difference might have been due to an unbalanced distribution of men and women between both modalities, such a significant difference in this Exp3 does not seem likely to be due to the same fact, especially since the number of male and female participants is very similar in both the audio-only modality and the audiovisual modality.

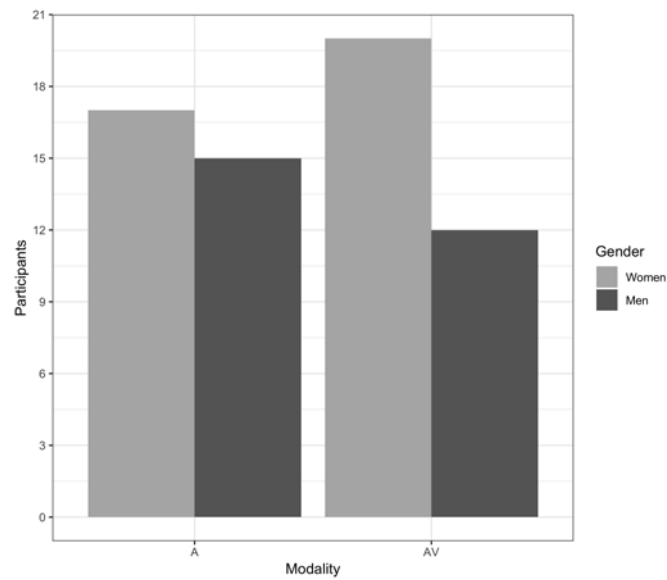
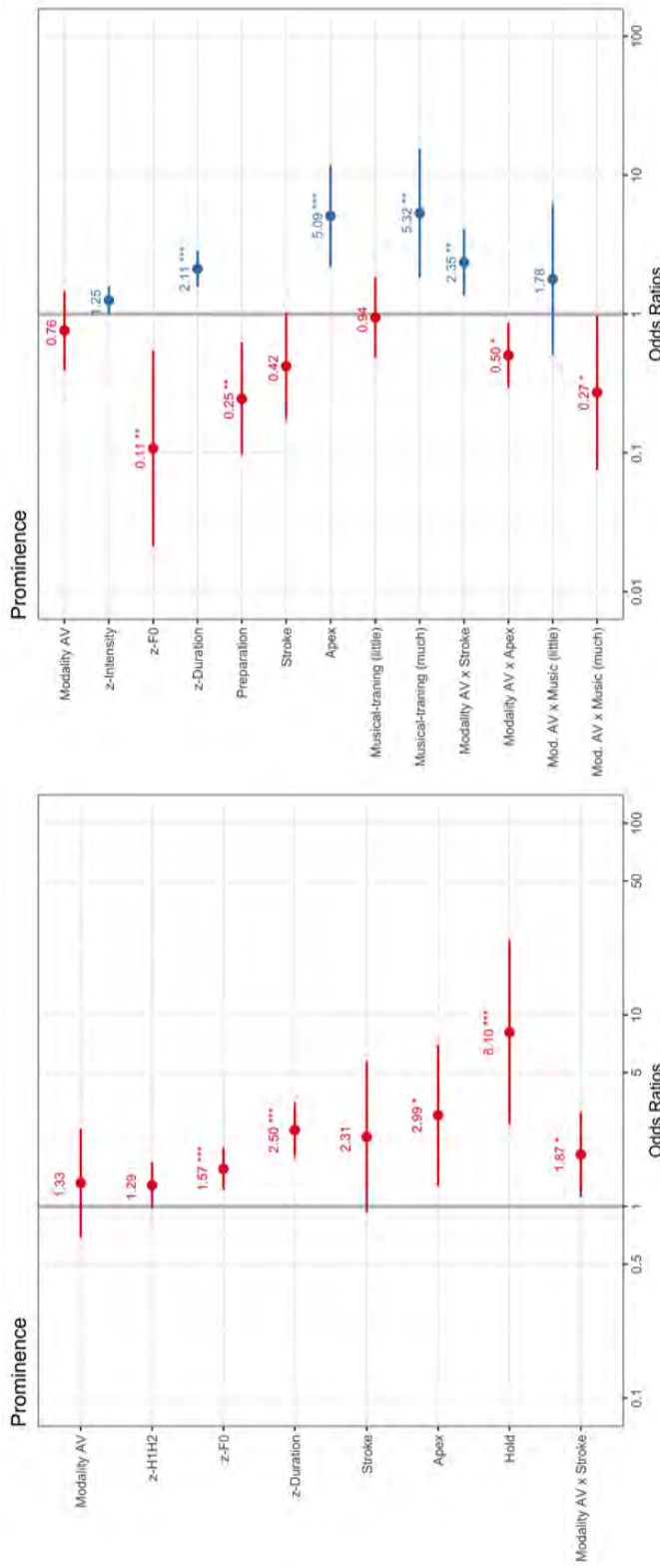


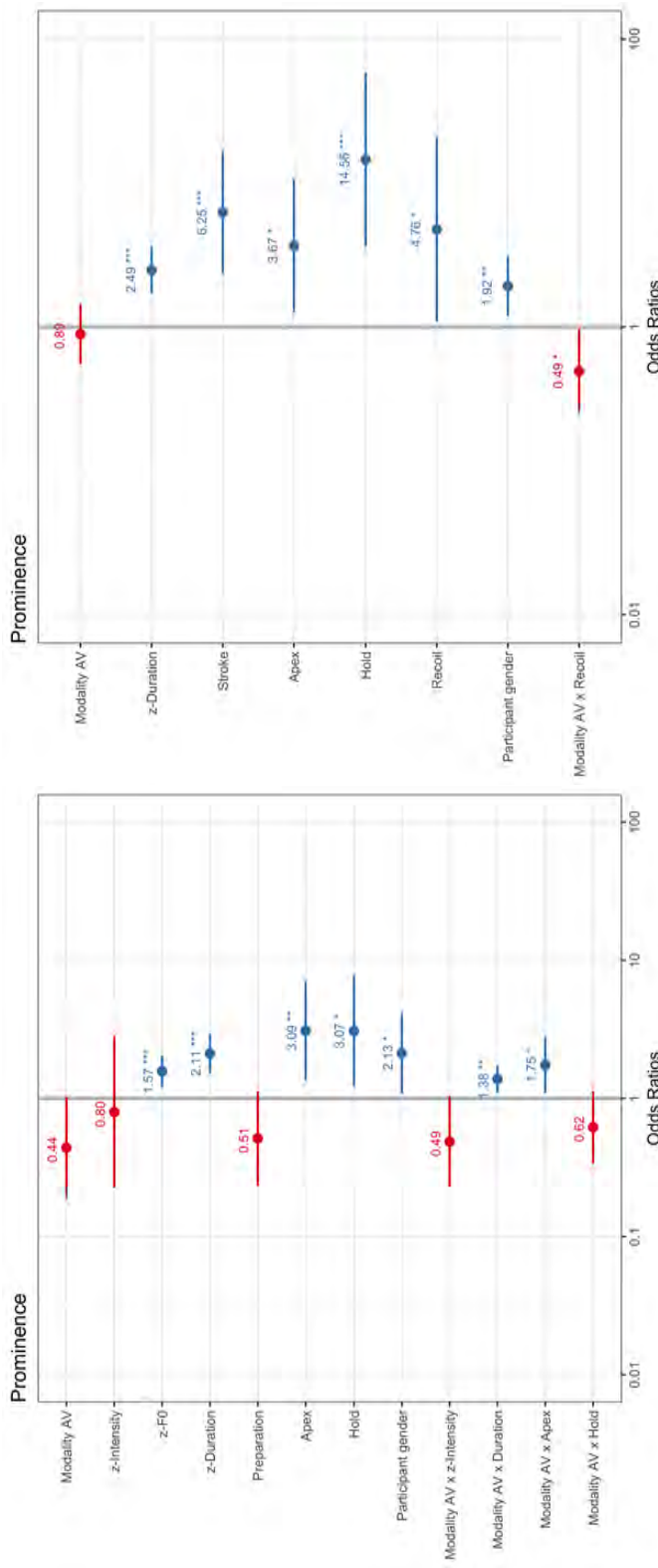
Figure 65: Number of participants in both modalities according to gender in Exp3.



(a) $g_{0,20}$ in Exp0

(b) $g_{1,13}$ in Exp1

Figure 66: Comparison of odds ratios for variables predicting prominence in the top-ranked models of (a) Exp0, (b) Exp1, (c) Exp2, and (d) Exp3.



(c) g₂15 in Exp2

(d) g₃18 in Exp3

Figure 66: Comparison of odds ratios for variables predicting prominence in the top-ranked models of (a) Exp0, (b) Exp1, (c) Exp2, and (d) Exp3.

5.4 Discussion

Perception of acoustic phrasal prominence is mediated by the visual cues of prominence, as observed in previous research (e.g. [Krahmer & Swerts, 2007](#); [Prieto et al., 2011](#)), and the interplay between both modalities has become to be known as audiovisual prosody ([Krahmer & Swerts, 2009](#)). Building on the methodology used in Experiment I, the aim of this Experiment II was to gain insight precisely into the multimodal perception of acoustic prominence and analyse how exactly the different acoustic correlates of prominence are used by listeners, with and without relying on visual cues, in Castilian Spanish. For this, four independent perceptual experiments were conducted via the Internet, each involving a different manipulation of the acoustic correlates of prominence: manipulation of f_0 in condition ‘Exp1’; of intensity in condition ‘Exp2’; and of both f_0 and intensity in ‘Exp3’—the control condition ‘Exp0’ did not involve any manipulation of the speech signal. Each experiment was also presented to participants in either the audio-only or the audiovisual modality. A total of 240 participants rated 12960 words, corresponding to four target sentences used as stimuli, so that 30 participants rated just one experimental condition in only one of the two modalities. The Internet has been previously used in similar perceptual studies successfully (e.g. [Kok et al., 2016](#); [Masson-Carro et al., 2017](#)). For example, in a study on the functional role of gestures, Kok and his colleagues ([2016](#)) administered a series of online videos to participants to rate the gestures they observed using a 7-point Likert scale. For this, participants were financially compensated through the online platform *Crowdfunder*².

Experiment II followed a similar way of administering stimuli to participants via the Internet, although it involved no remuneration, and it employed the same

² Now *Figure Eight*, <https://www.figure-eight.com/>.

methodology used in Experiment I (§ 3.3.2 and § 3.3.3). A small corpus of 50 sentences had been previously built for Experiment I using spontaneous speech samples extracted from a Spanish television talent show (*Operación Triunfo*, 1st edition). From this corpus, a set of 13 sentences were selected to be used as stimuli, so that the necessary time to complete this online Experiment II did not take much longer than 10 minutes.

Hypothesis 1: methodology

Differently from the methodology applied in this research, previous methodologies used animated agents (e.g. Al Moubayed & Beskow, 2009; Krahmer et al., 2002a) or multimodal stimuli elicited in experimental settings (e.g. Foxton et al., 2010; Krahmer & Swerts, 2007) to study multimodal prominence perception. Apart from these, very few studies employed spontaneous speech, from which only Swerts and Krahmer (2010, experiment 1: ‘auditory mark-up’) conducted a perceptual experiment, although only in the auditory modality. The limitations inherent to these methods, already mentioned (§ 3.3.1), have not previously allowed to study in detail how the different acoustic correlates of prominence relate to one another and what is the exact role played by gestures in this process.

Experiment I, using spontaneous speech to overcome some of the mentioned limitations, showed that spontaneous speech material could be used successfully in the study of the multimodal perception of prominence. Besides, it permitted to explore some of the issues that still remained unanswered and added ecological validity to the study of multimodal prominence perception. The results obtained in Experiment I were a first approximation to the two research questions addressed in this study. These results were partly replicated in this Experiment II, and they were also extended using a much more larger sample size of participants

and a smaller sample size of target sentences.

After describing the collected data, agreement among participants was calculated separately per condition and modality and expressed as mean Cohen's kappa. Participants' agreement values across conditions was generally low and ranged between $\kappa = .28$ and $\kappa = .42$ in the audio-only modality, and between $\kappa = .27$ and $\kappa = .38$ in the audiovisual modality. However, despite the low values reached in some experimental conditions, agreement reached the highest value ($\kappa = .42$) in the audio-only modality in the control condition Exp0—in which participants relied on all acoustic cues of prominence. This shows a slightly higher agreement than that obtained in studies under similar conditions, i.e. prominence rating in the audio-only modality of non-manipulated stimuli from a corpus of spontaneous speech, as in Mo et al. (2008) ($\kappa = .38$).

Certainly, Mo and her colleagues used Fleiss's kappa, a statistic that goes beyond pairwise comparisons and provides a single value for agreement among more than two raters. In this research, mean Cohen's kappa was preferred, since the responses of two trained listeners relying on all acoustic cues of prominence were used as a 'gold standard' to compare pairwise to the responses of each participant. This allowed to assess how agreement among participants deviated or got closer to this 'gold-standard' depending on whether the values of mean Cohen's kappa decreased or increased in the comparison when responses by the two trained listeners were computed (details are in 22). Despite the low agreement obtained in some conditions, such as in Exp1—neutralised f_0 —, participants agreement was generally closer to that of the two trained listeners in the audio-only modality than in the audiovisual modality. From a mere perceptual perspective, it might be possible that agreement was lower due to the higher cognitive load involved in processing the speech in the audiovisual modality.

Inter-rater agreement in this Experiment II (between $\kappa = .28$ and $\kappa = .42$)

was similar to that obtained in Experiment I ($\kappa = .39$). As just mentioned, despite being low, agreement is not lower than those obtained in studies under better and more homogenized perceptual conditions. This suggests that such low agreement cannot be exclusively ascribed to the difficulty of the task or the neutralization of the acoustic cues, but to the general difficulty that experiments on prominence perception seem to involve.

However, it is interesting to note that in this Experiment II the observed differences among modalities tended to decrease with the neutralization of acoustic cues, with a very similar agreement between modalities in Exp3 (only duration as cue). Actually, there was a consistently higher agreement in the audio-only modality than in the audiovisual modality, except in Exp2 (where intensity had been neutralized), where this trend was reversed (Figure 22). The lowest agreement was obtained in the audiovisual modality of Exp1 (where intensity had been neutralized). This seems to indicate that the visual cues of prominence did not suffice for participants to clearly agree on the prominent words of the sentences when the perceptual weight of f_0 was absent. Interestingly, such a low agreement was not observed in Exp3 (only duration), where participants' responses were more similar—and therefore with a higher inter-rater agreement—not only when compared to one another, but also when compared between modalities.

Also, in the audio-only modality the participants' responses tended to get closer to the 'gold standard' of the responses given by two trained listeners, while they got farther in two of the four experimental conditions (Exp0 and Exp1) in the audiovisual modality. This could be interpreted, as previously mentioned, as a difficulty to process the higher cognitive load involved in processing the speech in the audiovisual modality. In this respect, it is worth noting that the highest agreement was certainly in the audio-only modality of the control condition, Exp0, which makes it slightly higher than the value offered by Mo et

al. under similar conditions (2008).

As for model building and selection based on *AIC*, it was successfully applied, as in Experiment I. Although the introduction of information criteria approaches in statistical analysis that make use of mixed models is still relatively uncommon in the field of linguistics (e.g. Adamou et al., 2018), it has been gaining popularity over the last decade in other fields such as biology, ecology, or evolution, where it is now often applied (see e.g. Arnold, T. W., 2010; Grueber et al., 2011, for details).

In this Experiment II, all three experimental conditions, carried out as independent experiments on the Internet in each modality, were firstly compared against the control condition as done in Experiment I (§ 4.3.2.4 and § 4.3.2.3). Next to the control group, the ‘gold standard’ was also used to assess the performance of participants in the control condition, who could rely on all acoustic cues of prominence. Secondly, the between-subjects design of the experiment allowed to independently analyse the role of the different variables that made participants consider words prominent in each experimental condition.

For this, different generalised linear mixed models were fitted to assess the variables of interest, which included the gestures phases of preparation, stroke, apex, hold, and recoil, as well as the acoustic correlates of f_0 , intensity, spectral balance, and duration. From the set of built models, the minimal adequate model, i.e. that with the lowest *AIC* value, was reported (§ 5.3.2.3) and discussed (§ 5.3.2.3).

Hypothesis 2: acoustic correlates of prominence

The manipulations conducted on the speech signal allowed to administer the stimuli in three different conditions. In the case of f_0 , the intonation curve was kept within a range of 2 ST between the lowest and the highest f_0 values; and intensity was neutralised to a constant amplitude of 69 dB. Duration was not

manipulated, partly because there did not seem to be an easy way to neutralise its prominence-lending properties without incurring in a perceptual mismatch between the sound of the speech signal and the articulation movements performed by the speaker in the image.

Hypothesis 2.1 was supported by the results of this Experiment II, as it was in Experiment I. However, hypothesis 2.2 was rejected, since duration was consistently found to strongly influence participants' marking performance, and also drive the process of prominence perception, even when it was the only available cue. Together with duration, in conditions where f_0 was available to participants—Exp0 and Exp2—this acoustic correlate also had a strong effect on the way participants marked prominent words. In its turn, intensity seemed to have a lesser effect than f_0 and duration on participants' perception. Thus, when all acoustic cues were available in Exp0, intensity did not prove significant, and participants gave prominence marks led by the perceptual effects of f_0 and duration (see Figure 66 for a comparison).

However, also in the control condition Exp0, spectral balance seemed to play a more important role than intensity, even if it did not reach significance ($p = .07$, see Table 29). Interestingly, in Exp1, where f_0 had been neutralised, intensity did seem to be more relied on by participants, next to duration, despite the fact that this effect was more modest than that observed for the other acoustic correlates of prominence. More precisely, in the two models reported in this first condition Exp1, intensity fell short of significance in the first one ($p = .05$) and proved significant in the second one, while spectral balance did not seem to play any important role in this experimental condition (Figure 55). In the case of the second condition Exp2, both f_0 and duration, as the only acoustic cues available to participants, were significantly used by participants in the process of rating. Finally, in the third experimental condition Exp3, where both f_0 and intensity

had been neutralised, duration proved sufficient for participants to consistently mark prominent words.

This is in line with previous results supporting the cross-linguistic role played by duration in producing and perceiving phrasal prominence in combination with the perceptual effects of at least another correlate (Kohler, 2005; Mo, 2008b; Ortega-Llebaria, 2006; Vogel et al., 2016). For example, in a similar study on the perception of phrasal prominence conducted in English—but only in the auditory modality—, Mo (2008a) concluded that duration determined participants' marks of prominence; and she also reported a strong effect of spectral balance next to that of duration. In the case of Spanish, Ortega-Llebaria (2006) observed that syllables became longer when they were realised together with a pitch accent (Figure 3). Additionally, this correlation between f_0 and duration was also reported for Spanish by Vogel et al. (2016) in the context of unstressed vowels, who concluded that duration had the lead in cueing phrasal prominence (Figure 4).

As mentioned earlier, the results obtained in Experiment II show that the perceptual effect of intensity in combination with duration seems less strong than the joint effect of f_0 and duration. This might be due to the mentioned lengthening effect of pitch accents on stressed syllables. However, evidence for the combined role of duration and intensity/spectral balance on the production and perception of phrasal prominence has also been reported by several authors, (e.g. for English Kochanski et al., 2005; Mo, 2008a; Silipo & Greenberg, 1999, 2000; for Dutch Sluijter & van Heuven, 1996b; Sluijter et al., 1997). Although in both cases the importance of duration seems unquestioned, the support for either f_0 or intensity/spectral balance as an auxiliary cue to duration is reminiscent of the long-standing debate that confronted advocates of the melodic accent against those defending the role of loudness/articulatory effort (e.g. Sievers, 1901; Stet-

son, 1928; Navarro Tomás, 1964) (§ 2.1.4 and § 2.1.8.7).

Hypothesis 3: gestural correlates of prominence

The results obtained in this Experiment II show little difference between modalities when all experimental conditions were compared to the control as in Experiment I, in which an overall difference between modalities was observed. More precisely, Experiment I revealed a lower perceptual threshold in participants for f_0 and duration in the audiovisual modality (Figure 40a and Figure 40b). However, unlike Experiment I, this was not observed when the three experimental conditions were compared against the control condition in Experiment II; although an effect of modality on duration was found in the second condition Exp2 in the opposite direction. In this second condition, where intensity had been neutralised, the same increase in syllable length in both modalities had a stronger perceptual effect on participants relying on the visual cues of prominence (Figure 58a). In other words, for a word to be considered prominent participants needed it to be longer in the audiovisual modality than in the audio-only modality, despite counting on the visual cues of prominence. In this sense, previous studies have observed that the visual cues of prominence have an influence on the perception of both f_0 and intensity (Foxton et al., 2010; Scarborough et al., 2009), with a slightly stronger effect in the case of intensity, which Foxton et al. related to its more clearly perception as a result of visible articulatory effort.

However, the differences observed between Experiment I and Experiment II for the effect of the visual cues of prominence on the perception of the acoustic cues of f_0 and duration might be due to a difference in the number of stimuli and their range of acoustic values. On the other hand, the results obtained in Experiment II indicate that in the absence of intensity—as in Exp2—perception of syllable length seems to vary between modalities. In this sense, the neutralisation

of intensity might be responsible for this difference, especially because of the perceptual effect pointed out for the combination of intensity and duration (Turk & Sawusch, 1996).

Furthermore, the results for the effect of modality on the perception of prominence only overlapped partially in both Experiment I and Experiment II. Despite the lack of overall main effect of modality, hypothesis 3.1 was supported by the role played by the visual cues of prominence in the marks given by participants. On the one hand, there was a general tendency for participants to mark more words in the audiovisual modality, but unlike Experiment I, this was not found significant in Experiment II. On the other hand, when participants could not rely on all acoustic cues of prominence in Experiment I, this trend decreased and words were less likely to be considered prominent in the audiovisual modality. In Experiment I this had proved significant in the cases where f_0 had been neutralised in C1 and when both f_0 and intensity had been neutralised in C3. However, despite observing a similar trend in Experiment II in all experimental conditions, this effect only reached significance in the second condition Exp2—neutralised intensity (Figure 49).

One possible interpretation of this trend is that the uncertainty caused by the absence of clear acoustic cues—as when both f_0 and intensity were neutralised—result in more random marking of prominence. It is possible that the information available in the audiovisual modality makes participants concentrate their prominence marks around the more clearly perceived visual cues of prominence. The role played by the different gesture phases in the audiovisual modality showed a tendency for words to be considered less prominent when accompanied by strokes, holds, and recoils, but only the recoil phase of gestures was found significant. Several authors have underlined the role played by visual cues, especially beat gestures conducted with head, eyebrows, or hands, in enhancing the per-

ception of acoustic prominence (e.g. Al Moubayed & Beskow, 2009; Granström et al., 1999; House et al., 2001; Krahmer et al., 2002a,b; Krahmer & Swerts, 2007; Prieto et al., 2011). Additionally, some studies on the neural integration and processing of gesture and speech have also pointed out that one of the functions of beat gestures might be driving listeners' attention and helping them process relevant aspects of the spoken signal (Biau & Soto-Faraco, 2013; Biau et al., 2015).

In this sense, the contribution of visual cues to participants' processing of acoustic prominence has been mainly driven by the stroke phase of gestures, which rejects hypothesis 3.2, as reported in Experiment I. In the general comparison of the experimental conditions against the control group, strokes were observed to increase the chances of marking prominence in the audiovisual modality. This effect was also observed in a similar comparison in Experiment I (Figure 40f) and both in the control condition Exp0 (Figure 51d) and in the first condition Exp1 (Figure 54f) of Experiment II.

Interestingly, participants' marks in the audiovisual modality did not seem to be generally prompted by apexes, which actually reduced the chances of the words they co-occurred with to be considered prominent. This seeming hindering effect of apexes was observed both in Experiment I for the comparison of all conditions against the control (Figure 40e) and in Exp1 of Experiment II, where f_0 had been neutralised (Figure 54g). However, a contrary effect for the role of apexes was found when f_0 and duration were the only acoustic cues of prominence in Exp2. In this case, apexes actually made words more likely to be considered prominent in the audiovisual modality (Figure 58c). This might be due to the more preponderant role of f_0 as an acoustic cue of prominence, especially since a temporal coordination between apexes and f_0 peaks has been consistently observed in different studies (e.g. Esteve-Gibert & Prieto, 2013; Leonard & Cummins, 2010; Loehr, 2004). Curiously, in the control condition Exp0, where

participants relied on all acoustic cues of prominence including f_0 , this same effect of apexes was not observed; rather, words tended to be perceived more prominent in the audiovisual modality when they co-occurred with strokes, as just mentioned (Figure 51d).

Hypothesis 4: musical training

In this Experiment II also a different performance of participants according to their level of musical training was observed, which was not tested in Experiment I due to the small sample size of participants. Since behavioural and neural differences have been reported in numerous studies for the effect of musical training in the auditory domain (e.g. Hutka et al., 2015; Parbery-Clark et al., 2009; Thompson et al., 2004), information about participants' level of formal musical training was collected through a questionnaire after the experimental session.

A difference due to participants' level of musical training was only found in the first condition Exp1, in which the prominence-lending cues of f_0 had been manipulated by keeping the intonation curve within a 2-semitone range. Interestingly, no such effect was observed in the control condition or in Exp3, where f_0 , together with intensity, had also been neutralised. In Exp1, despite the neutralisation of f_0 , those participants having more than 5 years of formal musical training showed more probability to mark words as prominent than those without any musical training and with fewer than five years of formal musical training.

Musical training also had an influence on words considered prominent when modalities were compared, so that a similar group of highly-trained musicians showed much less probability to mark words as prominent in the audiovisual modality. It seems that participants with musical training strongly relied on their developed pitch-discriminatory skills—despite the narrow f_0 range—in order to

mark prominent words in the audio-only modality; however, a similar group of participants in the audiovisual modality did not. This is in line with the behavioural differences previously observed in skilled musicians when performing pitch discrimination tasks (e.g. [Hutka et al., 2015](#)). In these results, musicianship seems to play a role in certain adverse acoustic conditions, although the fact that this was restricted to the neutralised effect of f_0 only in the first experimental condition needs further research.

Behavioural differences previously observed in musicians have been related to distinct neural mechanisms from those of non-musicians both in music (e.g. [Amemiya et al., 2014](#); [Liang et al., 2016](#)) and speech (e.g. [Hutka et al., 2015](#)). Nevertheless, musicianship has also been observed to have an effect on the multimodal perception of speech. Generally, the *McGurk effect*—by which an auditory /ba/ combined with a visual /ga/ is generally perceived as /da/ ([McGurk & MacDonald, 1976](#))—has been observed to modulate the cortical processing of auditory signals at an early stage ([Colin et al., 2002](#); [van Wassenhove et al., 2005](#)); similarly, an early modulation of auditory speech information has been reported for manual gestures ([Biau & Soto-Faraco, 2013](#); [Biau et al., 2015, 2016](#)). However, people having extensive musical training have been reported to differently integrate auditory and visual input when compared to non-musicians ([Lee & Noppeney, 2011](#); [Paraskevopoulos et al., 2012](#)); and such a difference has also been observed for the McGurk effect in the multimodal perception of speech ([Proverbio et al., 2016](#)). In the study conducted by Proverbio and her colleagues, musicians were less sensitive to the McGurk effect than non-musicians, i.e. their musical training helped them to correctly identify phonemes despite incongruent visual information. In this sense, the findings obtained in this research are in line with the observation that musical training improves speech-in-noise perception (e.g. [Strait & Kraus, 2011](#)), and that the cognitive processing of pitch shows a shift

in the neural networks when perceiving speech in adverse acoustic conditions (e.g. Zendel et al., 2015).

Hypothesis 5: participant gender

Finally, it was surprising to also observe a difference due to gender in some experimental conditions. More precisely, in Exp2 and Exp3—with f_0 and duration, and only duration as respective acoustic cues—women were observed to be more likely to consider words as prominent, although this difference did not show any interaction with modality. At first sight, this difference was interpreted in Exp2 as possibly stemming from the unbalanced distribution of participants' gender across modalities, with 27 women versus 3 men in the audio-only modality (Figure 60). Since the results revealed a general trend for participants to mark fewer words in the audiovisual modality, the observed gender difference seemed to be due to an actual difference between modalities. Nevertheless, the same gender difference was found in the third experimental condition Exp3, even though no such an unbalanced gender distribution across modalities existed (Figure 65). In this case, unlike Exp2, no similar difference between modalities was observed.

Previous research has reported behavioural and neuroanatomical differences between men and women in the utilisation of visual cues from lip and facial movements when recognising different speech sounds (e.g. Dancer et al., 1994; Öhrström & Traunmüller, 2004; Ruytjens et al., 2006, 2007; Watson et al., 1996; see Alm & Behne, 2015, for a summary).

Although the existence of gender differences in language perception is far from settled, women have been observed to perform better at speech-reading than males (e.g. Dancer et al., 1994; Strelnikov et al., 2009), which has been related to the fact that women could be more active gazers than men (e.g. Johnson

et al., 1988). Additionally, women have been reported to be more sensitive to visual cues than men in audiovisual speech perception (Aloufy et al., 1996; Irwin et al., 2006; Öhrström & Traunmüller, 2004). For example, Öhrström and Traunmüller (2004) showed that women were significantly more influenced by the visual modality than men in perceiving incongruent Swedish vowels embedded in a syllable. Similar results were reported by Irwin et al. (2006), who studied the influence of visual speech for the syllable /ba/. Irwin et al. suggested that such a gender difference might be due to a different pattern in language processing, with a stronger activation in bilateral brain areas causing a more efficient audiovisual language processing in women (e.g. Baynes et al., 1994; Coney, 2002).

Furthermore, neuroanatomical differences point to a stronger activation of brain areas associated with speech perception in women (Ruytjens et al., 2006, 2007). In addition, it has been claimed that gender differences in audiovisual speech perception may emerge in the context of challenging stimuli (Jaeger et al., 1998), which can be related to the results observed in the third experimental condition, Exp3, in Experiment II, where duration was the only acoustic cue available to participants.

Limitations

One of the limitations of this Experiment II lies in the small number of sentences rated by participants. The very experimental design, which was imposed by the large number of participants taking part in the experiment and the completion of the experimental task via the Internet, made necessary to reduce the time of the experimental task and consequently reduce the number of stimuli. Additionally, unlike Experiment I, where participants rated a much larger number of stimuli under supervision, the experimental task in Experiment II was

Experiment II

not conducted in a sound-proof cabin in the laboratory. The uncertainty associated to the conditions in which participants conducted the experimental task is a downside to the use of the Internet in the methodology employed in this Experiment II.

General discussion and conclusions

Previous insights gained into the multimodal perception of acoustic prominence have been reinforced by the experimental research conducted in this study. The two experiments presented here have showed that it is possible to add ecological validity to prior studies by using spontaneous speech samples extracted from television. It is also the first time that a study like this one has been conducted for Castilian Spanish.

The results yielded by the two experiments conducted in the course of this research have been able to cast light on the questions that gave rise to it:

1. How do the different acoustic correlates relate to one another and to gestures in the perception of prominence?
2. How do gestures contribute to the perception of prominence?

These two questions arose from the shortcomings found in previous methodologies and were intended to analyse multimodal prominence perception in a very different context from that of studies using controlled speech with either animated agents or elicited gestures in the laboratory.

Probably, one of the most interesting conclusions that can be drawn from this research is the importance of the temporal aspect in the perception of prom-

inence, i.e. the effect that both the acoustic correlate of duration and the stroke phase of gestures may play in multimodal prominence perception when compared to other perceptually relevant information.

In this research, the marks of prominence given to words by participants in the different experimental conditions showed that identification of phrasal stress was possible even if duration was available as the only acoustic cue. In addition, whichever the acoustic cues available in each condition, they were consistently used by participants to provide marks of prominence, whether it was only duration, f_0 and duration or intensity and duration. However, differently from intensity, spectral balance did not play any role when combined with duration. Rather, spectral balance was more strongly relied on when all acoustic cues of prominence were present, even if it did not prove significant.

The results obtained in this research support previous findings for the role of duration as a main cue to phrasal stress. Initially, duration was considered in previous research second to f_0 in the perception of lexical stress (Bolinger, 1958; Fry, 1955, 1958). When the acoustic correlates of lexical stress and phrasal stress were later analysed separately, new results advanced that duration was a consistent role of lexical stress, while the perceptual effects of f_0 were considered to cue phrasal rather than lexical stress (e.g. Sluijter & van Heuven, 1996a; Sluijter et al., 1997). However, several authors also reported the importance of duration in cueing phrasal stress in German (Kohler, 2005), in English (Mo, 2008a), and in Spanish (Vogel et al., 2016).

In the case of Spanish, research has been influenced by a strong tradition of studies on lexical stress. Ortega-Llebaria pointed out the necessity to study the acoustic correlates of lexical stress in the context they occur (Ortega-Llebaria, 2006). Thus, she made evident that duration tended to increase in the presence of pitch accents and nuclear pitch accents for oxytone words (Ortega-Llebaria,

2006). The lengthening effect that the realisation of a pitch accent—thus, potentially signalling phrasal prominence—has on duration was found controversial, with cross-linguistic differences that have been accounted for on different grounds (Beckman & Edwards, 1994; Ortega-Llebaria & Prieto, 2007; Sluijter & van Heuven, 1996a,b) (§ 2.1.8.1). The perceptual results obtained here point to a consistent use of durational cues to detect phrasal prominence in the control condition, actually accounting for slightly stronger perceptual weight than f_0 . Thus, it was observed that the latter two cues were consistently used by participants, while overall intensity and spectral balance were not perceptually relevant to detect prominence under normal acoustic conditions. This partially confirms the results obtained by Vogel et al. (2016), who observed that duration and f_0 were consistent perceptual cues of unstressed vowels accompanied by a pitch accent, although they found a stronger perceptual effect than that found in this research for intensity rather of f_0 in the case of stressed vowels accompanied by a pitch accent (§ 4). This difference can be put down to the more fine-grained results they obtained for both stressed and unstressed vowels in Spanish sentences, while the present results did not target any specific environment but the general perceptual effect of each cue to detect prominence.

The approach used here is more in line with the one used by Mo (2008a) in her study on prominence perception, in which it was concluded that duration determined participants' marks of prominence, although Mo also found a strong effect of spectral balance next to that of duration. Even if the results presented here did not show any role for spectral balance or intensity to cue prominence under normal acoustic conditions, intensity was very likely a relevant cue when the perceptual effects of f_0 had been neutralised (Figure 53 and Figure 55). In this sense, it was found that the perceptual effect of intensity in combination with duration seems less strong than the joint effect of f_0 and duration. This might be

due to the mentioned lengthening effect of pitch accents on stressed syllables. In any case, it is worth noting that in the degraded-speech paradigm used here, it was observed that whichever the acoustic cues available to participants in each condition, they were used to provide marks of prominence, whether it was only duration, f_0 and duration or intensity and duration.

On the other hand, the results obtained in this research showed that the perceptual effect of the acoustic correlates of prominence was affected by the presence of visual information. This was the case of both f_0 and duration in certain contexts, showing that the perceptual threshold for these two acoustic correlates may vary in the presence of the visual cues of prominence. For example, in Experiment I the stressed syllable of words needed to be higher in pitch and longer in duration in the audio-only than in the audiovisual modality to be considered prominent. The fact that this difference was not observed in the same way in Experiment II might be due to the small sample of target sentences used in this second experiment. Nevertheless, these differences observed in the results of both experiments require further research.

As for the effect of the visual information on the marks of prominence given by participants, marking was different between modalities when the experimental conditions were compared to the control group. On the one hand, the results of both comparisons Experiment I and Experiment II showed differences in modality as a main effect. On the other hand, as observed in the comparison between both experiments, the visual cues of prominence generally tended to increase the chances of words to be perceived as prominent; while in adverse acoustic conditions, when some acoustic cues of prominence were missing, this trend was reversed, and the visual cues of prominence generally tended to make words less likely to be considered prominent. Possibly, the uncertainty caused by the absence of clear the acoustic cues resulted in more random marking of prominence;

thus, the visual information might help participants concentrate their prominence marks around the more clearly perceived visual cues of prominence. In other words, under normal acoustic conditions, the audiovisual modality makes the listener consider more words as prominent than the audio-only modality. However, in the experimental conditions, with fewer cues to detect prominence, the perception of prominence becomes more challenging in the audio-only modality, and the effect of the visual cues of prominence does not induce to consider more words as prominent. For example, in Experiment I, 3.6% more words were marked in the audiovisual modality in the control condition (Table 16), and this difference reached 5.7% in Experiment II (Table 23), while smaller, even negative differences were observed between modalities in the experimental conditions in both experiments.

There has been a large number of studies on the interaction between gesture and speech that have corroborated the strong connection between them. For example, not only are gestures temporally aligned to speech, but they also enhance the perceptual effects of the acoustic correlates of prominence (e.g. De Ruiter, 1998; Krahmer & Swerts, 2007, experiment 1; Krivokapić et al., 2015, 2016; Leonard & Cummins, 2010; Rochet-Capellan et al., 2008; Rusiewicz, 2010). Similarly, a more reduced number of studies (Krahmer & Swerts, 2007, experiments 2 and 3) have found that a visual cues have an influence on how speech stimuli are perceived acoustically. Such findings are supported by the results obtained in this research, showing that, under normal acoustic conditions, the gesture phase of strokes increased the chances of marking prominence in the audiovisual modality. Surprisingly, apexes did not seem to lead prominence perception in this case. By the same token, by using in a degraded-speech paradigm, it was made evident that the apex phase of gestures seems to account for less perceived prominence than strokes when f_0 was neutralised—and intensity and duration were

still present as acoustic cues. This might be due partly to the temporal coordination of both apexes and f_0 , which could play an important role in the perception of prominence. Similarly, when intensity was neutralised, and f_0 and duration were the only acoustic cues, apexes strongly contributed to lead the perception of prominence in the audiovisual modality. So this shows how the different phases of gestures contribute differently to enhance the perceptual effects of the acoustic correlates of prominence.

Initially, the perceptual effects of the visual component of speech focused on the so-called McGurk effect (McGurk & MacDonald, 1976), by which articulatory lip movements affect speech perception. Also the rest of the face (e.g. Pelachaud et al., 1996) was later reported to affect speech perception, and as a result, it was observed that both facial expressions and body movements play an important role in conveying functions traditionally associated to prosody, such as phrasing and emphasis. This interaction between the visual correlates of prominence and speech prosody was dubbed as ‘audiovisual speech’ (Swerts & Krahmer, 2005), and research on this visual component of communication has systematically explored the nature of this interaction (e.g. Al Moubayed et al., 2010; Granström et al., 1999; Kim et al., 2014; Krahmer & Swerts, 2007; Prieto et al., 2011; Scarborough et al., 2009). More precisely, visual cues were observed to result in a stronger production and perception of verbal prominence (Krahmer & Swerts, 2007), and facial gesturing was also found to systematically influence the perception of verbal prominence (Dohen & Løevenbruck, 2009; House et al., 2001; Swerts & Krahmer, 2008).

The temporal coordination of gesture and speech has previously been analysed in detail, and the phonological synchrony rule put forward by McNeill (McNeill, 1992) has been supported by different studies (e.g. Esteve-Gibert & Prieto, 2013; Krivokapić et al., 2015, 2016; Leonard & Cummins, 2010; Loehr, 2004).

A temporal alignment has been confirmed between pitch accents and the apex phase of gestures—the peak of effort that occurs at an instant in time, i.e. the “kinetic goal of the stroke” (e.g. [Jannedy & Mendoza-Denton, 2005](#); [Loehr, 2004](#)).

In this study, the criteria for the collection of a corpus with speech material from a TV talent show reflected very well the natural interweaving of gestures performed with hands, head, and eyebrows typically found in everyday spoken language. Thus, the corpus abounded in gestures, so that hardly any words in the corpus occurred without the presence of one. The annotation omitted the classification of gesture types, but included annotation of gesture phases. This annotation presented some difficulties, especially when gestures included several articulators. Thus, when the gesture was performed with hands together with any other articulator, the annotation followed the most visible movement of the hands, since the second articulator—or possibly the remaining two articulators, i.e. head and eyebrows—hardly presented any preparation or stroke phase, but the gesture was sudden and its apex coincided with that of the apex phase of the hand gesture.

In this respect, the prominence marks given by participants in the two experiments conducted in this research seem to have been consistently driven by the stroke phase of gestures rather than the apex when the visual cues of prominence were available to them, despite differences between Experiment I and Experiment II in the number of marks given to strokes ([Table 37b](#) and [Table 46a](#)). As previously mentioned, apexes had a stronger effect than strokes when intensity had been neutralised and f_0 and duration were the only acoustic cues of prominence. In addition, neither of these two gesture phases seemed to play any determinant role when both modalities were compared in the most adverse acoustic conditions—only duration as an acoustic cue—, which suggests that such challenging conditions made very difficult for participants to clearly rely on either

gesture phase to consider words as prominent.

Thus, it is possible to conclude that the temporal aspect common to both the acoustic correlate of duration and the stroke phase of gestures may play a fundamental role in multimodal prominence perception when compared to other perceptually relevant information. In this sense, previous studies have underlined the importance of prosodic lengthening in signalling focused constituents (Baumann et al., 2007, for German; Eady et al., 1986; Watson et al., 2008, for English; Jun & Lee, 1998, for Korean), an effect that has also been found to be correlated to higher f_0 . However, in the case of Spanish, lengthening has been observed in syllables carrying nuclear stress that also keep the typical low f_0 of declarative sentences (Escandell-Vidal, 2011). Such observations were made in cases of *verum focus*, which has been associated to the values of impatience and insistence introduced by the repetition of given information (Escandell-Vidal et al., 2014). The conclusion pointed out by Escandell and her colleagues about the independence of duration from f_0 when signalling prominence is supported by the results obtained in this research. Similarly, the stroke phase of gestures has traditionally be considered as the nucleus of the gesture and is temporally aligned with stressed syllables (Kendon, 1972; McNeill, 1992). Despite the fact that apexes have been seen to align more precisely with pitch accents (Loehr, 2004; Jannedy & Mendoza-Denton, 2005), the present research suggests that the longer duration of strokes afford more perceptual salience in the multimodal processing of speech.

These interesting findings were also complemented by the observation of a different distribution of gestures according to the articulator involved in their production when compared to the distribution of gestures reported in a previous study using spontaneous speech (Ambrazaitis & House, 2017). Previous methodological approaches had limited themselves to the study of isolated gestures,

mostly beat gestures performed with either head or eyebrows, and to a lesser extent hands (e.g. Al Moubayed et al., 2010; House et al., 2001; Krahmer et al., 2002a,b; Prieto et al., 2011; Swerts & Krahmer, 2010). In their study, Ambrazaitis and House (2017) analysed the use of eyebrow and head movements to convey information structure using recordings of TV newsreaders and found that eyebrow movements were rarely produced in isolation but occurred much more frequently together with a head movement. In Experiment I, the use of spontaneous speech material, similar to the study of Ambrazaitis and House, allowed to use stimuli containing more than one single gesture, thus reflecting more faithfully the natural interweaving of gestures performed with hands, head, and eyebrows typically found in everyday spoken language. However, the results obtained here showed that gestures were mostly produced by combining especially hands and head (51.1% of the occurrences); and, to a much lesser extent, hands, eyebrows, and head (11.1%). In the case of gestures that were produced with a single body part, it was mostly the hands that performed the gesture (23.5%). In this sense, differently from the results offered by Ambrazaitis and House, very few occurrences of gestures were produced with eyebrows and head (1.1%) (Figure 33). This difference might be due to the limited expressivity of the source of spontaneous speech chosen by Ambrazaitis and House—TV newsreaders—, especially when compared to more expressive samples from the speech material used in this study, i.e. spontaneous speech from television talent shows. Therefore, further research is needed to analyse the role of each articulator separately and in combination in the multimodal perception of prominence, possibly with a similar methodology using spontaneous speech stimuli obtained from a similar source.

In addition, the results obtained here clearly show that the musical training of participants should be taken into account in any study on prominence perception depending on the perceptual effects of f_0 . Although the exact advantages of a

musically-trained ear in the perception of multimodal prominence needs further investigation, it is evident that the prominence-lending properties of f_0 can be exploited by individuals with solid musical training, especially when they rely on minimal variations of pitch to detect prominence.

Previously, behavioural differences have been observed in musicians when compared to non-musicians. This has been related to distinct neural mechanisms resulting from consistent musical training (e.g. Amemiya et al., 2014; Hutka et al., 2015; Liang et al., 2016). Furthermore, consistent musical training has also been reported to have an effect on the multimodal perception of speech. The *McGurk effect* (McGurk & MacDonald, 1976) has been observed to modulate the cortical processing of auditory signals at an early stage (Colin et al., 2002; van Wassenhove et al., 2005); and by the same token, an early modulation of auditory speech information has been reported for manual gestures (Biau & Soto-Faraco, 2013; Biau et al., 2015, 2016). Nevertheless, people having consistent musical training integrate auditory and visual input in a different way when compared to non-musicians (Lee & Noppeney, 2011; Paraskevopoulos et al., 2012), and such a difference has also been observed for the McGurk effect in the multimodal perception of speech (Proverbio et al., 2016). Thus, the findings obtained in Experiment II are in line with the observation that musical training improves speech-in-noise perception (e.g. Strait & Kraus, 2011), and that the cognitive processing of pitch shows a shift in the neural networks when perceiving speech in adverse acoustic conditions (e.g. Zendel et al., 2015). Nevertheless, a more clear picture of the role of consistent musical training in the multimodal perception of prominence will be yielded by future research.

On the other hand, the different way of conducting the marking task observed in women and men in certain conditions point to a stronger reliance on the visual cues of prominence by women, especially in adverse acoustic conditions,

although no differences between modalities were found. Previous research has reported behavioural and neuroanatomical differences between men and women for the audiovisual perception of speech (Dancer et al., 1994; Öhrström & Traunmüller, 2004; Ruytjens et al., 2006, 2007; Watson et al., 1996). More precisely, women have been observed to perform better at speech-reading than males, which has been related to the fact that women could be more active gazers than men (e.g. Johnson et al., 1988). Additionally, women have been reported to be more sensitive to visual cues than men in audiovisual speech perception (Aloufy et al., 1996; Irwin et al., 2006; Öhrström & Traunmüller, 2004). As for neuroanatomical differences, a stronger activation of brain areas associated with speech perception has been reported for women (Ruytjens et al., 2006, 2007); and it has been claimed that gender differences in audiovisual speech perception may emerge in the context of challenging stimuli (Jaeger et al., 1998), which can be related to the results observed in the third experimental condition Exp3, where duration was the only acoustic cue available to participants. However, a more detailed account of such a behavioural difference deserves further analysis.

As for the methodological details of this research, it is worth mentioning that the estimation of generalised linear mixed models (GLMMs) and their comparison by means of the Akaike Information Criterion *AIC* (Akaike, 1973) has proved crucial to determine which variables determined the marks of prominence given by participants. Despite the fact that such a statistical approach based on information criteria is still relatively uncommon in the field of linguistics (e.g. Adamou et al., 2018), it has been gaining popularity over the last decade in other fields such as biology, ecology, or evolution, where it is now often applied (see e.g. Arnold, T. W., 2010; Grueber et al., 2011, for details).

The implications of these insights are expected to fruitfully contribute to the debate on the role played by the different acoustic and visual cues in the percep-

tion of phrasal prominence. Also important is the fact that the results derived from this experimental research have been obtained for Castilian Spanish, which is certainly to be of great interest for the field of Hispanic linguistics. Thus, these results not only offer a better understanding of the multimodal perception of prominence, but also suggest that musical training and gender can play an important role in this process. Even though further research is needed to provide a more solid account of some of the elements involved in the auditory and visual perception of speech, it has been possible to observe that, in the course of this research, a different methodology could help to answer some questions that needed to be addressed.

More generally, this research is added to the efforts of current studies investigating the visual aspect of speech and communication, and they offering a more complete and complex picture than that we have until recently had. The visual component of speech serves important communicative functions that cannot be neglected, with important implications in pragmatic issues. Body movements performed as eyebrow raises, head nods, and manual gestures are a crucial part of everyday interactions, and are tightly linked not only to prosodic functions, such as prominence, phrasing, intonation, salience, turn-taking, and grounding in face-to-face interactions (e.g. [Krahmer & Swerts, 2007](#); [Kushch & Prieto Vives, 2016](#); [Nakano et al., 2003](#); [Prieto et al., 2011](#); [Srinivasan & Massaro, 2003](#)), but often they are also a necessary expressive resource in the conveyance of ideas, abstract and concrete alike (e.g. [Cienki, 2005](#); [Novack & Goldin-Meadow, 2017](#)). These visual cues are also fundamental in signalling affective speech, including emotions considered as universal—happiness, surprise, fear, sadness, anger, disgust, and interest—as well as emotions emerging in social contexts that may include uncertainty and frustration, so commonly expressed in human interactions (e.g. [Barkhuysen et al., 2005](#); [Ekman, 1999](#); [Swerts & Krahmer, 2005](#)).

With the insights gained here, this research intends to humbly contribute to a more comprehensive knowledge of the broader phenomenon of communication, in which the role of the visual component of speech has been put in relation with both the ontogenesis and the phylogenesis of language. The integration of gesture and speech is more and more often taken into account in studies on language development (e.g. Iguualada et al., 2015; Iverson & Goldin-Meadow, 2005). Some of the communicative milestones achieved by infants are mediated by the use of hand movements and the joint use of deictic gestures and speech in social interactions. Infants' ability to selectively use this multimodal communicative strategy is crucial in drawing the adult's attention towards referential cues, an aspect that has proved decisive in the acquisition and development of the semi-otic component of language (e.g. Carpenter et al., 1998; Rodríguez et al., 2015).

In the same way as Labov's (1968) concept of *speech community* implies the "uniformity of abstract patterns of variation", one might wonder whether similar patterns of variation also emerge in gesturing, and thus, whether it is possible to speak of 'gesture communities'. Cross-cultural variation in gestures accompanying speech, as well as emblems, is well-attested (e.g. Kita, 2009)—which seems to confirm the popular belief that Spaniards or Italians do gesticulate more than northern Europeans. Thus, the question arises whether speech communities do also show similar patterns of variation in their gestural behaviour. The repertoire of body movements involved in gesticulation seems to have a highly idiosyncratic character that may also be formed in narrow contact with the repertoire of gestures displayed by those individuals taking part in one's daily communicative interactions (e.g. Mol et al., 2012; Noland, 2009; Ricci Bitti, 2013; Schwarts, 1995).

This social aspect of human communication—in which gestures are not to be considered as less important than speech—has led to reconsider the role of

gestures in its relation to language (e.g. McNeill, 1992; Kita, 2000). In this sense, several events in the last decades have revived the idea of a possible gestural evolution of language (e.g. Armstrong et al., 1995). On the one hand, the reflections made by Bickerton in the 90's on infant speech, pidgin and creole languages, and primate communication brought into the middle of the linguistic debate the qualitative leap existing between primate communication and our articulated language, a leap known as the *continuity paradox* (Bickerton, 1990). The traditional view that primate vocal calls were the antecedent of articulated speech was contested when the gestural communication of primates started to be better understood and primates were observed to intentionally use arm and hand gestures to influence the behaviour of other members of their groups (e.g. Tomasello & Call, 1997). On the other hand, the discovery of mirror neurons also showed that the motor neural networks controlling the movements of both arm and vocal tract partially overlap in Broca's area (Gentilucci et al., 2008), a structure that is related, on evolutionary principles, to the equivalent area in the brain of primates and responsible for motor functions controlling arm and mouth (Gallese et al., 1996; Mukamel et al., 2010).

This sensorimotor link between gesture and speech and its evolutionary implications are complex. Such a tight relationship is reflected in the shared brain mechanisms in language and motor functions such as the activation of certain brain areas during language tasks (e.g. premotor area), which are typically associated with motor functions; the activation of brain areas during motor tasks that are typically associated with language behaviour (e.g. Broca's area); and the patterns of clinical conditions affecting to both language and motor function (Iverson & Thelen, 1999).

An important number of studies published since the second half of the 20th have not only made that gestures are considered more seriously in their close re-

lationship with speech in communication, but they have also made that gestures become to be seen an essential piece to solve the riddle of the emergence of human language. In this sense, this experimental research represents a necessary step towards a more complete understanding of the multimodal perception of speech and, by focusing on the perception the so-called audiovisual prosody, it has also contributed to offer a more complete picture of the role that gestures play in communication. This dissertation has offered some important insights into the research topic as well as it has also opened up new perspectives to be taken up in future research. The attempt to overcome some limitations inherent to previous methodologies by proposing a new methodological approach has proved satisfactory, but it has also made evident that there is room for fine-tuning and for more original ideas that permit to study address problems from new angles.

Appendix A.

Experiment I

Sentences of the corpus: trial sentences correspond to sentences 1 to 10. Sentences 11 to 21 are the non-manipulated stimuli, and the remaining sentences, 21 to 50, served as stimuli target sentences after neutralising their acoustic cues of prominence (f_0 and intensity). As they were presented to participants, no punctuation marks were used.

Trial sentences:

- (1) A veces pequeñas enfermedades ocasionales como caídas un no sé qué unas anginas.
- (2) Cada uno valoramos la situación de distinta manera.
- (3) Eso eso es una cosa que es sola y no lo digo por nadie ni lo quiero decir en general lo dejo lo dejo caer.
- (4) Y y yo lo entiendo porque a mí me pasa y por eso te he dicho por eso te he dicho que no o sea que hemos hecho exactamente lo mismo.
- (5) Y es fantástico que además lo asumas y que encima vayas y le pidas disculpas fantástico me parece maravilloso.
- (6) Y hay días que estamos tristes y estamos tos tristes.
- (7) Ha habido un mogollón eh de cosas a la vez que te han dicho.
- (8) Igual él tiene una vida más independiente y esas manías igual que le puede pasar a ella.
- (9) Pero claro hay veces que dudo y esa inseguridad está porque quizás a veces lo he sentido así.
- (10) A veces soy muy echá palante pero es lo que tú dices yo soy así por algo.

Non-manipulated stimuli:

- (11) Cómo se puede sentir alguien que tenga dificultad de presentarse de una manera extrovertida pues agradable etcétera que sea alguien más tímido.
- (12) Para ver cómo somos para ver si somos conflictivos si podemos valer si no.
- (13) Nosotros no nos hemos portao mal hemos interrumpido no hemos no nos hemos comportao bien en el aspecto ese de que igual entrábamos cuando no teníamos que entrar.
- (14) No te preocupes no sé qué dándola ánimos y ahora es justamente la al al revés no.
- (15) Porque confundimos la mano izquierda con la hipocresía y el ser hipócrita es una cosa y el ser diplomático es otra muy diferente.
- (16) Tú lo has dicho explícito yo no lo he dicho explícito.
- (17) Que haya algo una idea de alguien que esté plasmada en una estrofa o una línea o algo así para que todos tengan aquí la cuestión.
- (18) Pero claro hay veces que dudo y esa inseguridad está porque quizás a veces lo he sentido así.
- (19) También podían haber esto hecho esto cerrado y que los lunes saliera una gala no.
- (20) Hay días que que estás contento y está todo el mundo contento.

Manipulated stimuli¹:

- (21) Yo no tengo que pagar lo que hacen otros es lo único que digo.
- (22) Y me veía y la y la observaba y me cortaba o sea no me salía.
- (23) A que me nominan por no concentrarme porque es que hay que medir lo que dices.

¹ This material is publicly available at <http://dx.doi.org/10.17632/jkvftnpr5j.1>

- (24) Sencillamente me parece una persona absolutamente que nunca tiene una mala palabra para nadie.
- (25) Aparte del sobreesfuerzo ha tenido que luchar contra ella o sea contra ella misma.
- (26) Yo me lo he currao igual que Rosa sabes y yo lo he pasao mal igual que Rosa.
- (27) En ese tipo de relaciones es la donde más fácil te puedes encontrar la mentira.
- (28) Cantar lógicamente es lo más importante es el punto número uno si no no estaríamos aquí no.
- (29) Te lo te lo he dicho o no te lo he dicho al igual que le he dicho joder esto te ha salido muy bien te sale bonito.
- (30) Y cuando leía los esto decía esto me lo ha escrito to mi padre.
- (31) Pero una vez nos lo han dicho es que no ha vuelto a suceder.
- (32) Fíjate tú te sentiste inhibido por una personalidad pues extrovertida.
- (33) Yo me puedo llevar mejor no sé con cierta gente que con otra gente en un momento determinado.
- (34) Todos tenemos más afinidad con unos que con otros.
- (35) O sea a ver pongamos las cosas en su sitio vale perfecto que tú no tienes que reaccionar así.
- (36) Yo creo que ellas buscan cómo eres si les puedes interesar una persona que están tratando contigo a lo mejor tres años grabando ciertos discos.
- (37) Siempre he pecado mucho de lo que de lo que le ha pasado a a Javián y he perdido un montón de gente por eso.

- (38) Estés delante de ocho millones estés delante de veinte siempre habrá uno que no le vas a gustar y eso te tienes que acostumbrar.
- (39) Y lo primero que has ido es a saca a preguntarle a Verónica si le ha sentao mal.
- (40) Tienes que tener un comportamiento básico tienes que saberte comportarte lógicamente.
- (41) Pues eso es lo que estoy intentando hacer por eso estoy concentrao y mirando el papel a ver si puedo poner algo.
- (42) El otro día estábamos leyendo los emails los de allí y yo decía a mí quién me va a escribir.
- (43) Porque yo la conozco a Verónica y Verónica tiene mucho carácter y sabe perfectamente cuando le gusta algo y cuando no.
- (44) Yo no soy el que está interrumpiendo al compañero son ellos díselo a ellos como me lo dirías a mí y no me lo digas a mí.
- (45) Vale él es muy impulsivo y también eso también le sale del corazón.
- (46) O cualquier cosa se lo digo a este porque sé que me va a decir la verdad si se lo digo al otro por quedar bien puede que no me la diga.
- (47) Y te identificas con la persona con la que le preguntas para al mismo tiempo esa persona tenga confianza en ti y revelarte.
- (48) La semana pasada y esta semana Rosa es diferente.
- (49) Y más si lo has hecho con conciencia de lo que estabas haciendo.
- (50) Yo mira yo de hecho había apostado en primer lugar por los dos David por Alex o por Rosa tenía esos cuatro.

Experiment I

Sentence	Words	Seconds	Speech rate
1	13	2.51	5.18
2	16	3.66	4.37
3	16	3.40	4.71
4	14	4.94	2.83
5	13	4.66	2.79
6	18	4.11	4.38
7	15	4.84	3.10
8	17	4.21	4.04
9	28	4.68	5.98
10	14	4.34	3.23
11	13	2.96	4.39
12	10	5.26	1.90
13	18	4.29	4.20
14	9	2.70	3.33
15	18	5.05	3.56
16	25	9.49	2.63
17	24	4.71	5.10
18	24	5.99	4.01
19	18	3.38	5.33
20	11	4.11	2.68
21	20	3.61	5.54
22	20	6.67	3.00
23	20	5.70	3.51
24	27	4.97	5.43
25	13	2.96	4.39
26	32	5.00	6.40
27	23	5.49	4.19
28	9	3.27	2.75
29	13	2.65	4.91
30	20	8.52	2.35

Table A1: *Details of the 50 sentences from the corpus used as stimuli in Experiment I.*

Experiment I

Speaker	Words rated as prominent	Total words	% Prominence
1	244	900	27.11
2	91	264	34.47
3	326	1404	23.22
4	327	1392	23.49
5	157	612	25.65
6	109	408	26.72
7	176	636	27.67
8	84	300	28.00
9	60	216	27.78
10	51	240	21.25

(a)

Listener	Words rated as prominent	Total words	% Prominence
1	156	531	29.38
2	193	531	36.35
3	148	531	27.87
4	90	531	16.95
5	148	531	27.87
6	160	531	30.13
7	108	531	20.34
8	185	531	34.84
9	75	531	14.12
10	135	531	25.42
11	82	531	15.44
12	145	531	27.31

(b)

Table A2: *Details of prominence marks per speaker (a) and per listener (participant) (b).*

Experiment I

Word	Class	Prominence marks	Total words	%
mí	Pron	36	48	75.0
mucho	Adv	24	32	75.0
otros	Pron	23	32	71.8
Rosa	Noun	44	64	68.7
mal	Adv	22	32	68.7
más	Adv	40	64	62.5
lógicamente	Adv	19	32	59.3
bien	Adv	19	32	59.3
uno	Adj	19	32	59.3
ellos	Pron	18	32	56.2
Verónica	Noun	25	48	52.0

(a)

Word class	Prominence marks	Total words	%
Interjection	31	48	64.5
Adjective	234	478	48.9
Noun	383	846	45.2
Adverb	251	590	42.5
Negation	98	256	38.2
Verb	530	2234	23.7
Pronoun	364	1818	20
Conjunction	89	892	9.9
Determiner	36	402	8.9
Preposition	74	906	8.1

(b)

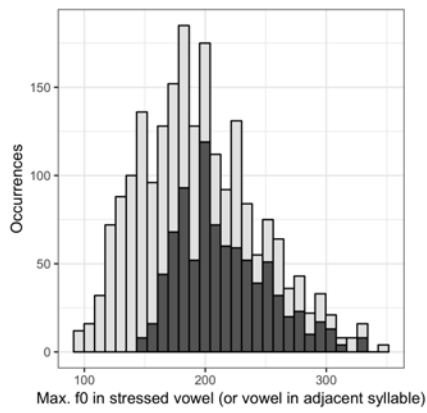
Table A3: Details of precise words and word categories receiving the highest number of prominence marks.

Experiment I

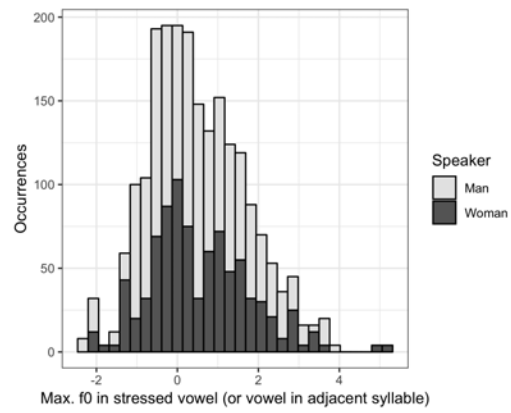
Sentence	HAND	EYEBROWS	HEAD	HA.EY	HA.HE	EY.HE	HA.EY.HE
1	3	0	0	0	2	0	2
2	3	0	0	0	2	0	0
3	4	0	0	0	1	0	0
4	0	1	0	0	3	0	0
5	1	0	0	0	6	0	0
6	0	0	1	0	3	0	0
7	0	0	0	0	4	0	0
8	2	0	1	0	0	0	0
9	0	0	0	0	4	0	0
10	1	1	1	1	6	0	1
11	0	1	0	0	5	0	2
12	0	0	0	0	3	0	1
13	0	0	1	0	4	0	1
14	0	1	0	0	7	0	0
15	0	0	1	0	1	0	2
16	2	0	0	0	1	0	0
17	2	0	1	0	0	0	1
18	1	0	0	0	4	0	1
19	4	0	0	0	3	0	0
20	1	0	2	0	0	0	1
21	0	0	0	0	0	0	1
22	3	0	0	0	2	0	0
23	1	0	0	0	4	1	3
24	1	0	1	0	3	0	1
25	3	0	0	0	5	0	0
26	1	0	1	0	3	0	1
27	3	0	3	0	4	1	1
28	1	0	1	0	4	0	0
29	1	0	3	0	1	0	0
30	2	0	0	0	2	0	0
Total	40	4	17	1	87	2	19

Table A4: Occurrences of gestures performed with different articulators per sentence ($n = 170$). Abbreviations: hand and eyebrows (HA.EY); hand and head (HA.HE); eyebrows and head (EY.HE); hand, eyebrows, and head (HA.EY.HE).

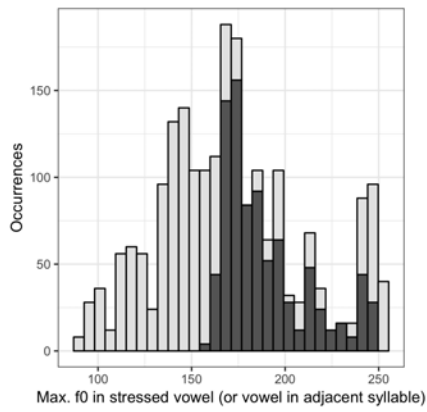
Experiment I



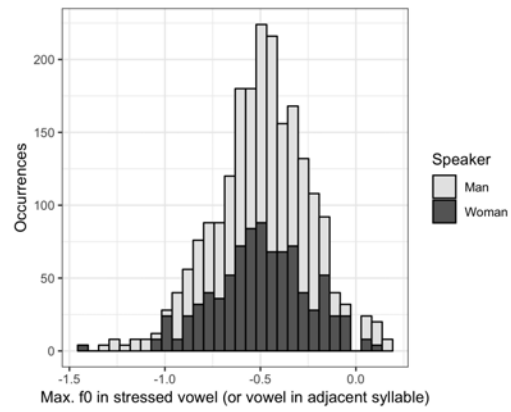
(a) Non-manipulated stimuli



(b) Normalisation in non-manipulated stimuli



(c) Manipulated stimuli



(d) Normalisation in manipulated stimuli

Figure A1: Histograms of f_0 measured as maximum frequency of stressed vowel (or vowel in adjacent syllable) for non-manipulated stimuli (a) and its normalization per speaker and sentence as z-scores (b). Distribution corresponding to neutralised f_0 is showed in (c) together with its normalization (d). Histograms (a) and (b) are representative of conditions C0 and C2. Histograms (c) and (d) correspond to conditions C1 and C3.

Experiment I

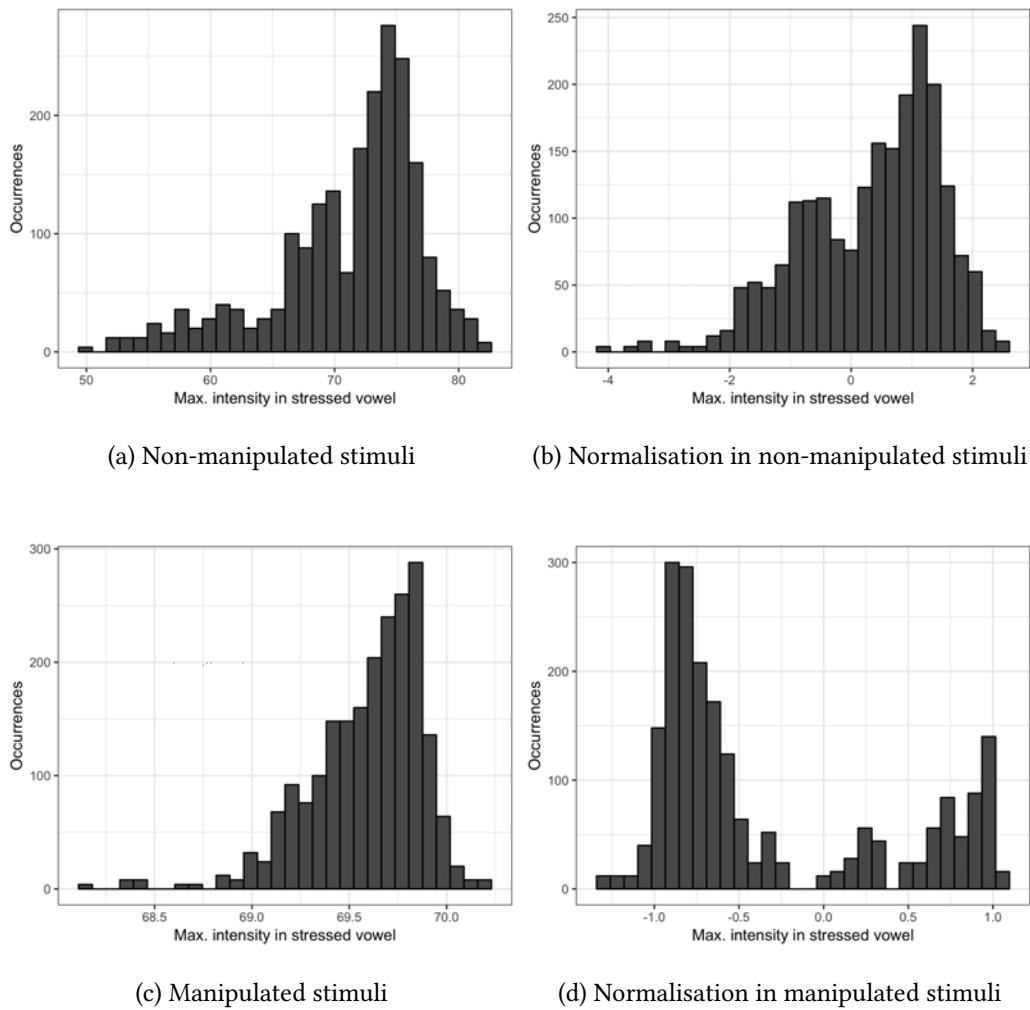
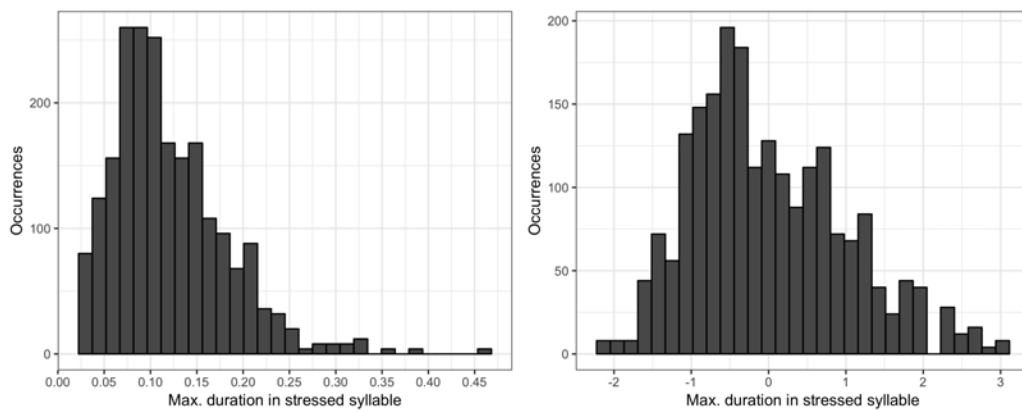


Figure A2: Histograms of intensity measured as maximum amplitude of stressed vowel for non-manipulated stimuli (a) and its normalization per sentence as z-scores (b). Distribution corresponding to neutralised intensity is showed in (c). Histograms (a) and (b) are representative of conditions C0 and C2. Histograms (c) and (d) correspond to conditions C1 and C3..



(a) Non-manipulated stimuli

(b) Normalisation in non-manipulated stimuli

Figure A3: Histograms of duration measured as mean duration of the stressed syllable (a) and its normalization per sentence as z-scores (b). Both histograms (a) and (b) are representative of duration in all conditions.

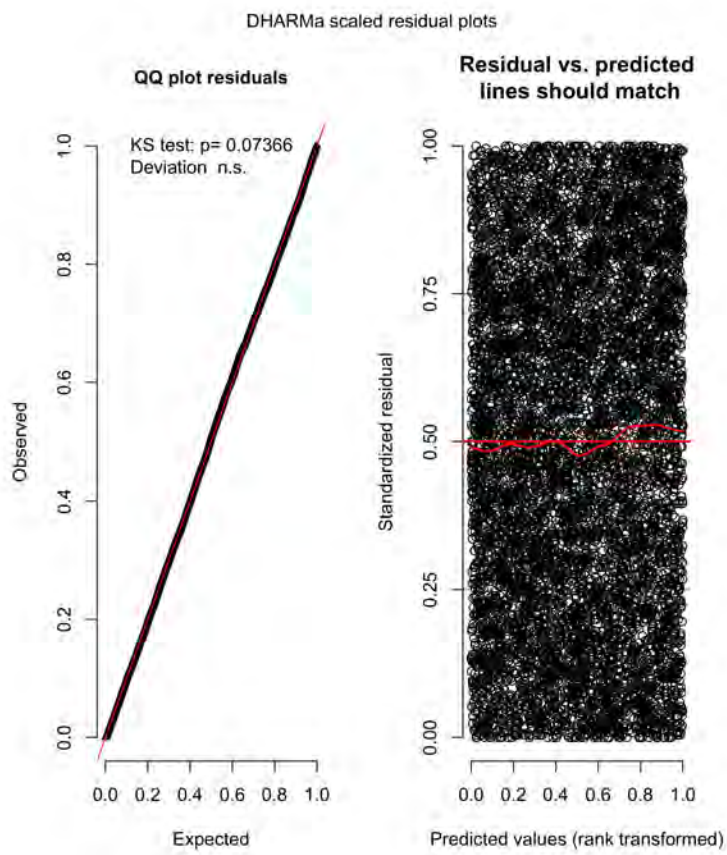




Figure A4: Details of the residuals of the fitted model M18. Plots have been created with the R package DHARMA (Hartig, 2018) to assess goodness-of-fit. On the left Q-Q plot showing no deviation of residuals from normality: Kolmogorov-Smirnov test (KS test) showed that deviation was non-significant (ns). On the right, nothing unusual in the plot showing residuals vs. predicted values.

Experiment I

Facultad de Filología UNED
Senda del Rey, 7
28040 Madrid

Laboratorio de Fonética  Antonio Quilis 

Yo, doy mi consentimiento para participar en el proyecto de investigación titulado *Estudio del énfasis del español*.

Reconozco que:
He sido informado del procedimiento del experimento y se me ha dado la oportunidad de discutir la información y mi participación con los investigadores.

Me han explicado los procedimientos requeridos para el proyecto y el tiempo involucrados y he recibido respuesta a cualquier pregunta que tenía.

Doy mi consentimiento para escuchar y responder a los sonidos o imágenes que oiga o vea en la pantalla del ordenador. También doy mi consentimiento para registrar mis respuestas.

Entiendo que mi participación es confidencial y que la información recabada durante este estudio podría ser publicada aunque no será revelada mi identidad.

Entiendo que puedo retirarme del estudio en cualquier momento y que esto no afectará a mi relación con el investigador ahora o en el futuro. En este caso, no recibiré la contraprestación económica de 10€.

Firmado:

Nombre:

Fecha:

Figure A5: Informed consent form presented to participants before taking part in the experiment.

Appendix B.

Experiment II

Experiment II

Sentences from the initial corpus that were used in Experiment II, as showed to the participants without punctuation marks.

Trial sentences:

- (1) Y me veía y la y la observaba y me cortaba o sea no me salía.
- (2) O sea a ver pongamos las cosas en su sitio vale perfecto que tú no tienes que reaccionar así.
- (3) En ese tipo de relaciones es la donde más fácil te puedes encontrar la mentira.
- (4) Yo me lo he currao igual que Rosa sabes y yo lo he pasao mal igual que Rosa.

Non-manipulated sentences:

- (1) Vale él es muy impulsivo y también eso también le sale del corazón.
- (2) Tú lo has dicho explícito yo no lo he dicho explícito.

Manipulated sentences:

- (1) Y lo primero que has ido es a saco a preguntarle a Verónica si le ha sentao mal.
- (2) Sencillamente me parece una persona absolutamente que nunca tiene una mala palabra para nadie.
- (3) Todos tenemos más afinidad con unos que con otros.
- (4) Y más si lo has hecho con conciencia de lo que estabas haciendo.

Filler sentences:

- (1) Porque confundimos la mano izquierda con la hipocresía y el ser hipócrita es una cosa y el ser diplomático es otra muy diferente.

Experiment II

- (2) Te lo te lo he dicho o no te lo he dicho al igual que le he dicho joder esto te ha salido muy bien te sale bonito.
- (3) Yo mira yo de hecho había apostado en primer lugar por los dos David por Alex o por Rosa tenía esos cuatro.

Experiment II

Age	Men				Women				Total
	Exp0	Exp1	Exp2	Exp3	Exp0	Exp1	Exp2	Exp3	
18-29	4	4	4	2	20	19	10	8	71
30-39	4	1	7	5	15	12	16	10	70
40-49	2	6	5	11	10	12	12	9	67
50-59	0	2	3	5	4	4	3	5	26
60	1	0	0	2	0	0	0	3	6
Total	11	13	19	25	49	47	41	35	240

(a)

Musical training	Condition				Total
	Exp0	Exp1	Exp2	Exp3	
None	34	33	39	36	142
Little	19	14	11	10	54
Much	7	13	10	14	44
Total	60	60	60	60	240

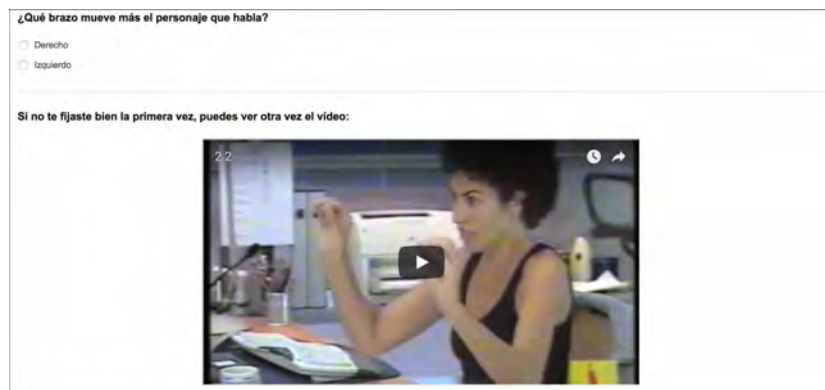
(b)

Table B1: *Details of participants per condition in Experiment II: (a) age range and (b) musical training.*

Experiment II



(d)



(e)

Figure B1: Sample screens from (a) to (e) showing the instructions given to participants.

Instructions:

(a) ‘This is a little perception experiment where you will watch some video-clips and then answer some easy questions. You will notice that we have taken extracts from a TV programme that you might know. For this reason, we hope the quality of image and sound will not be a problem for you to answer correctly the questions.’¹

(b) ‘In some cases, you will have to decide what words have a stronger em-

¹ This last instruction was intended to prepare the participants for the conducted stimuli manipulations.

Experiment II

phasis. Remember that some words stand out when they are uttered in the sentence. For example so: Press play to see someone say this sentence: “And I saw myself and I observed... her, and I lost my nerve, like, I couldn’t make it”.

(c) ‘Now you can watch the video again to select the words that, in your opinion, are uttered with a stronger emphasis’.

(d) ‘In some other cases, after listening to the sentence you will have to answer a question about what you have just seen, so you will have to pay attention. As in this example. Press play on the videoclip:’

(e) ‘What arm does the speaking person move more? If you didn’t notice the first time, you can watch the video again:’

Sentence	Words	Seconds	Speech rate
1	14	2.51	5.18
2	9	2.7	3.33
3	18	3.38	5.33
4	13	2.65	4.91

(a)

Sentence	Words rated as prominent	Total words	% Prominence
1	822	3360	24.46
2	761	2160	35.23
3	955	4320	22.10
4	664	3120	21.28

(b)

Table B2: *Details of target sentences used in Experiment II.*

Experiment II

Word	Class	Prominence marks	Total words	%
saco	Noun	205	240	85.4
nunca	Neg	197	240	82.0
unos	Pron	184	240	76.6
otros	Pron	178	240	74.1
conciencia	Noun	159	240	66.2
mal	Adv	130	240	54.1
más	Adv	257	480	53.5
todos	Pron	121	240	50.4
nadie	Neg	119	240	49.5
Verónica	Noun	115	240	47.9

(a)

Word class	Prominence marks	Total words	%
Neg	316	480	65.8
Noun	760	1440	52.7
Adv	593	1440	41.1
Num	78	240	32.5
Pron	565	2640	21.4
Verb	561	3120	21.2
Prep	192	1680	11.4
Conj	117	1440	8.1
Art	20	480	4.1

(b)

Table B3: *Details of precise words and word categories receiving the highest number of prominence marks.*

Experiment II

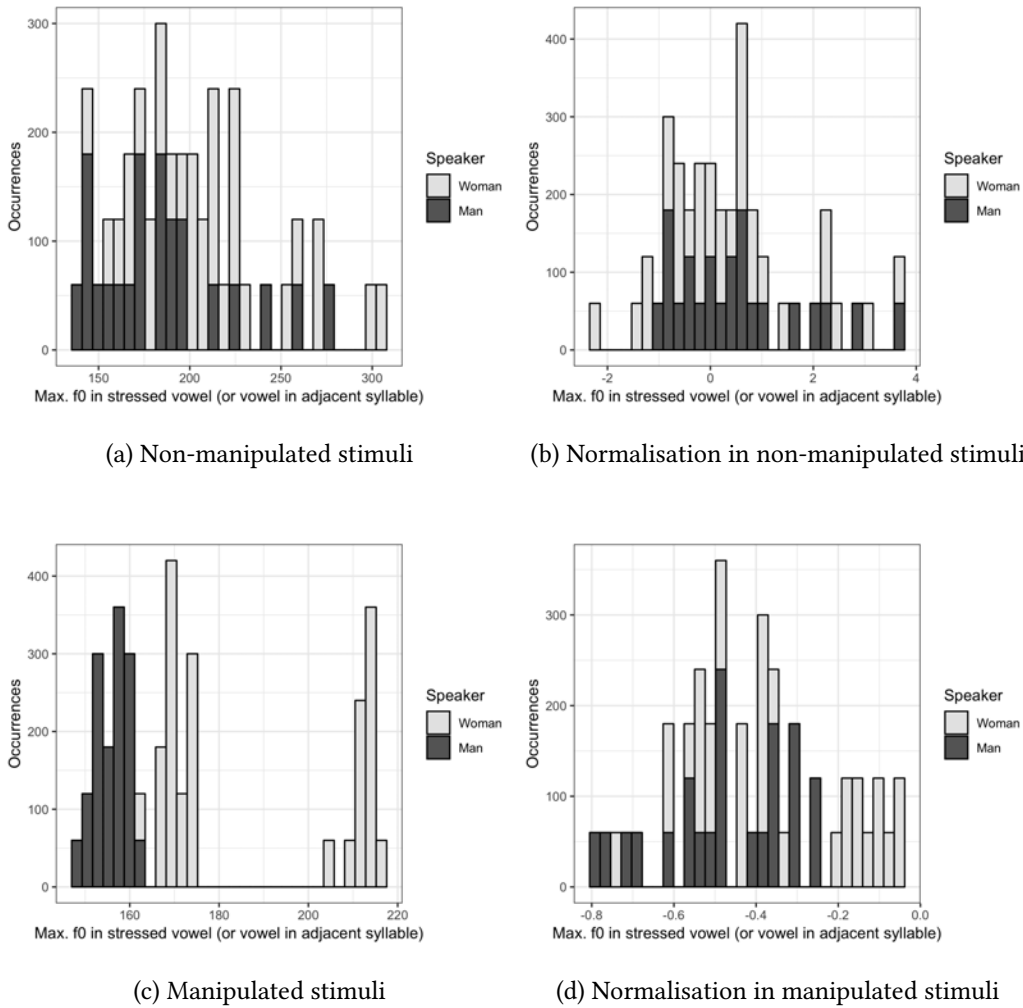


Figure B2: Histograms of f_0 measured as maximum frequency of stressed vowel (or vowel in adjacent syllable) for non-manipulated stimuli (a) and its normalization per speaker and sentence as z-scores (b). Distribution corresponding to neutralised f_0 is showed in (c) together with its normalization (d). Histograms (a) and (b) are representative of conditions Exp0 and Exp2. Histograms (c) and (d) correspond to conditions Exp1 and Exp3.

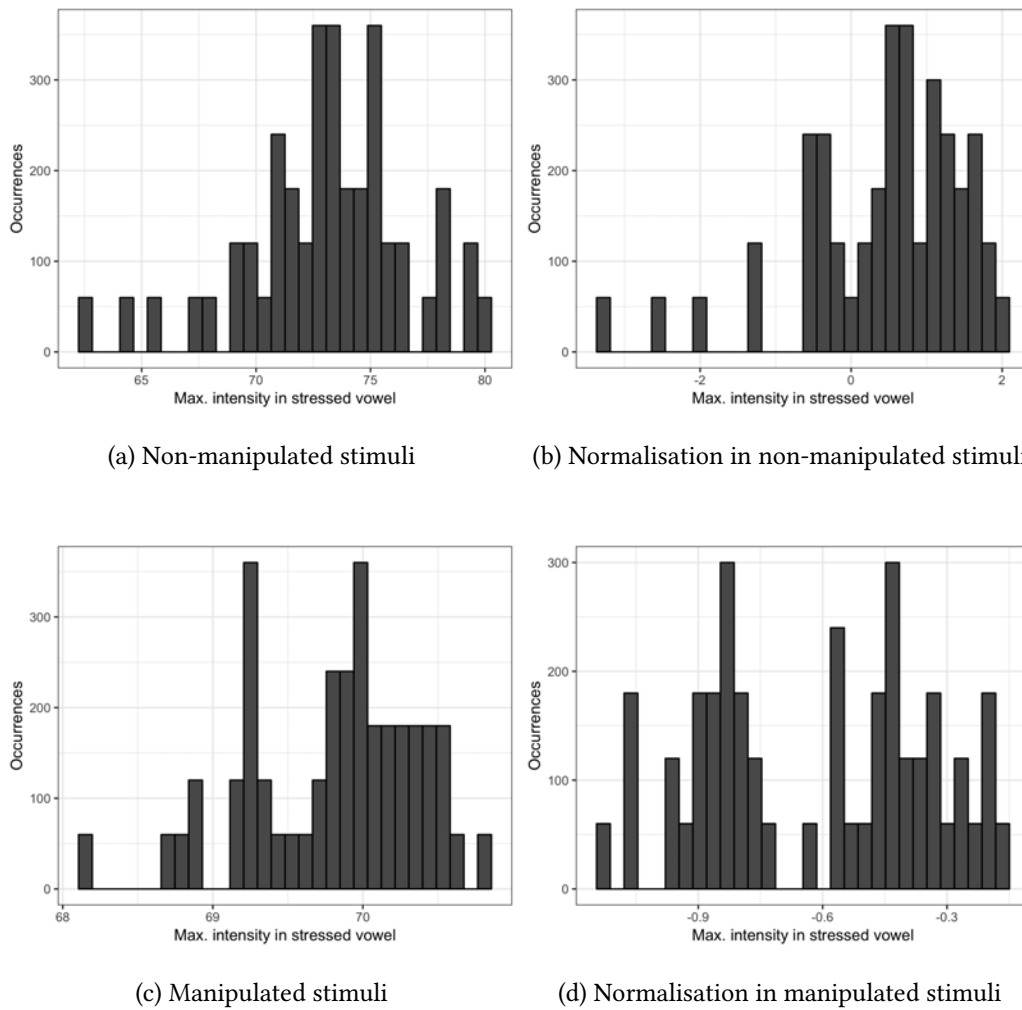
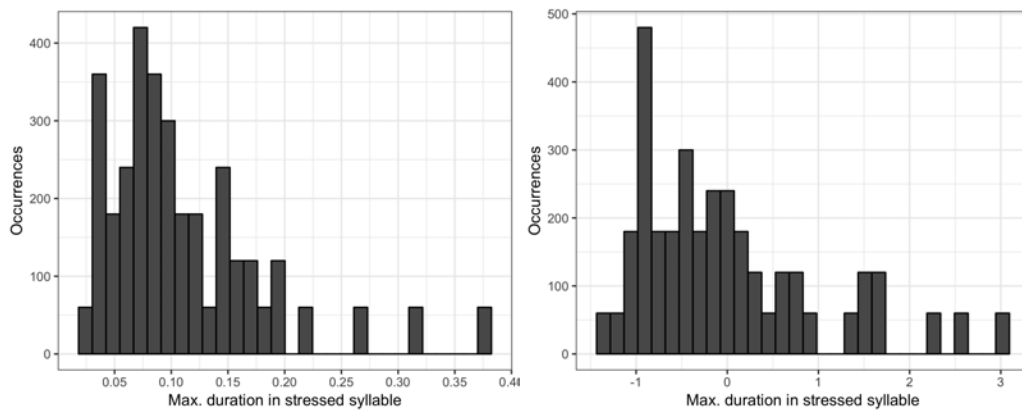


Figure B3: Histograms of intensity measured as maximum amplitude of stressed vowel for non-manipulated stimuli (a) and its normalization per sentence as z-scores (b). Distribution corresponding to neutralised intensity is showed in (c). Histograms (a) and (b) are representative of conditions Exp0 and Exp2. Histograms (c) and (d) corresponded to conditions Exp1 and Exp3.

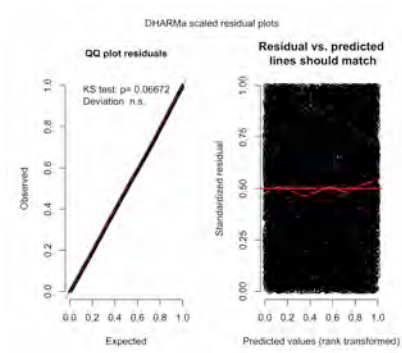


(a) Non-manipulated stimuli

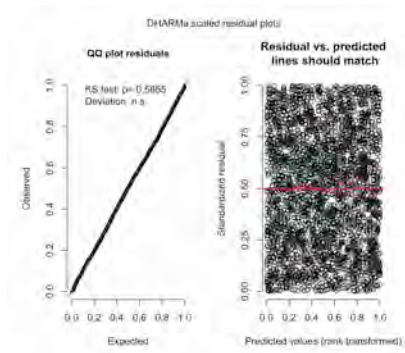
(b) Normalisation in non-manipulated stimuli

Figure B4: Histograms of duration measured as mean duration of the stressed syllable (a) and its normalization per sentence as z-scores (b). Both histograms (a) and (b) are representative of duration in all conditions.

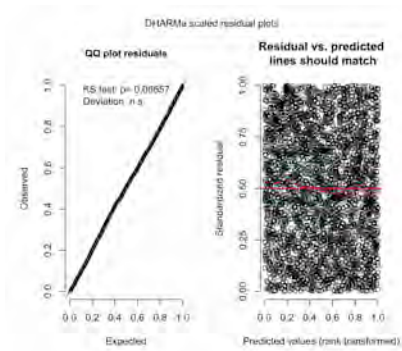
Experiment II



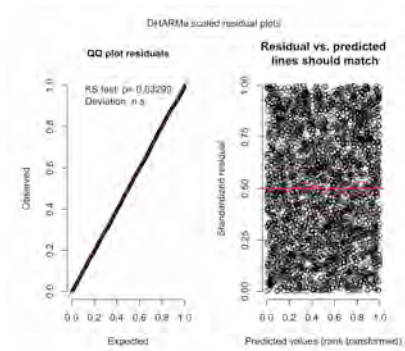
(a) Model G17



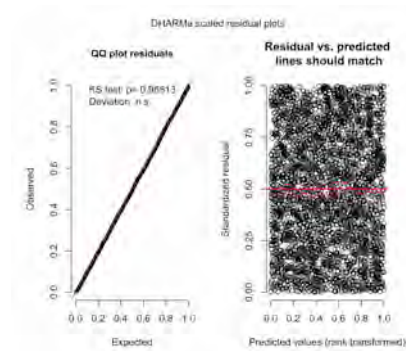
(b) Model g_{020}



(c) Model g_{113}



(d) Model g_{215}



(e) Model g_{318}

Figure B5: Details of the residuals of the minimal adequate models fitted in Experiment II. Plots have been created with the R package DHARMA (Hartig, 2018) to assess goodness-of-fit.

Bibliography

- Adamou, E., Gordon, M., & Gries, S. T. (2018). Prosodic and morphological focus marking in Ixcatec (Otomanguan). In E. Adamou, K. Huade, & M. Vanhove (Eds.) *Information Structure in Lesser-described Languages (Studies in Language Companion Series, 199)*, (pp. 51–84). Amsterdam: John Benjamins.
- Agelfors, E., Beskow, J., Dahlquist, M., Granström, B., Lundeberg, M., Spens, K., & Öhman, T. (1998). Synthetic faces as a lipreading support. In *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP98)*, (pp. 3047–3050). Sydney, Australia.
- Aitkin, M. (1996). A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and computing*, 6(3), 251–262.
- Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In B. N. Petrov, & F. Csaki (Eds.) *Proceedings of the 2nd International Symposium on Information Theory*, (pp. 267–281). Budapest: Akadémiai Kiadó.
- Al Moubayed, S., & Beskow, J. (2009). Effects of visual prominence cues on speech intelligibility. In *Proceedings of the International Conference on Auditory Visual Speech Processing (AVSP09)*, (pp. 43–46). Norwich, England.
- Al Moubayed, S., Beskow, J., & Granström, B. (2010). Auditory visual prominence. *Journal of Multimodal User Interfaces*, 3(4), 299–309.

Bibliography

- Al Moubayed, S., Beskow, J., Granström, B., & House, D. (2011). Audio-visual prosody: Perception, detection, and synthesis of prominence. In A. Esposito, A. M. Esposito, R. M. Martone, V. C. Müller, & G. Scarpetta (Eds.) *Towards Autonomous, Adaptive, and Context-Aware Multimodal Interfaces. Theoretical and Practical Issues*, (pp. 55–71). Berlin/Heidelberg/New York: Springer-Verlag.
- Alm, M., & Behne, D. (2015). Do gender differences in audio-visual benefit and visual influence in audio-visual speech perception emerge with age? *Frontiers in Psychology*, *6*:1014.
- Aloufy, S., Lapidot, M., & Myslobodsky, M. (1996). Differences in susceptibility to the “blending illusion” among native Hebrew and English speakers. *Brain and Language*, *53*(1), 51–57.
- Altenberg, B. (1987). *Prosodic patterns in spoken English: studies in the correlation between prosody and grammar for text-to-speech conversation*. Lund: Lund University Press.
- Ambrazaitis, G., & House, D. (2017). Multimodal prominences: Exploring the patterning and usage of focal pitch accents, head beats and eyebrow beats in Swedish television news readings. *Speech Communication*, *95*, 100–113.
- Amemiya, K., Karino, S., Ishizu, T., Yumoto, M., & Yamasoba, T. (2014). Distinct neural mechanisms of tonal processing between musicians and non-musicians. *Clinical Neurophysiology*, *125*(4), 738–747.
- Ananthakrishnan, S., & Narayana, S. S. (2008). Automatic prosodic event detection using acoustic, lexical, and syntactic evidence. *IEEE Transactions on Audio, Speech, and Language Processing*, *16*(1), 216–228.
- Armstrong, D. F., Stokoe, W. C., & Wilcox, S. E. (1995). *Gesture and the Nature of Language*. Cambridge: Cambridge University Press.
- Arnold, J. E. (2010). How speakers refer: The role of accessibility. *Language and Linguistic Compass*, *4*(4), 187–203.

Bibliography

- Arnold, T. W. (2010). Uninformative parameters and model selection using Akaike's Information Criterion. *Journal of Wildlife Management*, 74(6), 1175–1178.
- Aronson, E., Wilson, T., & Akert, R. (2010). *Social Psychology*. New Jersey: Pearson Education Inc.
- Arvaniti, A., Ladd, D. R., & Mennen, I. (1998). Stability of tonal alignment: the case of Greek prenuclear accents. *Journal of Phonetics*, 36(1), 3–25.
- Astésano, C., Magne, C., Morel, M., Coquillon, A., Espesser, R., Besson, M., & Lacheret-Dujour, A. (2004). Marquage acoustique du focus contrastif non codé syntaxiquement en français. In *Proceedings of the Journées d'Etude sur la Parole (JEP)*. Fez, Morocco.
- Aylett, M., & Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, 47(1), 31–56.
- Baayen, R. H., Davidson, D. J., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4), 390–412.
- Bagdasarian, S. A., & Vanyan, L. V. (2011). On the interrelation of rhythm and phrasal accent: A contrastive study. In *Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS11)*, (pp. 272–275). Hong Kong, China.
- Barbieri, F., Buonocore, A., Dalla Volta, R., & Gentilucci, M. (2009). How symbolic gestures and words interact with each other. *Brain and Language*, 110(1), 1–11.
- Barbosa, P. A., Eriksson, A., & Åkesson, J. (2013). Cross-linguistic similarities and differences of lexical stress realisation in Swedish and Brazilian Portuguese. In E. L. Asu, & P. Lippus (Eds.) *Nordic Prosody. Proceedings of the 11th conference, Tartu 2012*, (pp. 97–106). Frankfurt am Main: Peter Lang.

Bibliography

- Bard, E. G., Anderson, A. H., Sotillo, C., Aylett, M., Doherty-Sneddon, G., & Newlands, A. (2000). Controlling the intelligibility of referring expressions in dialogue. *Journal of Memory and Language*, 42(1), 1–22.
- Barkhuysen, P., Krahmer, E., & Swerts, M. (2005). Problem detection in human-machine interactions based on facial expressions of users. *Speech Communication*, 45(3), 343–359.
- Barkhuysen, P., Krahmer, E., & Swerts, M. (2008). The interplay between the auditory and visual modality for end-of-utterance detection. *Journal of the Acoustical Society of America*, 123(1), 354–365.
- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3), 255–278.
- Bates, D., Kliegl, R., Vasishth, S., & Baayen, R. H. (2015a). Parsimonious mixed models. *ArXiv e-prints*, ArXiv:1506.04967.
- Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015b). Fitting linear mixed-effects models using *lme4*. *Journal of Statistical Software*, 67(1), 1–48.
- Bateson, G. (1958). Language and Psychotherapy, Frieda Fromm-Reichmann's Last Project. *Psychiatry*, 21(1), 91–95.
- Baumann, S., Becker, J., & Mücke, D. (2007). Tonal and articulatory marking of focus in German. In *Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS07)*. Saarbrücken, Germany: University of Saarbrücken.
- Baumann, S., & Roth, A. (2014). Prominence and coreference – on the perceptual relevance of F0 movement, duration and intensity. In *Proceedings of the 7th International Conference on Speech Prosody (SP2014)*, (pp. 227–231). Dublin, Ireland.

Bibliography

- Baynes, K., Funnell, M. G., & Fowler, C. A. (1994). Hemispheric contributions to the integration of visual and auditory information in speech perception. *Perception and Psychophysics*, 55(6), 633–641.
- Beckman, M. E. (1986). *Stress and Non-Stress Accent*. Berlin/Boston: De Gruyter.
- Beckman, M. E., & Edwards, J. (1994). Articulatory evidence for differentiating stress categories. In P. A. Keating (Ed.) *Phonological Structure and Phonetic Form: Phonology and Phonetic Evidence*, (pp. 7–33). Cambridge: Cambridge University Press.
- Beckman, M. E., & Pierrehumbert, J. B. (1986). Intonational structure in Japanese and English. *Phonology Yearbook*, 3, 255–309.
- Bednarek, M. (2013). “There’s no harm, is there, in letting your emotions out”: a multimodal perspective on language, emotion and identity in MasterChef Australia. In N. Lorenzo-Dus, & P. Garcés-Conejos Blitvich (Eds.) *Real Talk: Reality Television and Discourse Analysis in Action*, (pp. 88–114). London: Palgrave Macmillan UK.
- Bell, A., Brenier, J., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, 60(1), 92–111.
- Bello, A. (1847/1860). *Gramática de la lengua castellana destinada al uso de los americanos*. Santiago de Chile: Imprenta del Progreso, 5^a ed.
- Bergmann, K., Aksu, V., & Kopp, S. (2011). The relation of speech and gestures: Temporal synchrony follows semantic synchrony. In *Proceedings of the 2nd Workshop on Gesture and Speech in Interaction (GESPIN - 2011)*. Bielefeld, Germany.
- Bergmann, K., & Kopp, S. (2006). Verbal or visual? How information is distributed across speech and gesture in spatial dialogue. In D. Schlangen, & R. Fernández (Eds.) *Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue*, (pp. 90–97). Potsdam: Universitätsverlag.

Bibliography

- Bernardis, P., & Gentilucci, M. (2006). Speech and gesture share the same communication system. *Neuropsychologia*, *44*(2), 178–190.
- Beskow, J. (1995). Rule-based visual speech synthesis. In *Proceedings of the 4th European Conference on Speech Communication and Technology (Eurospeech - 1995)*, (pp. 299–302). Madrid, Spain.
- Beskow, J., Granström, B., & House, D. (2006). Visual correlates to prominence in several expressive modes. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH - 2006*, (pp. 1272–1275). Pittsburgh, USA.
- Biau, E., Morís Fernández, L., Holle, H., Avila, C., & Soto-Faraco, S. (2016). Hand gestures as visual prosody: BOLD responses to audio–visual alignment are modulated by the communicative nature of the stimuli. *Neuroimage*, *132*, 129–137.
- Biau, E., & Soto-Faraco, S. (2013). Beat gestures modulate auditory integration in speech perception. *Brain and Language*, *124*(2), 143–152.
- Biau, E., Torralba, M., Fuentemilla, L., de Diego Balaguer, R., & Soto-Faraco, S. (2015). Speaker's hand gestures modulate speech perception through phase resetting of ongoing neural oscillations. *Cortex*, *68*, 76–85.
- Bickerton, D. (1990). *Language and Species*. Chicago: University Chicago Press.
- Birch, S., & Clifton Jr., C. (1995). Focus, accent, and argument structure: Effects on language comprehension. *Language and Speech*, *38*(4), 365–392.
- Birdwhistell, R. (1952). *Introduction to kinesics. An annotation system for analysis of body motion and gesture*. Louisville, KY: University of Louisville.
- Birdwhistell, R. (1970). *Kinesics and context*. Philadelphia: University of Pennsylvania Press.

Bibliography

- Bocci, G., & Avesani, C. (2011). Phrasal prominences do not need pitch movements: Postfocal phrasal heads in Italian. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH - 2011*, (pp. 1357–1360). Florence, Italy.
- Bock, J. K., & Mazzella, J. R. (1983). Intonational marking of given and new information: Some consequences for comprehension. *Memory and Cognition*, 11(1), 64–76.
- Boersma, P., & Weenink, D. (2018). Praat: Doing phonetics by computer. Computer program. Version 6.0.43, retrieved 8 September 2018 from <http://www.praat.org/>.
- Bolinger, D. L. (1955). Intersections of stress and intonation. *Word*, 11(2), 195–203.
- Bolinger, D. L. (1958). A theory of pitch accent in English. *Word*, 14(2-3), 109–149.
- Bolinger, D. L. (1961). Contrastive accent and contrastive stress. *Language*, 37, 83–96.
- Bolinger, D. L. (1972). Accent is predictable (if you're a mind-reader). *Language*, 48(3), 633–644.
- Bolinger, D. L., & Hodapp, M. (1961). Acento melódico. Acento de intensidad. *Boletín del Instituto de Filología de la Universidad de Chile*, 13, 33–48.
- Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J.-S. S. (2009). Generalized linear mixed models: A practical guide for ecology and evolution. *Trends in Ecology and Evolution*, 24(3), 127–135.
- Bonifaccio, G. (1616). *L'arte de' Cenni [The art of signs]*. Vicenza: Francesco Grossi.
- Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421), 9–25.
- Bressem, J. (2013). A linguistic perspective on the notation of form features in gestures. In C. Müller, A. Cienki, E. Fricke, S. H. Ladewig, D. McNeill, & S. Tessendorf (Eds.) *Body – Language – Communication. An International Handbook on Multimodality in*

Bibliography

- Human Interaction (Handbooks of Linguistics and Communication Science 38.1.)*, (pp. 1079–1098). Berlin/Boston: De Gruyter.
- Broughton, R. (1870). *Red as a rose is she: a novel*. London: Richard Bentley.
- Brown, G. (1983). Prosodic structure and the given/new distinction. In D. R. Ladd, & A. Cutler (Eds.) *Prosody: Models and Measurements*, (pp. 67–78). Berlin/Heidelberg/New York: Springer-Verlag.
- Bruce, G. (1977). *Swedish word accents in sentence perspective*. Lund: CWK Gleerup.
- Brugman, H., & Russel, A. (2004). Annotating multimedia/multi-modal resources with elan. In *Fourth International Conference on Language Resources and Evaluation (LREC - 2004)*. Nijmegen: Max Planck Institute for Psycholinguistics.
- Buhmann, J., Caspers, J., van Heuven, V. J., Hoekstra, H., Martens, J.-P., & Swerts, M. (2002). Annotation of prominent words, prosodic boundaries and segmental lengthening by non-expert transcribers in the Spoken Dutch Corpus. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC02)*, (pp. 779–785).
- Bull, P., & Connelly, G. (1985). Body movement and emphasis in speech. *Journal of Nonverbal Behavior*, 9(3), 169–187.
- Bulwer, J. (1644). *Chirologia or the Natural Language of the Hand, etc. [and] Chironomia, or, The art of manual Rhetoric, etc..* London: Thomas Harper.
- Burling, R. (1993). Primate calls, human language, and nonverbal communication. *Current Anthropology*, 34(1), 25–53.
- Burling, R. (2000). Comprehension, production and conventionalisation in the origins of language. In C. Knight, M. Studdert-Kennedy, & J. Hurford (Eds.) *The Evolutionary Emergence of Language*, (pp. 27–39). Cambridge: Cambridge University Press.

Bibliography

- Burnham, K. P., & Anderson, D. R. (2002). *Model Selection and Multimodel Inference: a Practice Information-Theoretic Approach*. Berlin/Heidelberg/New York: Springer-Verlag.
- Butterworth, B. (1989). Lexical access in speech production. In W. Marslen-Wilson (Ed.) *Lexical Representation and Process*, (pp. 108–135). Cambridge, MA: MIT Press.
- Butterworth, B., & Beattie, G. (1978). Gesture and silence as indicators of planning in speech. In R. N. Campbell, & P. T. Smith (Eds.) *Recent Advances in the Psychology of Language: Formal and Experimental Approaches. NATO Conference Series, vol 4b*, (pp. 347–360). Boston: Springer-Verlag.
- Cafaro, A., Vilhjálmsón, H. H., Bickmore, T., Heylen, D., Jóhannsdóttir, K. R., & Valgarðsson, G. S. (2012). First impressions: Users' judgments of virtual agents' personality and interpersonal attitude in first encounters. In Y. Nakano, M. Neff, A. Paiva, & M. Walker (Eds.) *Intelligent Virtual Agents. IVA 2012. Lecture Notes in Computer Science, vol. 7502*, (pp. 67–80). Berlin/Heidelberg/New York: Springer-Verlag.
- Campbell, N., & Beckman, M. E. (1997). Accent, stress, and spectral tilt. *Journal of the Acoustical Society of America*, *101*(5), 3195–3195.
- Canellada, M. J., & Madsen, J. K. (1987). *Pronunciación del español. Lengua hablada y literaria*. Madrid: Castalia.
- Carpenter, M., Nagell, K., & Tomasello, M. (1998). Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs of the Society for Research in Child Development*, *63*(4), i–vi, 1–174.
- Cartmill, E. A., Beilock, S., & Goldin-Meadow, S. (2012). A word in the hand: Action, gesture and mental representation in humans and non-human primates. *Philosophical Transactions of the Royal Society of London. Series B - Biological Sciences*, *367*(1585), 129–143.

Bibliography

- Castiglione, B. (1528). *Il libro del cortegiano* [*The book of the courtier*]. Firenze: Filippo di Giunta, eredi.
- Cavé, C., Guaitella, I., Bertrand, R., Santi, S., Harlay, F., & Espesser, R. (1996). About the relationship between eyebrow movements and f_0 variations. In *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP96)*, vol. 2175–2179. Philadelphia, USA.
- Cerrato, L. (2007). *Investigating Communicative Feedback Phenomena across Languages and Modalities*. Ph.D. thesis, KTH Computer Science and Communication.
- Chiarcos, C., Barry, C., & Grabski, M. (Eds.) (2011). *Saliency: Multidisciplinary perspectives on its function in discourse*. Berlin/New York: de Gruyter.
- Chomsky, N., & Halle, M. (1968). *The sound pattern of English*. New York: Harper & Row.
- Christenfeld, N., Schachter, S., & Bilous, F. (1991). Filled pauses and gestures: It's not coincidence. *Journal of Psycholinguistic Research*, 20(1), 1–10.
- Cienki, A. (2005). Image schemas and gesture. In B. Hampe (Ed.) *From Perception to Meaning: Image Schemas in Cognitive Linguistics*, (pp. 421–442). Berlin: De Gruyter.
- Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, 12(4), 335–359.
- Cocks, N., Dipper, L., Pritchard, M., & Morgan, G. (2013). The impact of impaired semantic knowledge on spontaneous iconic gesture production. *Aphasiology*, 27(9), 1050–1069.
- Cohen, A., Collier, R., & 't Hart, J. (1982). Declination: Construct or intrinsic feature of speech pitch? *Phonetica*, 39(4-5), 254–273.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46.

Bibliography

- Cole, J. (2015). Prosody in context: a review. *Language, Cognition and Neuroscience*, 30(1-2), 1–31.
- Cole, J., Mahrt, T., & Hualde, J. I. (2014). Listening for sound, listening for meaning: Task effects on prosodic transcription. In *Proceedings of the 7th International Conference on Speech Prosody (SP2014)*, (pp. 859–863). Dublin, Ireland.
- Cole, J., Mo, Y., & Hasegawa-Johnson, M. (2010). Signal-based and expectation-based factors in the perception of prosodic prominence. *Laboratory Phonology*, 1(2), 425–452.
- Colin, C., Radeau, M., Soquet, A., Demolin, D., Colin, F., & Deltenre, P. (2002). Mismatch negativity evoked by the McGurk-MacDonald effect: A phonetic representation within short-term memory. *Clinical Neurophysiology*, 113(4), 495–506.
- Condillac, E. B. d. (1746). *Essai sur l'origine des connaissances humaines [Essay on the origin of human knowledge]*. Amsterdam: Pierre Mortier.
- Condon, W. S. (1976). An analysis of behavioral organization. *Sign Language Studies*, 13, 285–318.
- Condon, W. S., & Ogston, W. D. (1966). Sound film analysis of normal and pathological behavior patterns. *Journal of Nervous and Mental Disease*, 143(4), 338–347.
- Condon, W. S., & Ogston, W. D. (1967). A segmentation of behavior. *Journal of Psychiatric Research*, 5(3), 221–235.
- Coney, J. (2002). Lateral asymmetry in phonological processing: Relating behavioral measures to neuroimaged structures. *Brain and Language*, 80(3), 355–365.
- Contreras, H. (1963). Sobre el acento español. *Boletín del Instituto de Filología de la Universidad de Chile*, 15, 223–237.
- Contreras, H. (1964). ¿tiene el español un acento de intensidad? *Boletín del Instituto de Filología de la Universidad de Chile*, 16, 237–239.

Bibliography

- Corballis, M. C. (2002). *From Hand to Mouth: The Origins of Language*. Princeton/Oxford: Princeton University Press.
- Crichtley, M. (1939). *The Language of Gesture*. London: Edward Arnold.
- Crivelli, C., Jarillo, S., Russell, J. A., & Fernández-Dols, J. M. (2016). Reading emotions from faces in two indigenous societies. *Journal of Experimental Psychology: General*, 145(7), 830–843.
- Cuervo, R. J. (1874). *Gramática de la lengua castellana destinada al uso de los americanos*. Ed. con notas de R. J. Cuervo. Bogotá: Hermanos Echeverría.
- Cummins, F., & Port, R. (1998). Rhythmic constraints on stress timing in English. *Journal of Phonetics*, 26(2), 145–171.
- Cutler, A. (1976). Phoneme-monitoring reaction time as a function of preceding intonation contour. *Perception and Psychophysics*, 20(1), 55–60.
- Cutler, A., & Darwin, C. J. (1981). Phoneme-monitoring reaction time and preceding prosody: Effects of stop closure duration and of fundamental frequency. *Perception & Psychophysics*, 29(3), 217–224.
- Cutler, A., & Foss, D. J. (1977). On the role of sentence stress in sentence processing. *Language and Speech*, 20(1), 1–10.
- Dahan, D., Tanenhaus, M. K., & Chambers, C. G. (2002). Accent and reference resolution in spoken-language comprehension. *Journal of Memory and Language*, 47(2), 292–314.
- Dancer, J., Krain, M., Thompson, C., Davis, P., & Glen, J. (1994). A cross-sectional investigation of speechreading in adults: Effects of age, gender, practice, and education. *Volta Review*, 96, 31–40.
- Danesi, M. (1993). *Vico, Metaphor, and the Origin of Language*. Bloomington: Indiana University Press.

Bibliography

- Darwin, C. R. (1871). *The Descent of Man and Selection in Relation to Sex*. London: John Murray.
- Darwin, C. R. (1872). *The Expression of the Emotions in Man and Animals*. London: John Murray.
- de Gelder, B., & Vroomen, J. (2000). The perception of emotions by ear and by eye. *Cognition and Emotion*, 14(3), 289–311.
- de Jorio, A. (1832). *La mimica degli antichi investigata nel gestire napoletano [Gestural expression of the ancients in the light of Neapolitan gesturing]*. Naples: Fibreno.
- de la Mota, C. (1997). Prosody of sentences with contrastive new information in Spanish. In A. Botinis, G. Kouroupetroglou, & G. Carayiannis (Eds.) *Intonation: theory, models and applications*, (pp. 75–78). Athens, Greece.
- De Ruiter, J. (2000). The production of gesture and speech. In D. McNeill (Ed.) *Language and Gesture*, (pp. 248–311). Cambridge: Cambridge University Press.
- De Ruiter, J. P. (1998). *Gesture and speech production*. Ph.D. thesis, Katholieke Universiteit Nijmegen.
- Delattre, P. (1966). A comparison of syllable length conditioning among languages. *IRAL - International Review of Applied Linguistics in Language Teaching*, 4(1-4), 183–198.
- Delattre, P. (1969). An acoustic and articulatory study of vowel reduction in four languages. *IRAL - International Review of Applied Linguistics in Language Teaching*, 7(4), 295–325.
- Delgado-Martins, M. R. (1973). Análise acústica das vogais orais tônicas em português [acoustic analysis of stressed vowels in Portuguese]. *Boletim de Filologia (University of Lisbon)*, 2, 303–3014.
- Della Casa, G. (1558). *Il Galateo, overo de' costumi [Galateo: the rules of polite behavior]*. Venice: Nicolo Bevilacqua.

Bibliography

- Diderot, D. (1751). *Lettre sur les sourds et muets à l'usage de ceux qui entendent et qui parlent* [Letter on the deaf and dumb]. s. l.: s. n.
- Dittmann, A. T., & Llewelyn, L. G. (1969). Body movement and speech rhythm in social conversation. *Journal of Personality and Social Psychology*, 23(2), 283–292.
- Dogil, G. (1999). The phonetic manifestation of word stress in Lithuanian, Polish, German and Spanish. In H. van der Hulst (Ed.) *Word prosodic systems in the languages of Europe*, (pp. 273–311). Berlin: De Gruyter.
- Dogil, G., & Williams, B. (1999). The phonetic manifestation of word stress. In H. van der Hulst (Ed.) *Word prosodic systems in the languages of Europe*, (pp. 273–311). Berlin: De Gruyter.
- Dohen, M., & Løevenbruck, H. (2005). Audiovisual production and perception of contrastive focus in French: A multispeaker study. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH - 2005*, (pp. 2413–2416). Lisbon, Portugal.
- Dohen, M., & Løevenbruck, H. (2009). Interaction of audition and vision for the perception of prosodic contrastive focus. *Language and Speech*, 52(2-3), 177–206.
- Donati, C., & Nespør, N. (2003). From focus to syntax. *Lingua*, 113(11), 1119–1142.
- D'Imperio, M. (1998). Acoustic-perceptual correlates of sentence prominence in Italian questions and statements. *Journal of the Acoustical Society of America*, 104(3), 1779–1779.
- Eady, S., Cooper, W. E., Klouda, G., Mueller, P., & Lotts, D. (1986). Acoustical characteristics of sentential focus: Narrow vs. broad and single vs. dual focus environments. *Language and Speech*, 29(3), 233–251.
- Eberhardt, M., & Downs, C. (2015). “(r) you saying yes to the dress?”: Rhoticity on a bridal reality television show. *Journal of English Linguistics*, 43(2), 118–142.

Bibliography

- Efron, D. (1941/1972). *Gesture, Race, and Culture*. New York: King's Crown Press.
- Ekman, P. (1999). Emotional and conversational nonverbal signals. In L. S. Messing, & R. Campbell (Eds.) *Gesture, Speech, and Sign*, (pp. 45–55). New York: Oxford University Press.
- Ekman, P., & Friesen, W. V. (1969). The Repertoire of Nonverbal Behavior: Categories, Origins, Usage, and Coding. *Semiotica*, 1(1), 49–98.
- Enríquez, E., Casado, C., & Santos, A. (1989). La percepción del acento en español. *Lingüística Española Actual (LEA)*, 11, 241–269.
- Épée, C.-M. A. d. l. (1776). *Institution des sourds et muets, par la voie des signes méthodiques; ouvrage qui contient le projet d'une langue universelle, par l'entremise des signes naturels assujettis á une méthode [The instruction of the deaf and dumb by means of methodical signs]*. Paris: Nyon.
- Eriksson, A., Bertinetto, P. M., Heldner, M., Nodari, R., & Lenoci, G. (2016). The acoustics of word stress in Italian as a function of stress level and speaking style. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH - 2016*, (pp. 1059–1063). San Francisco, USA.
- Eriksson, A., & Heldner, M. (2015). The acoustics of word stress in English as a function of stress level and speaking style. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH - 2015*, (pp. 41–45). Dresden, Germany.
- Eriksson, A., Thunberg, G. C., & Traunmüller, H. (2001). Syllable prominence: A matter of vocal effort, phonetic distinctness and top-down processing. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH - 2001*, (pp. 399–402). Aalborg, Denmark.
- Escandell-Vidal, V. (2011). Verum focus y prosodia: cuando la duración (sí que) importa. *Oralia*, 14, 181–201.

Bibliography

- Escandell-Vidal, V., & Leonetti, M. (2014). Fronting and irony in Spanish. In A. Dufter, & A. S. Octavio de Toledo y Huerta (Eds.) *Left Sentence Peripheries in Spanish: Diachronic, Variationist and Comparative Perspectives (Linguistics Today, 214)*, (pp. 309–342). Amsterdam: John Benjamins.
- Escandell-Vidal, V., Marrero Aguiar, V., & Pérez Ocón, P. (2014). Prosody, information structure and Evaluation. In G. Thompson, & L. Alba-Juez (Eds.) *Evaluation in Context (Pragmatics and Beyond New Series, 242)*, (pp. 153–178). Amsterdam: John Benjamins.
- Estebas-Vilaplana, E., & Prieto, P. (2010). Castilian Spanish intonation. In P. Prieto, & P. Roseano (Eds.) *Transcription of intonation of the Spanish Language*, (pp. 17–48). Munich: Lincom Europa.
- Esteve-Gibert, N., & Prieto, P. (2013). Prosodic structure shapes the temporal realization of intonation and manual gesture movements. *Journal of Speech, Language, and Hearing Research, 56*(3), 850–864.
- Face, T. (2001). Focus and early peak alignment in Spanish intonation. *Probus, 13*(2), 223–246.
- Face, T. (2003). Intonation in Spanish declaratives: Differences between lab speech and spontaneous speech. *Catalan Journal of Linguistics, 2*, 115–131.
- Face, T. L. (2000). The role of syllable weight in the perception of Spanish stress. In H. Campos, E. Herburger, A. Morales-Front, & T. J. Walsh (Eds.) *Hispanic Linguistics at the Turn of the Millennium. Papers from the 3rd Hispanic Linguistics Symposium*, (pp. 1–13). Somerville: Cascadilla Press.
- Falk, S. (2014). On the notion of salience in spoken discourse - prominence cues shaping discourse structure and comprehension. *TIPA. Travaux interdisciplinaires sur la parole et le langage, 30*. [mis en ligne le 18 décembre 2014]. URL: <http://journals.openedition.org/tipa/1303>.

Bibliography

- Fant, G., Kruckenberg, A., & Nord, L. (1991). Durational correlates of stress in Swedish, French and English. *Journal of Phonetics*, 19(3-4), 351–365.
- Ferré, G. (2010). Timing relationships between speech and co-verbal gestures in spontaneous French. In *Language Resources and Evaluation, Workshop on Multimodal Corpora, May 2010*, vol. W6, (pp. 86–91). Malta.
- Ferreira, L. (2008). *High initial tones and plateaux in Spanish and Portuguese neutral declaratives: Consequences to the relevance of f_0 , duration and vowel quality as stress correlates*. Ph.D. thesis, University of Illinois at Urbana-Champaign.
- Féry, C., & Krifka, M. (2008). Unity and Diversity of Languages. In P. van Sterkenburg (Ed.) *Information Structure: Notional Distinctions, Ways of Expression*, (pp. 123–136). Amsterdam: John Benjamins.
- Féry, C., & Kügler, F. (2008). Pitch accent scaling on given, new and focused constituents in German. *Journal of Phonetics*, 36(4), 680–703.
- Feyereisen, P. (1987). Gestures and speech, interactions and separations: A reply to McNeill (1985). *Psychological Review*, 94(4), 493–498.
- Field, A., Miles, J., & Field, Z. (2012). *Discovering Statistics using R*. London: SAGE.
- Figueras, C., & Santiago, M. (1993a). Investigaciones sobre la naturaleza del acento a través del Visi-pitch. *Estudios de fonética experimental*, 5, 21–45.
- Figueras, C., & Santiago, M. (1993b). Producción del rasgo acentual mediante síntesis de voz. *Estudios de fonética experimental*, 5, 113–128.
- Firbas, J. (1964). On defining the theme in functional sentence perspective. *Travaux linguistiques de Prague*, 1, 267–280.
- Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A - Mathematical, Physical and Engineering Sciences*, 222(594-604), 309–368.

Bibliography

- Flecha-García, M. L. (2006). *Eye-brow raising in dialogue: Discourse structure, utterance function, and pitch accents*. Ph.D. thesis, University of Edinburgh.
- Flecha-García, M. L. (2007). Non-verbal communication in dialogue: Alignment between eyebrow raises and pitch accents in English. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 29, (p. 1753). Austin, USA.
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 75(5), 378–382.
- Fónagy, I. (1958). Elektrophysiologische beiträge zur Akzentfrage. *Phonetica*, 2, 12–58.
- Fosler-Lussier, E., & Morgan, N. (1999). Effects of speaking rate and word frequency on pronunciations in conventional speech. *Speech Communication*, 29(2-4), 137–158.
- Fournier, D. A., Skaug, H. J., Ancheta, J., Ianelli, J., Magnusson, A., Maunder, M. N., Nielsen, A., & Sibert, J. (2012). AD Model Builder: Using automatic differentiation for statistical inference of highly parameterized complex nonlinear models. *Optimization Methods and Software*, 27(2), 233–249. Computer program. URL: <http://www.admb-project.org/>.
- Fowler, C. A., & Housum, J. (1987). Talkers' signaling of "new" and "old" words in speech and listeners' perception and use of the distinction. *Journal of Memory and Language*, 26(5), 489–504.
- Foxton, J. M., Riviere, L.-D., & Barone, P. (2010). Cross-modal facilitation in speech prosody. *Cognition*, 115(1), 71–78.
- Freedman, N., & Hoffman, S. P. (1967). Kinetic behavior in altered clinical states: Approach to objective analysis of motor behavior during clinical interviews. *Perceptual and Motor skills*, 24, 527–539.
- Frota, S. (2002). Nuclear falls and rises in European Portuguese: A phonological analysis of declarative and question intonation. *Probus*, 14(1), 113–146.

Bibliography

- Fry, D. B. (1955). Duration and intensity as physical correlates of linguistic stress. *Journal of the Acoustical Society of America*, 27(4), 765–768.
- Fry, D. B. (1958). Experiments in the perception of stress. *Language and Speech*, 1(2), 126–152.
- Fry, D. B. (1965). The dependence of stress judgments on vowel formant structure. In *Proceedings of the 5th International Congress Phonetic Sciences (ICPhS65)*, (pp. 306–311). Münster.
- Fudge, E. C. (1969). Syllables. *Journal of Linguistics*, 5(2), 253–286.
- Gahl, S., & Garnsey, S. M. (2004). Knowledge of grammar, knowledge of usage: Syntactic probabilities affect pronunciation variation. *Language*, 80(4), 748–775.
- Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain*, 119(2), 593–609.
- Gallinares, J. (1944). Nuevos concepto de la acentuación española. *Boletín de Filología de la Universidad de Montevideo*, 4, 33–48.
- Garrido, J. M., Llisterri, J., de la Mota, C., & Ríos, A. (1993). Prosodic differences in reading style: Isolated vs. contextualized sentences. In *Proceedings of the 3rd European Conference on Speech Communication and Technology (Eurospeech - 1993)*, (pp. 573–575). Berlin, Germany.
- Garrido, J. M., Llisterri, J., de la Mota, C., & Ríos, A. (1995). Estudio comparado de las características prosódicas de la oración simple en español en dos modalidades de lectura. In A. Elejabeitia, & A. Iribar (Eds.) *Phonetica. Trabajos de fonética experimental*, (pp. 177–193). Bilbao: Laboratorio de fonética, Universidad de Deusto.
- Gentilucci, M., & Corballis, M. C. (2006). From manual gesture to speech: A gradual transition. *Neuroscience & Biobehavioral Reviews*, 30(7), 949–960.

Bibliography

- Gentilucci, M., Dalla Volta, R., & Gianelli, C. (2008). When the hands speak. *Journal of Physiology*, 102, 21–30.
- Genzel, S., Ishihara, S., & Surányi, B. (2015). The prosodic expression of focus, contrast and givenness: A production study of Hungarian. *Lingua*, 165(B), 183–204.
- Gérando, J.-M. B. d. (1800). *Des signes et de l'art de penser considérés dans leurs rapports mutuels [On signs and on the art of thinking considered in their relations to each other]*. Paris: Goujon fils, Fuchs, Henrichs.
- Gibbon, D. (1998). Intonation in German. In D. Hirst, & A. di Cristo (Eds.) *Intonation systems: A survey of twenty languages*, (pp. 78–95). Cambridge: Cambridge University Press.
- Gili-Fivela, B. (2006). Scaling e allineamento dei bersagli tonali: L'identificazione di due accenti discendenti [the coding of pitch target alignment and scaling: Identification of two falling pitch accents]. In S. Atti del Convegno Nazionale AISV (Associazione Italiana di Scienze della Voce) (Ed.) *Analisi prosodica: teorie, modelli e sistemi di annotazione*, (pp. 214–232). Torriana, RN: EDK.
- Goodall, P. (2014). *The History of Strategic Thought*. Xlibris UK.
- Gould, S. J. (1980). *The Panda's Thumb*. New York/London: W. W. Norton & Co.
- Grabe, E., Post, B., Nolan, F., & Farrar, K. (2000). Pitch accent realization in four varieties of British English. *Journal of Phonetics*, 28(2), 161–185.
- Granström, B., House, D., & Lundeberg, M. (1999). Prosodic cues in multimodal speech perception. In *Proceedings of the 14th International Congress of Phonetic Sciences (ICPhS99)*, vol. 1, (pp. 655–658). San Francisco, USA.
- Gregory, M. L., & Altun, Y. (2004). Using conditional random fields to predict pitch accents in conversational speech. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics, ACL - 2004*, vol. P04-1086. Barcelona, Spain.

Bibliography

- Gregory, M. L., Raymond, W. D., Bell, A., Fosler-Lussier, E., & Jurafsky, D. (1999). The effects of collocational strength and contextual predictability in lexical production. In *Proceedings of the Chicago Linguistic Society*, vol. 35, (pp. 151–166). Chicago, USA.
- Gries, S. T. (2015). The most under-used statistical method in corpus linguistics: Multi-level (and mixed-effects) models. *Corpora*, 10(1), 95–125.
- Grosz, B. J., Joshi, A. K., & Weinstein, S. (1995). Centering: A framework for modeling the local discourse. *Computational Linguistics*, 21(2), 203–225.
- Grosz, B. J., & Sidner, C. L. (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3), 175–204.
- Grueber, C. E., Nakagawa, S., Laws, R. J., & Jamieson, I. G. (2011). Multimodel inference in ecology and evolution: Challenges and solutions. *Journal of Evolutionary Biology*, 24(4), 699–711.
- Gueorguieva, R., & Krystal, J. H. (2004). Move over ANOVA: Progress in analyzing repeated-measures data and its reflection in papers published in the Archives of General Psychiatry. *Archives of General Psychiatry*, 61(3), 310–317.
- Gundel, J. K. (1974). *The role of topic and comment in linguistic theory*. Ph.D. thesis, University of Texas.
- Gundel, J. K., & Fretheim, T. (2004). Topic and focus. In L. R. Horn, & G. Ward (Eds.) *The Handbook of Pragmatics*, (pp. 175–196). Malden, MA: Blackwell Publishers.
- Gussenhoven, C. (1983a). Focus, mode, and the nucleus. *Journal of Linguistics*, 19(2), 377–417.
- Gussenhoven, C. (1983b). *A semantic analysis of the nuclear tones of English*. Bloomington, IN: Indiana University Linguistics Club.
- Gussenhoven, C. (1983c). Testing the reality of focus domains. *Language and Speech*, 26(1), 61–80.

Bibliography

- Gussenhoven, C. (1984). *On the Grammar and Semantics of Sentence Accents*. Cinnaminson, NJ: Foris.
- Gussenhoven, C. (1992). Sentence accents and argument structure. In I. M. Roca (Ed.) *Thematic Structure: Its Role in Grammar*, (pp. 79–106). Berlin/New York: Foris.
- Gussenhoven, C. (2004). *The Phonology of Tone and Intonation*. Research Surveys in Linguistics. Cambridge: Cambridge University Press.
- Gussenhoven, C. (2005). Transcription of Dutch Intonation. In S.-A. Jun (Ed.) *Prosodic Typology: The Phonology of Intonation and Phrasing*, (pp. 118–145). Oxford: Oxford University Press.
- Gussenhoven, C., Repp, B., Rietveld, A., Rump, H., & Terken, J. (1997). The perceptual prominence of fundamental frequency peaks. *Journal of the Acoustical Society of America*, 102(5), 3009–22.
- Gussenhoven, C., & Rietveld, A. (1988). Fundamental frequency declination in Dutch: Testing three hypotheses. *Journal of Phonetics*, 16, 355–369.
- Gussenhoven, C., & Rietveld, A. (1998). On the speaker-dependence of the perceived prominence of F0 peaks. *Journal of Phonetics*, 26(4), 371–380.
- Hadar, U., Steiner, T. J., Grant, E. C., & Rose, F. C. (1983). Head movement correlates of juncture and stress at sentence level. *Language and Speech*, 26(2), 117–129.
- Hadar, U., Steiner, T. J., & Rose, C. F. (1985). Head movement during listening turns in conversation. *Journal of Nonverbal Behavior*, 9(4), 214–228.
- Halle, M., & Vergnaud, J.-R. (1987). Stress and the cycle. *Linguistics Inquiry*, 18(1), 45–84.
- Halliday, M. A. K. (1967a). *Intonation and Grammar in British English*. The Hague/Paris: Mouton.

Bibliography

- Halliday, M. A. K. (1967b). Notes on transitivity and theme in English, part II. *Journal of Linguistics*, 3(2), 199–244.
- Hartig, F. (2018). Dharma: Residual diagnostics for hierarchical (multi-level / mixed) regression models. R package version 0.2.0, URL: <http://florianhartig.github.io/DHARMA/>.
- Hedberg, N., & Fadden, L. (2007). The information structure of *it*-clefts, *wh*-clefts and reverse *wh*-clefts in English. In N. Hedberg, & R. Zacharski (Eds.) *The Grammar-Pragmatics Interface. Essays in Honor of Jeanette K. Gundel (Pragmatics & Beyond, 155)*, (pp. 19–48). Amsterdam: John Benjamins.
- Heldner, M. (2001). Spectral emphasis as a perceptual cue to prominence. *TMH-QPSR*, 42(1), 51–57.
- Heldner, M. (2003). On the reliability of overall intensity and spectral emphasis as acoustic correlates of focal accents in Swedish. *Journal of Phonetics*, 31(1), 39–62.
- Heldner, M., & Strangert, E. (1997). To what extent is perceived focus determined by f_0 -cues? In *Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech - 1997)*, (pp. 875–877). Rhodes, Greece.
- Hellmuth, S. (2007). The relationship between prosodic structure and pitch accent distribution: Evidence from Egyptian Arabic. *The Linguistic Review*, 24(2), 289–214.
- Heylen, D. (2008). Listening Heads. In I. Wachsmuth, & G. Knoblich (Eds.) *Modeling Communication with Robots and Virtual Humans*, (pp. 241–259). Berlin/Heidelberg/New York: Springer-Verlag.
- Hirschberg, J. (1993). Pitch accent in context predicting intonational prominence from text. *Artificial Intelligence*, 63(1-2), 305–340.
- Hockett, C. F. (1958). *A Course in Modern Linguistics*. New York: Macmillan.

Bibliography

- Hoetjes, M., Krahmer, E., & Swerts, M. (2014). Does our speech change when we cannot gesture? *Speech Communication, 57*, 257–267.
- Hostetter, A. B., & Alibali, M. W. (2007). Raise your hand if you're spatial: Relations between verbal and spatial skills and gesture production. *Gesture, 7*(1), 73–95.
- House, D., Beskow, J., & Granström, B. (2001). Timing and interaction of visual cues for prominence in audiovisual speech perception. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH - 2001*, (pp. 387–390). Aalborg, Denmark.
- Hualde, J. I. (2002). Intonation in Romance: Introduction to the special issue. *Probus, 14*(1), 1–7.
- Hurvich, C., & Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika, 76*(2), 297–293.
- Huss, V. (1975). Neutralisierung englischer Akzentunterschiede in der Nachkontur [Neutralisation of English stress contrasts in post-nuclear position]. *Phonetica, 32*, 278–291.
- Huss, V. (1978). English word stress in the post-nuclear position. *Phonetica, 35*(2), 86–105.
- Hutka, S., Bidelman, G. M., & Moreno, S. (2015). Pitch expertise is not created equal: Cross-domain effects of musicianship and tone language experience on neural and behavioural discrimination of speech and music. *Neuropsychologia, 71*, 52–63.
- Igualada, A., Bosch, L., & Prieto, P. (2015). Language development at 18 months is related to multimodal communicative strategies at 12 months. *Infant Behavior and Development, 39*, 42–52.
- Irwin, J. R., Whalen, D. H., & Fowler, C. A. (2006). A sex difference in visual influence on heard speech. *Perception and Psychophysics, 68*(4), 582–592.

Bibliography

- Isenberg, D., & Gay, T. (1978). Acoustic correlates of perceived stress in an isolated synthetic disyllable. *Journal of the Acoustical Society of America*, 64(S21).
- Ishi, C. T., Ishiguro, H., & Hagita, N. (2014). Analysis of relationship between head motion events and speech in dialogue conversation. *Speech Communication*, 57, 233–243.
- Iverson, J. M., & Goldin-Meadow, S. (2005). Gesture paves the way for language development. *Psychological Science*, 16(5), 367–371.
- Iverson, J. M., & Thelen, E. (1999). Hand, mouth and brain: The dynamic emergence of speech and gesture. *Journal of Consciousness Studies*, 6(11-12), 19–40.
- Jack, R. E., Garrod, O. G., Yu, H., Caldara, R., & Schyns, P. G. (2012). Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences of the United States of America*, 109(19), 7241–7244.
- Jackendoff, R. (1972). *Semantic Interpretation in Generative Grammar*. Cambridge, MA: MIT Press.
- Jaeger, J. J., Lockwood, A. H., van Valin, R. D., Kemmerer, D. L., Murphy, B. W., & Wack, D. S. (1998). Sex differences in brain regions activated by grammatical and reading tasks. *NeuroReport*, 9(12), 2803–2807.
- Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Archives of General Psychiatry*, 59(4), 434–446.
- Jannedy, S., & Mendoza-Denton, N. (2005). Structuring information through gesture and intonation. *Interdisciplinary Studies on Information Structure*, 3, 199–244.
- Jiménez-Bravo, M., & Marrero, V. (submitted). Multimodal perception of acoustic prominence in Spanish: A methodological proposal. *Speech Communication*.
- Johnson, F. M., Hicks, L. H., Goldberg, T., & Myslobodsky, M. S. (1988). Sex differences in lipreading. *Bulletin of the Psychonomic Society*, 26(2), 106–108.

Bibliography

- Jong, K. d. (1994). Initial tones and prominence in Seoul Korean. *Ohio State University Working Papers in Linguistics*, 43, 1–14.
- Jun, S.-A. (2005). Korean intonational phonology and prosodic transcription. In S.-A. Jun (Ed.) *Prosodic Typology: The Phonology of Intonation and Phrasing*, (pp. 201–229). Oxford: Oxford University Press.
- Jun, S.-A., & Lee, H.-J. (1998). Phonetic and phonological markers of contrastive focus in Korean. In *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP 98)*, (pp. 1295–1298). Sydney, Australia.
- Jurafsky, D., Bell, A., Gregory, M., & Raymond, W. D. (2001). Probabilistic relations between words: Evidence from reduction in lexical production. In J. L. Bybee, & P. J. Hopper (Eds.) *Frequency and the Emergence of Linguistic Structure (Typological Studies in Language, 45)*, (pp. 229–254). Amsterdam: John Benjamins.
- Kakouros, S. (2017). *Cognitive and probabilistic basis of prominence perception in speech*. Ph.D. thesis, School of Electrical Engineering, Aalto University.
- Kakouros, S., & Räsänen, O. (2015). Automatic detection of sentence prominence in speech using predictability of word-level acoustic features. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH - 2015*, (p. 568–572). Dresden, Germany.
- Karpiński, M., Jarmolowicz-Nowikow, E., & Malisz, Z. (2009). Aspects of gestural and prosodic structure of multimodal utterances in Polish task-oriented dialogues. *Speech and Language Technology*, 11, 113–122.
- Kendon, A. (1972). Some relationships between body motion and speech: An analysis of an example. In A. W. Siegman, & B. Pope (Eds.) *Studies in Dyadic Communication*, (pp. 177–210). New York: Pergamon.

Bibliography

- Kendon, A. (1980). Gesticulation and speech: Two aspects of the process of utterance. In M. R. Key (Ed.) *The Relationship of Verbal and Nonverbal Communication*, (pp. 207–227). The Hague: Mouton.
- Kendon, A. (1982). Coordination and framing in face-to-face interaction. In M. Davis (Ed.) *Interaction Rhythms*, (pp. 351–363). New York: Human Sciences Press.
- Kendon, A. (1988). How gestures can become like words. In F. Poyatos (Ed.) *Cross-cultural perspectives in nonverbal communication*, (pp. 133–141). Toronto: Hogrefe.
- Kendon, A. (2004). *Gesture: Visible Action as Utterance*. Cambridge: Cambridge University Press.
- Kießling, A., Kompe, R., Batliner, A., Niemann, H., & Nöth, E. (1996). Classification of boundaries and accents in spontaneous speech. In *Proceedings of the CRIM/FORWISS Workshop*, (pp. 104–113). Montreal.
- Kim, J., Cvejić, E., & Davis, C. (2014). Tracking eyebrows and head gestures associated with spoken prosody. *Speech Communication*, 57, 317–330.
- Kimbara, I. (2006). On gestural mimicry. *Gesture*, 6(1), 36–61.
- Kipp, M. (2003). *Gesture Generation by Imitation: From Human Behavior to Computer Character Animation*. Ph.D. thesis, Saarland University.
- Kita, S. (1993). *Language and thought interface: A study of spontaneous gestures and Japanese mimetics*. Ph.D. thesis, University of Chicago.
- Kita, S. (2000). How representational gestures help speaking. In D. McNeill (Ed.) *Language and Gesture*, (pp. 162–185). Cambridge: Cambridge University Press.
- Kita, S. (2009). Cross-cultural variation of speech-accompanying gesture: A review. *Language and Cognitive Processes*, 24(2), 145–167.

Bibliography

- Kita, S., & Davies, T. S. (2009). Competing conceptual representations trigger co-speech representational gestures. *Language and Cognitive Processes*, 24(5), 795–804.
- Kita, S., van Gijn, I., & van der Hulst, H. (1998). Movement phases in signs and co-speech gestures, and their transcription by human coders. In I. Wachsmuth, & M. Fröhlich (Eds.) *Gesture and Sign Language in Human-Computer Interaction*, (pp. 23–35). Berlin/Heidelberg/New York: Springer-Verlag.
- Klatt, D. H. (1973). Discrimination of fundamental frequency contours in synthetic speech: Implications for models of pitch perception. *Journal of the Acoustical Society of America*, 53(1), 8–16.
- Kochanski, G., Grabe, E., Coleman, J., & Rosner, B. (2005). Loudness predicts prominence: Fundamental frequency lends little. *Journal of the Acoustical Society of America*, 118(2), 1038–1054.
- Kohler, K. J. (1991). A model of German intonation. In K. J. Kohler (Ed.) *Studies in German Intonation*, (pp. 295–360). Kiel: IPDS.
- Kohler, K. J. (2003). Neglected categories in the modelling of prosody pitch timing and non-pitch accents. In *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS03)*, (pp. 2925–2928). Barcelona, Spain.
- Kohler, K. J. (2005). Form and function of non-pitch accents. In K. J. Kohler, F. Kleber, & P. Benno (Eds.) *AIPUK*, vol. 35a, (pp. 97–123). Kiel: IPDS.
- Kohler, K. J. (2006). Paradigms in experimental prosodic analysis: From measurement to function. In S. e. Sudhoff (Ed.) *Methods of Empirical Prosody Research*, (pp. 123–152). Berlin/New York: De Gruyter.
- Kohler, K. J. (2008). The perception of prominence patterns. *Phonetica*, 65(4), 257–269.
- Kohler, K. J. (2012). The perception of lexical stress in German: Effects of segmental duration and vowel quality in different prosodic patterns. *Phonetica*, 69(1), 1–26.

Bibliography

- Kohler, K. J., Pätzold, M., & P., S. A. (1997). The Kiel corpus of read/spontaneous speech – Acoustic data base, processing tool and analysis results. From the acoustic data collection to a labelled speech data bank of spoken standard German. In A. P. Simpson, K. J. Kohler, & T. Rettstadt (Eds.) *AIPUK*, vol. 32, (pp. 97–123). Kiel: IPDS.
- Kok, K., Bergmann, K., Cienki, A., & Kopp, S. (2016). Mapping out the multifunctionality of speakers' gestures. *Gesture*, 15(1), 37–59.
- Krahmer, E., Ruttkay, Z., Swerts, M., & Wesselink, W. (2002a). Pitch, eyebrows and the perception of focus. In *Proceedings of the 1st International Conference on Speech Prosody (SP2002)*, (pp. 443–446). Aix-en-Provence, France.
- Krahmer, E., Ruttkay, Z., Swerts, M., & Wesselink, W. (2002b). Perceptual evaluation of audiovisual cues for prominence. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH - 2002*, (pp. 1933–1936). Denver, USA.
- Krahmer, E., & Swerts, M. (2004). More about brows: A cross-linguistic analysis-by-synthesis study. In Z. Ruttkay, & C. Pelachaud (Eds.) *From brows to trust: Evaluating Embodied Conversational Agents. Human-Computer Interaction Series, No. 7*, (pp. 191–216). Dordrecht: Kluwer Academic Publishers.
- Krahmer, E., & Swerts, M. (2007). The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. *Journal of Memory and Language*, 57(3), 396–414.
- Krahmer, E., & Swerts, M. (2009). Audiovisual prosody—Introduction to the special issue. *Language and Speech*, 52(2-3), 129–133.
- Krauss, R. M., & Hadar, U. (1999). The role of speech-related arm/hand gestures in word retrieval. In R. N. Campbell, & L. Messing (Eds.) *Gesture, Speech, and Sign*, (pp. 93–116). Oxford: Oxford University Press.

- Krivokapić, J., Tiede, M. K., & Tyrone, M. E. (2015). A kinematic analysis of prosodic structure in speech and manual gestures. In *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS15)*, (pp. 1240–1244). Glasgow, UK: University of Glasgow.
- Krivokapić, J., Tiede, M. K., & Tyrone, M. E. (2017). A kinematic study of prosodic structure in articulatory and manual gestures: Results from a novel method of data collection. *Laboratory Phonology*, 8(1), 1–26.
- Krivokapić, J., Tiede, M. K., Tyrone, M. E., & Goldenberg, D. (2016). Speech and manual gesture coordination in a pointing task. In *Proceedings of the 8th International Conference on Speech Prosody (SP2016)*, (pp. 1240–1244). Boston, USA.
- Kruijff-Korbayová, I., & Steedman, M. (2003). Discourse and information structure. *Journal of Logic, Language and Information*, 12(3), 249–259.
- Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22(1), 79–86.
- Kushch, O., & Prieto Vives, P. (2016). The effects of pitch accentuation and beat gestures on information recall in contrastive discourse. In *Proceedings of the 8th International Conference on Speech Prosody (SP2016)*, (pp. 922–925). Boston, USA.
- Laan, G. P. M. (1997). The contribution of intonation, segmental durations, and spectral features to the perception of a spontaneous and a read speaking style. *Speech Communication*, 2(1), 43–65.
- Labov, W. (1968). The reflection of social processes in linguistic structures. In J. A. Fishman (Ed.) *Readings in the Sociology of Language*, (pp. 240–251). The Hague: Mouton.
- Ladd, D. R. (2008). *Intonational phonology*. Cambridge, MA: Cambridge University Press, 2nd ed.

Bibliography

- Ladefoged, P., Draper, M. H., & Whitteridge, D. (1958). Syllables and stress. In *Miscellaneous Phonetica III*, (pp. 1–14). London: International Phonetic Association, University College.
- Lea, W. A. (1977). Acoustic correlates of stress and juncture. In L. M. Hyman (Ed.) *Studies in stress and accent. Southern California Occasional Papers in Linguistics*, 4, (pp. 83–119). Los Angeles, CA: University of southern California.
- Lea, W. A. (1980). Prosodic aids to speech recognition. In W. A. Lea (Ed.) *Trends in Speech Recognition*, (pp. 83–119). New Jersey: Prentice Hall.
- Lee, H., & Noppeney, U. (2011). Long-term music training tunes how the brain temporally binds signals from multiple senses. *Proceedings of the National Academy of Sciences of the United States of America*, 108(51), E1441–E1450.
- Lehiste, I., & Peterson, G. E. (1959). Vowel amplitude and phonemic stress in American English. *Journal of the Acoustical Society of America*, 31(4), 428–435.
- Leonard, T., & Cummins, F. (2010). The temporal relation between beat gestures and speech. *Language and Cognitive Processes*, 26(10), 1457–1471.
- Leonetti, M., & Escandell-Vidal, V. (2009). Fronting and *verum* focus in spanish. In A. Dufter, & D. Jacob (Eds.) *Focus and Background in Romance Languages (Studies in Language Companion Series, 112)*, (pp. 155–204). Amsterdam: John Benjamins.
- Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.
- Levelt, W. J. M., Richardson, G., & La Heij, W. (1985). Pointing and voicing in deictic expressions. *Journal of Memory and Language*, 24(2), 133–164.
- Levy, R. P., & Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. In B. Schölkopf, J. C. Platt, & T. Hoffman (Eds.) *Advances in Neural Information Processing Systems*, (pp. 849–856). Cambridge, MA: MIT Press.

Bibliography

- Lewis, F., Butler, A., & Gilbert, L. (2011). A unified approach to model selection using the likelihood ratio test. *Methods in Ecology and Evolution*, 2(2), 155–162.
- Liang, C., Earl, B., Thompson, I., Whitaker, K., Cahn, S., Xiang, J., Fu, Q.-J., & Zhang, F. (2016). Musicians are better than non-musicians in frequency change detection: Behavioral and electrophysiological evidence. *Frontiers in Neuroscience*, 10:464.
- Liberman, A. M., & Prince, A. (1977). On stress and linguistic rhythm. *Linguistic Inquiry*, 8, 249–336.
- Liberman, M. (1979). *The intonational system of English*. New York: Garland Press.
- Lieberman, P. (1960). Some acoustic correlates of word stress in American English. *Journal of the Acoustical Society of America*, 32(4), 451–454.
- Llisterri, J., a, M. J., de la Mota, C., Riera, M., & Ríos, A. (2003b). The perception of lexical stress in Spanish. In *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS03)*, (pp. 2023–26). Barcelona, Spain.
- Llisterri, J., Machuca, M. J., de la Mota, C., Riera, M., & Ríos, A. (2003a). Algunas cuestiones en torno al desplazamiento acentual en español. In E. Z. Herrera, & P. Martín Butragueño (Eds.) *La tonía. Dimensiones fonéticas y fonológicas*, (pp. 163–185). Ciudad de México: Colegio de México.
- Llisterri, J., Machuca, M. J., de la Mota, C., Riera, M., & Ríos, A. (2005). La percepción del acento léxico en español. In *Filología y lingüística. Estudios ofrecidos a Antonio Quilis*, vol. 1, (pp. 271–297). Madrid: CSIC - UNED - Universidad de Valladolid.
- Llisterri, J., Machuca, M. J., Ríos, A., & Schwab, S. (2016). La percepción del acento léxico en un contexto oracional. *Loquens*, 3(2), e033.
- Loehr, D. (2007). Aspects of rhythm in gesture and speech. *Gesture*, 7(2), 179–214.
- Loehr, D. P. (2004). *Gesture and Intonation*. Ph.D. thesis, Georgetown University.

Bibliography

- Loehr, D. P. (2012). Temporal, structural, and pragmatic synchrony between intonation and gesture. *Laboratory Phonology*, 3(1), 71–89.
- Loftus, E. F., Miller, D. G., & Burns, H. J. (1978). Semantic integration of verbal information into a visual memory. *Journal of Experimental Psychology*, 4(1), 19–31.
- Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, 10(4), 325–337. Computer program. URL: <https://www.mrc-bsu.cam.ac.uk/software/bugs/the-bugs-project-winbugs/>.
- Mahrt, T., Cole, J., Fleck, M., & Hasegawa-Johnson, M. (2012). F0 and the perception of prominence. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH - 2012*, (pp. 2421–2424). Portland, USA.
- Mahrt, T., Huang, J.-T., Mo, Y., Fleck, M., Hasegawa-Johnson, M., & Cole, J. (2011). Optimal models of prosodic prominence using the Bayesian information criterion. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH - 2011*, (pp. 2037–2040). Florence, Italy.
- Mallery, G. (1881). Sign language among North American Indians compared with that among other peoples and deaf mutes. In *First Annual Report of the Bureau of Ethnology to the Secretary of the Smithsonian Institution, 1879-1880*, (pp. 263–552). Washington: Government Printing Office.
- Manolescu, A., Olson, D., & Ortega-Llebaria, M. (2009). Cues to contrastive focus in Romanian. In M. Vigário, S. Frota, & M. J. Freitas (Eds.) *Phonetics and Phonology: Interactions and Interrelations. Current Issues in Linguistic Theory*, vol. 306, (pp. 71–90). Amsterdam: John Benjamins.
- Martínez Amador, E. M. (1954). *Diccionario gramatical*. Barcelona: Ramón Sopena.

Bibliography

- Masson-Carro, I., Goudbeek, M., & Kraemer, E. (2017). How what we see and what we know influence iconic gesture production. *Journal of Nonverbal Behavior*, 41(4), 367–394.
- Mazerolle, M. J. (2017). AICcmodavg: Model selection and multimodel inference based on (Q)AIC(c). R package version 2.1-1, URL: <https://cran.r-project.org/package=AICcmodavg>.
- McClave, E. (1991). *Intonation and Gesture*. Ph.D. thesis, Georgetown University.
- McClave, E. (1994). Gestural beats: The rhythm hypothesis. *Journal of Psycholinguistic Research*, 23(1), 45–66.
- McClave, E. (1998). Pitch and manual gestures. *Journal of Psycholinguistic Research*, 27(1), 69–89.
- McClave, E. (2000). Linguistic functions of head movements in the context of speech. *Journal of Pragmatics*, 32(7), 855–878.
- McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264, 746–748.
- McNeill, D. (1985). So you think gestures are nonverbal? *Psychological Review*, 92(3), 350–371.
- McNeill, D. (1992). *Hand and Mind: What Gestures Reveal about Thought*. Chicago: University of Chicago Press.
- McNeill, D. (2000). Introduction. In D. McNeill (Ed.) *Language and Gesture*, (pp. 1–10). Cambridge: Cambridge University Press.
- McNeill, D. (2005). *Gesture and Thought*. Chicago: University Chicago Press.
- McNeill, D. (2012). *How Language Began*. Cambridge, MA: Cambridge University Press.

Bibliography

- McQuown, N. A. (1971). *The Natural History of an Interview*. Chicago: University of Chicago Library.
- Meister, I. G., Boroojerdi, B., Foltys, H., Sparing, R., Huber, W., & Töpper, R. (2003). Motor cortex hand area and speech: Implications for the development of language. *Neuropsychologia*, 41(4), 401–406.
- Mirković, J., & Gaskell, M. G. (2016). Does sleep improve your grammar? Preferential consolidation of arbitrary components of new linguistic knowledge. *PLoS ONE*, 11(4), e0152489.
- Mittelberg, I., & Evola, V. (2013). Iconic and representational gestures. In C. Müller, A. Cienki, E. Fricke, S. H. Ladewig, D. McNeill, & J. Bressemer (Eds.) *Body – Language – Communication. An International Handbook on Multimodality in Human Interaction (Handbooks of Linguistics and Communication Science 38.2.)*, (pp. 1732–1746). Berlin/Boston: De Gruyter.
- Mixdorff, H. (1998). *Intonation Patterns of German - Model-based. Quantitative Analysis and Synthesis of F0-Contours*. Ph.D. thesis, TU Dresden.
- Mixdorff, H., & Widera, C. (2001). Perceived prominence in terms of a linguistically motivated quantitative intonation model. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH - 2001*, (pp. 403–406). Aalborg, Denmark.
- Mo, Y. (2008a). Acoustic correlates of prosodic prominence for naïve listeners of American English. *Annual Meeting of the Berkeley Linguistics Society*, 34(1), 257–267.
- Mo, Y. (2008b). Duration and intensity as perceptual cues for naïve listeners' prominence and boundary perception. In *Proceedings of the 4th International Conference on Speech Prosody (SP2008)*, (pp. 739–742). Campinas, Brazil.

Bibliography

- Mo, Y., Cole, J., & Lee, E.-K. (2008). Naïve listeners' prominence and boundary perception. In *Proceedings of the 4th International Conference on Speech Prosody (SP2008)*, (pp. 735–738). Campinas, Brazil.
- Mol, H. C., & Uhlenbeck, G. M. (1956). The linguistic relevance of intensity in stress. *Lingua*, 5, 205–213.
- Mol, L., Krahmer, E., Maes, A., & Swerts, M. (2012). Adaptation in gesture: Converging hands or converging minds? *Journal of Memory and Language*, 66(1), 249–264.
- Montgomery, K. J., Isenberg, N., & Haxby, J. V. (2007). Communicative hand gestures and object-directed hand movements activated the mirror neuron system. *Social Cognitive and Affective Neuroscience*, 2(2), 114–122.
- Morrel-Samuels, P., & Krauss, R. M. (1992). Word familiarity predicts temporal asynchrony of hand gestures and speech. *Journal of Experimental Psychology: Human Learning and Memory*, 18(3), 615–622.
- Morton, J., & Jassem, W. (1965). Acoustic correlates of stress. *Language and Speech*, 8(3), 159–181.
- Mukamel, R., Ekstrom, A. D., Kaplan, J., Iacoboni, M., & Fried, I. (2010). Single-neuron responses in humans during execution and observation of actions. *Current Biology*, 20(8), 750–756.
- Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility. *Psychological Science*, 15(2), 133–137.
- Nakano, Y. I., Reinstein, G., Stocky, T., & Cassell, J. (2003). Towards a model of face-to-face grounding. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, (pp. 354–365). Sapporo, Japan.
- Nakatani, L., & Aston, C. (1978). Perceiving stress patterns of words in sentences. *Journal of the Acoustical Society of America*, 63, S55.

Bibliography

- Navarro Tomás, T. (1918). *Manual de pronunciación española*. Madrid: RFE.
- Navarro Tomás, T. (1944). *Manual de entonación española*. New York: Hispanic Institute in the United States.
- Navarro Tomás, T. (1964). La medida de la intensidad. *Boletín del Instituto de Filología de la Universidad de Chile*, 16, 231–235.
- Nenkova, A., Brenier, J., Kothari, A., Calhoun, S., Whitton, L., Beaver, D., & Jurafsky, D. (2007). To Memorize or to Predict: Prominence Labeling in Conversational Speech. In *Human Language Technology Conference of the North American Chapter of the association of Computational Linguistics*, (pp. 9–16). Rochester, NY, USA.
- Nobe, S. (1996). *Representational gestures, cognitive rhythms, and acoustic aspects of speech: A network/threshold model of gesture production*. Ph.D. thesis, University of Chicago.
- Noland, C. (2009). *Agency and Embodiment. Performing Gestures/Producing Culture*. Cambridge, MA: Harvard University Press.
- Nooteboom, S. G., & Kruyt, J. G. (1987). Accents, focus distribution, and the perceived distribution of given and new information: An experiment. *Journal of the Acoustical Society of America*, 82(2), 1512–1524.
- Novack, M. A., & Goldin-Meadow, S. (2017). Gesture as Representational Action: A paper about function. *Psychonomic Bulletin and Review*, 24(3), 652–665.
- Öhrström, N., & Traunmüller, H. (2004). Audiovisual perception of Swedish vowels with and without conflicting cues. In P. Branderud, & H. Traunmüller (Eds.) *Proceedings of FONETIK 2004: The 17th Swedish Phonetics Conference*, (pp. 40–43). Dept. of Linguistics, Stockholm University.

Bibliography

- Ortega-Llebaria, M. (2006). Phonetic cues to stress and accent in Spanish. In M. Díaz-Campos (Ed.) *Selected Proceedings of the 2nd Conference on Laboratory Approaches to Spanish Phonology*, (pp. 104–118). Somerville, USA: Cascadilla Press.
- Ortega-Llebaria, M., & Prieto, P. (2007). Disentangling stress from accent in Spanish: Production patterns of the stress contrast in deaccented syllables. In P. Prieto, J. Mascará, & M.-J. Solé (Eds.) *Segmental and Prosodic Issues in Romance Phonology (Current Issues in Linguistic Theory, 282)*, (pp. 155–176). Amsterdam: John Benjamins.
- Ortega-Llebaria, M., & Prieto, P. (2009). Perception of word stress in Castilian Spanish. In M. Vigário, S. Frota, & M. J. Freitas (Eds.) *Phonetics and Phonology: Interactions and interrelations*, (pp. 35–50). Amsterdam: John Benjamins.
- Ortega-Llebaria, M., & Prieto, P. (2011). Acoustic correlates of stress in central Catalan and Castilian Spanish. *Language and Speech*, 54(1), 73–97.
- Ortega-Llebaria, M., Prieto, P., & Vanrell, M. (2007). Perceptual evidence for direct acoustic correlates of stress in Spanish. In *Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS07)*, (pp. 1121–1124). Saarbrücken, Germany.
- Özyürek, A., Kita, S., Allen, S., Furman, R., & Brown, A. (2005). How does linguistic framing of events influence co-speech gestures? Insights from cross-linguistic variations and similarities. *Gesture*, 5(1-2), 219–240.
- Özyürek, A., Willems, R. M., Kita, S., & Hagoort, P. (2007). On-line integration of semantic information from speech and gesture: insights from event-related brain potentials. *Journal of Cognitive Neuroscience*, 19(4), 605–616.
- Pamies, A. (1994). *Acento, ritmo y lenguaje*. Ph.D. thesis, Universidad de Granada.
- Pamies, A. (1997). Consideraciones sobre la marca acústica del acento fonológico. *Estudios de Fonética Experimental*, 8, 11–50.

Bibliography

- Pamies, A. (2003). The relation between stress and tonal peaks. In *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS03)*, (pp. 2013–2034). Barcelona, Spain.
- Pamies, A., Fernández, A. M., Martínez, E., Ortega, A., & Amorós, M. C. (2002). Umbrales tonales en el español peninsular. In J. Díaz García (Ed.) *Actas del II Congreso de Fonética Experimental*, (pp. 272–278). Sevilla, Spain.
- Paraskevopoulos, E., Kuchenbuch, A., S.C., H., & Pantev, C. (2012). Musical expertise induces audiovisual integration of abstract congruency rules. *Clinical Neurophysiology*, 32(50), 18196–18203.
- Parbery-Clark, A., Skoe, E., Lam, C., & Kraus, N. (2009). Musician enhancement for speech-in-noise. *Ear and Hearing*, 30(6), 653–661.
- Parrell, B., Goldstein, L., Lee, S., & Byrd, D. (2014). Spatiotemporal coupling between speech and manual motor actions. *Journal of Phonetics*, 42, 1–11.
- Patel, A. D., Peretz, I., Tramo, M., & Labrecque, R. (1998). Processing prosodic and music patterns: A neuropsychological investigation. *Brain and Language*, 61(1), 123–144.
- Pelachaud, C., Badler, N., & Steedman, M. (1996). Generating facial expressions for speech. *Cognitive Science*, 20(1), 1–46.
- Peters, B. (2005). The database – the Kiel corpus of spontaneous speech. In K. J. Kohler, F. Kleber, & P. Benno (Eds.) *AIPUK*, vol. 35a, (pp. 1–6). Kiel: IPDS.
- Pierrehumbert, J. B. (1979). The perception of fundamental frequency declination. *Journal of the Acoustical Society of America*, 62(2), 363–369.
- Pierrehumbert, J. B. (1980). *The phonology and phonetics of English intonation*. Ph.D. thesis, Massachusetts Institute of Technology.
- Pierrehumbert, J. B., & Beckman, M. E. (1988). *Japanese Tone Structure*. Cambridge, MA: MIT Press.

Bibliography

- Pika, S., Liebal, K., Call, J., & Tomasello, M. (2007). The gestural communication of apes. In K. Liebal, C. Müller, & S. Pika (Eds.) *Gestural Communication in Nonhuman and Human Primates (Benjamins Current Topics, 10)*, (pp. 35–49). Amsterdam: John Benjamins.
- Pike, L. (1945). *The intonation of American English*. Ann Arbor, MI: University of Michigan Press.
- Pitt, M., Dille, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., & Fosler-Lussier, E. (2007). *Buckeye Corpus of Conversational Speech (2nd release)*. Columbus: Department of Psychology, Ohio State University (Distributor). URL: <https://www.buckeyecorpus.osu.edu>.
- Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC - 2003)*. Vienna, Austria. Computer program. URL: <https://www.http://mcmc-jags.sourceforge.net/>.
- Pluymaekers, M., Ernestus, M., & Baayen, R. H. (2005). Lexical frequency and acoustic reduction in spoken Dutch. *Journal of the Acoustical Society of America*, 118(4), 2561–2569.
- Portele, T., & Heuft, B. (1997). Towards a prominence-based synthesis system. *Speech Communication*, 21(1), 61–72.
- Potisuk, S., Gandour, J., & Harper, M. (1996). Acoustic correlates of stress in Thai. *Phonetica*, 53(4), 200–220.
- Powell, M. J. D. (2009). The BOBYQA algorithm for bound constrained optimization without derivatives. *Department of Applied Mathematics and Theoretical Physics, Cambridge England, technical report NA2009/06*.

Bibliography

- Prieto, P., D'Imperio, M., & Gili-Fivela, B. (2005). Pitch accent alignment in romance: Primary and secondary associations with metrical structure. *Language and Speech*, 48(4), 359–396.
- Prieto, P., & Ortega-Llebaria, M. (2006). Stress and accent in Catalan and Spanish: Patterns of duration, vowel quality, overall intensity, and spectral balance. In *Proceedings of the 3rd International Conference on Speech Prosody (SP2006)*, (pp. 337–340). Dresden, Germany.
- Prieto, P., & Ortega-Llebaria, M. (2009). Do complex pitch gestures induce syllable lengthening in Catalan and Spanish? In M. Vigário, S. Frota, & M. J. Freitas (Eds.) *Phonetics and Phonology: Interactions and Interrelations (Current Issues in Linguistic Theory, 306)*, (pp. 51–70). Amsterdam: John Benjamins.
- Prieto, P., Puglesi, C., Borràs-Comes, J., Arroyo, E., & Blat, J. (2011). Crossmodal prosodic and gestural contribution to the perception of contrastive focus. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH - 2011*, (pp. 977–980). Florence, Italy.
- Prieto, P., Puglesi, C., Borràs-Comes, J., Arroyo, E., & Blat, J. (2015). Exploring the contribution of prosody and gesture to the perception of focus using an animated agent. *Journal of Phonetics*, 49, 41–54.
- Prieto, P., Vanrell, M. d. M., Astruc, L., Payne, E., & Post, B. (2012). Phonotactic and phrasal properties of speech rhythm: Evidence from Catalan, English, and Spanish. *Speech Communication*, 54(6), 681–702.
- Prince, A., & Smolensky, P. (1993). Optimality theory: Constraint interaction in generative grammar. *Rutgers Center for Cognitive Science Technical Report TR-2*.
- Prince, E. F. (1981). Topicalization, focus-movement, and Yiddish-movement: A pragmatic differentiation. In *Proceedings of the 7th Annual Meeting of the Berkeley Linguistics Society*, (pp. 249–264). Berkeley, USA.

Bibliography

- Proverbio, A. M., Massetti, G., Rizzi, E., & Zani, A. (2016). Skilled musicians are not subject to the McGurk effect. *Scientific Reports*, 6, 30423.
- Quak, M., London, R. E., & Talsma, D. (2015). A multisensory perspective of working memory. *Frontiers in Human Neuroscience*, 9:197.
- Quené, H. (2008). Multilevel modeling of between-speaker and within-speaker variation in spontaneous speech tempo. *Journal of the Acoustical Society of America*, 123(2), 1103–1113.
- Quené, H., & Kager, R. (1993). Prosodic sentence analysis without exhaustive parsing. In V. J. van Heuven, & L. C. W. Pols (Eds.) *Analysis and synthesis of speech: Strategic research towards high-quality TTS generation*, (pp. 115–130). Berlin/New York: De Gruyter.
- Quené, H., & van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language*, 59(4), 413–425.
- Quilis, A. (1965). Phonologie de la quantité en espagnol. *Phonetica*, 13, 82–85.
- Quilis, A. (1971). Caracterización fonética del acento español. *Travaux de Linguistique et de Littérature*, 9, 53–72.
- Quilis, A. (1981). *Fonética acústica de la lengua española*. Madrid: Gredos.
- Quilis, A., & Esgueva, M. (1983). Realización de los fonemas vocálicos españoles en posición fonética normal. In M. Esgueva, & M. Cantarero (Eds.) *Estudios de Fonética I*, (pp. 159–252). Madrid: Consejo Superior de Investigaciones Científicas (CSIC).
- R Development Core Team (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Computer program. Version 3.5.1 retrieved 2 July 2018 from <https://www.R-project.org/>.

Bibliography

- Ramus, F., Nespors, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, 73(3), 265–292.
- Real Academia Española (1959). *Gramática de la lengua española*. Madrid: Espasa-Calpe.
- Ricci Bitti, P. E. (2013). Facial expression and social interaction. In C. Müller, A. Cienki, E. Fricke, S. H. Ladewig, D. McNeill, & J. Bressemer (Eds.) *Body – Language – Communication. An International Handbook on Multimodality in Human Interaction (Handbooks of Linguistics and Communication Science 38.2)*, (pp. 1342–1349). Berlin/Boston: De Gruyter.
- Richards, S. A. (2008). Dealing with overdispersed count data in applied ecology. *Journal of Applied Ecology*, 45(1), 218–227.
- Rietveld, A., & Gussenhoven, C. (1985). On the relation between pitch excursion size and prominence. *Journal of Phonetics*, 13(3), 299–308.
- Rietveld, A., & Koopmans-van Beinum, F. (1987). Vowel reduction and stress. *Speech Communication*, 6(3), 217–229.
- Ríos, A. (1991). *Caracterización acústica del ritmo del castellano. Trabajo de investigación de tercer ciclo*. Ph.D. thesis, Universidad Autónoma de Barcelona.
- Rochet-Capellan, A., Laboissière, R., Galván, A., & Schwartz, J.-L. (2008). The speech focus position effect on jaw-finger coordination in a pointing task. *Journal of Speech, Language, and Hearing Research*, 51(6), 1507–1521.
- Rodríguez, C., Moreno-Núñez, A., Basilio, M., & Sosa, N. (2015). Ostensive gestures come first: Their role in the beginning of shared reference. *Cognitive Development. Special Issue: Semiotic and Cognition in Human Development*, 36, 142–149.
- Rosenberg, A., & Hirschberg, J. (2009). Detecting pitch accents at the word, syllable and vowel level. In *Human Language Technologies – The 2009 Annual Conference of the*

Bibliography

- North American Chapter of the Association for Computational Linguistics (HLT NAACL - 2009)*, (pp. 81–84). Boulder CO, USA.
- Roustan, B., & Dohen, M. (2010). Co-production of contrastive prosodic focus and manual gestures: Temporal coordination and effects on the acoustic and articulatory correlates of focus. In *Proceedings of the 5th International Conference on Speech Prosody (SP2010)*, 100110, (pp. 1–4). Chicago, USA.
- RStudio Team (2016). *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA. Computer program. Version 1.1.456 retrieved 19 July 2018 from <http://www.rstudio.com/>.
- Rue, H., Martino, S., & Chopin, N. (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2), 319–392.
- Ruiz, M., & Pereira, Y. (2010). Acento léxico: tendencias de los correlatos acústicos. *Onomázein*, 22(2), 43–58.
- Rusiewicz, H. L. (2010). *The role of prosodic stress and speech perturbation on the temporal synchronization of speech and deictic gestures*. Ph.D. thesis, University of Pittsburgh.
- Ruytjens, L., Albers, F., van Dijk, P., Wit, H., & Willemsen, A. (2006). Neural responses to silent lipreading in normal hearing male and female subjects. *European Journal of Neuroscience*, 24(6), 1835–1844.
- Ruytjens, L., Georgiadis, J. R., Holstege, G., Wit, H. P., Albers, F. W., & Willemsen, A. T. (2007). Functional sex differences in human primary auditory cortex. *European Journal of Nuclear Medicine and Molecular Imaging*, 34(12), 2073–2081.
- Santerre, L., & Bothorel, A. (1969). Mesure et interprétation de la ligne d'intensité sur les sonogrames. *Travaux de l'institut de phonétique de Strasbourg*, 2, 82–90.

Bibliography

- Scarborough, R., Keating, P., Mattys, S. L., Cho, T., & Alwan, A. (2009). Optical phonetics and visual perception of lexical and phrasal stress in English. *Language and Speech*, 52(2-3), 135–175.
- Schafer, A., Carlson, K., Clifton Jr., C., & Frazier, L. (2000). Focus and the interpretation of pitch accent: Disambiguating embedded questions. *Language and Speech*, 43(1), 75–105.
- Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika*, 78(4), 719–727.
- Schefflen, A. E. (1964). The Significance of Posture in Communication Systems. *Psychiatry*, 27(4), 316–331.
- Schegloff, E. A. (1984). On some gestures' relation to talk. In J. M. Atkinson (Ed.) *Structures of Social Action: Studies in Conversation Analysis*, (pp. 266–295). Cambridge: Cambridge University Press.
- Schielzeth, H., & Nakagawa, S. (2012). Nested by design: Model fitting and interpretation in a mixed model era. *Methods in Ecology and Evolution*, 4(1), 14–24.
- Schmerling, S. F. (1976). *Aspects of English sentence stress*. Austin, TX: University of Texas Press.
- Schröder, M., Cowie, R., Douglas-Cowie, E., Westerdijk, M., & Gielen, S. (2001). Acoustic correlates of emotion dimensions in view of speech synthesis. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH - 2001*, (pp. 87–90). Aalborg, Denmark.
- Schwartz, D. L. (1995). The emergence of abstract representations in dyad problem solving. *Journal of the Learning Science*, 4(3), 321–354.
- Schwarzschild, R. (1999). GIVENness, AvoidF and other constraints on the placement of accent. *Natural Language Semantics*, 7(2), 141–177.

Bibliography

- Selkirk, E. O. (1984). *Phonology and syntax: the relation between sound and structure*. Cambridge, MA: MIT Press.
- Sgall, P., & Hajičová, E. (1977). Focus on focus (Part I). *The Prague Bulletin of Mathematical Linguistics*, 28, 5–54.
- Sgall, P., & Hajičová, E. (1978). Focus on focus (Part II). *The Prague Bulletin of Mathematical Linguistics*, 29, 23–41.
- Shattuck-Hufnagel, S., Yasinnik, Y., Veilleux, N., & Renwick, M. (2007). A method for studying the time alignment of gestures and prosody in American English: 'hits' and pitch accents in academic-lecture-style speech. In A. Esposito, M. Bratanić, E. Keller, & M. Marinaro (Eds.) *Fundamentals of Verbal and Nonverbal Communication and the Biometric Issue, NATO Publishing Sub-Series E: Human and Societal Dynamics, vol. 18*, (pp. 1079–1098). Amsterdam: IOS Press.
- Shukla, S. (1996). Śikṣa:s, pra:tiśa:khyas, and the Vedic accent. In K. R. Jankowsky (Ed.) *Multiple Perspectives on the Historical Dimensions of Language*, (pp. 269–279). Münster: Nodus Publikationen.
- Sievers, E. (1901). *Grundzüge der Phonetik. Bibliothek indogermanischer Grammatiken 1*. Leipzig: Breitkopf und Härtel.
- Silipo, R., & Greenberg, S. (1999). Automatic transcription of prosodic stress for spontaneous English discourse. In *Proceedings of 14th International Congress of Phonetic Sciences (ICPhS99)*, (pp. 2351–2354). San Francisco, USA.
- Silipo, R., & Greenberg, S. (2000). Prosodic stress revisited: Reassessing the role of fundamental frequency. In *Proceedings of the NIST Speech Transcription Workshop*. College Park, USA.
- Silverman, K., Beckman, M. E., Pitrelli, J. F., Ostendorf, M., W. Wightman, C. W., Price, P. J., Pierrehumbert, J. B., & Hirschberg, J. (1992). ToBI: A standard for labeling Eng-

Bibliography

- lish prosody. In *Proceedings of the 2nd International Conference on Spoken Language Processing (ICSLP92)*, (pp. 981–984). Banff, Canada.
- Singmann, H., & Kellen, D. (in press). An introduction to mixed models for experimental psychology. In D. H. Spieler, & E. Schumacher (Eds.) *New Methods in Neuroscience and Cognitive Psychology*. Psychology Press.
- Skopeteas, S., & Fanselow, G. (2010). Focus types and argument asymmetries: a cross-linguistic study in language production. In C. Breul, & E. Göbbel (Eds.) *Contrastive information structure (Linguistics Today Series, 165)*, (pp. 169–197). Amsterdam: John Benjamins.
- Sluijter, A. M. C., & van Heuven, V. J. (1996a). Spectral balance as an acoustic correlate of linguistic stress. *Journal of the Acoustical Society of America*, 100(4), 2471–85.
- Sluijter, A. M. C., & van Heuven, V. J. (1996b). Acoustic correlates of linguistic stress and accent in Dutch and American English. In *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP96)*, (pp. 630–633). Philadelphia, USA.
- Sluijter, A. M. C., van Heuven, V. J., & Pacilly, J. J. A. (1997). Spectral balance as a cue in the perception of linguistic stress. *Journal of the Acoustical Society of America*, 101(1), 503–513.
- Solé, M. J. (1984). Experimentos sobre la percepción del acento. *Estudios de Fonética Experimental*, 1, 134–243.
- Sonderegger, M. (2012). *Phonetic and phonological dynamics on reality television*. Ph.D. thesis, University of Chicago.
- Sorin, C. (1981). Functions, roles and treatments of intensity in speech. *Journal of Phonetics*, 9, 359–374.

Bibliography

- Srinivasan, R., & Massaro, D. (2003). Perceiving prosody from the face and voice: Distinguishing statements from echoic questions in English. *Language and Speech*, 46(1), 1–22.
- Stetson, R. H. (1928). *Motor Phonetics*. Amsterdam: North-Holland.
- Stokoe, W. C. (2001). *Language in hand: Why sign came before speech*. Washington, DC: Gallaudet University Press.
- Strait, D. L., & Kraus, N. (2011). Can you hear me now? Musical training shapes functional brain networks for selective auditory attention and hearing speech in noise. *Frontiers in Psychology*, 13(2), 113.
- Streefkerk, B. (2002). *Acoustic and lexical/syntactic correlates*. Ph.D. thesis, LOT (Landelijke Oriëntatiecursus Theaterscholen), Utrecht.
- Streefkerk, B. M., Pols, L. C. W., & ten Bosch, L. F. M. (1997). Prominence in read-aloud sentences, as marked by listeners and classified automatically. In *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam*, 21, (pp. 101–116). Amsterdam, the Netherlands: University of Amsterdam.
- Streefkerk, B. M., Pols, L. C. W., & ten Bosch, L. F. M. (1998). Automatic detection of prominence (as defined by listeners' judgements) in read aloud sentences. In *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP 98)*, (pp. 683–686). Sydney, Australia.
- Strelnikov, K., Rouger, J., Lagleyre, S., Fraysse, B., Deguine, O., & Barone, P. (2009). Improvement in speech-reading ability by auditory training: evidence from gender differences in normally hearing, deaf and cochlear implanted subjects. *Neuropsychologia*, 47(4), 972–979.
- Sweet, H. (1877). *A Handbook of Phonetics*. Oxford: Clarendon.

Bibliography

- Swerts, M. (1997). Prosodic features at discourse boundaries of different strength. *Journal of the Acoustical Society of America*, 101(1), 514–521.
- Swerts, M., & Krahmer, E. (2004). Congruent and incongruent audiovisual cues to prominence. In *Proceedings of the 2nd International Conference on Speech Prosody (SP2004)*, (pp. 69–72). Nara, Japan.
- Swerts, M., & Krahmer, E. (2005). Audiovisual prosody and feeling of knowing. *Journal of Memory and Language*, 53(1), 81–94.
- Swerts, M., & Krahmer, E. (2008). Facial expression and prosodic prominence: Effects of modality and facial area. *Journal of Phonetics*, 36(2), 219–238.
- Swerts, M., & Krahmer, E. (2010). Visual prosody of newsreaders: Effects of information structure, emotional content and intended audience on facial expressions. *Journal of Phonetics*, 38(2), 197–206.
- Swerts, M., Krahmer, E., & Avesani, C. (2002). Prosodic marking of information status in Dutch and Italian: A comparative analysis. *Journal of Memory and Language*, 30(4), 629–654.
- Szaszák, G., & Beke, A. (2017). An empirical approach for comparing syntax and prosody driven prominence marking. *The Phonetician*, 114(1), 46–57.
- Szendrői, K. (2001). *Focus and the Syntax-Phonology Interface*. Ph.D. thesis, University College London.
- 't Hart, J. (1981). Differential sensitivity to pitch distance, particularly in speech. *Journal of the Acoustical Society of America*, 69(3), 811–821.
- 't Hart, J., Collier, R., & Cohen, A. (1990). *A perceptual study of intonation: An experimental-phonetic approach to speech melody*. Cambridge: Cambridge University Press.

Bibliography

- Tamburini, F. (2003). Prosodic prominence detection in speech. In *Conference: Signal Processing and Its Applications, 2003. Proceedings. Seventh International Symposium*, vol. 1, (pp. 385–388). Paris, France.
- ten Bosch, L. F. M. (1993). On the automatic classification of pitch movements. In *Proceedings of the 3rd European Conference on Speech Communication and Technology (Eurospeech - 1993)*, (pp. 781–784). Berlin, Germany.
- Terken, J. (1991). Fundamental frequency and perceived prominence of accented syllables. *Journal of the Acoustical Society of America*, 89(4), 1768–1776.
- Terken, J. (1994). Fundamental frequency and perceived prominence of accented syllables. (II. Nonfinal accents). *Journal of the Acoustical Society of America*, 95(6), 3662–3665.
- Terken, J. (1996). Variation of accent prominence within the phrase: Models and spontaneous speech data. In Y. Sagisaka, C. N., & N. Higuchi (Eds.) *Computing Prosody. Approaches to a Computational Analysis and Modelling of the Prosody of Spontaneous Speech*, (pp. 95–111). Berlin/Heidelberg/New York: Springer-Verlag.
- Terken, J., & Hermes, D. (2000). The perception of prosodic prominence. In M. Horne (Ed.) *Prosody: Theory and experiment. Studies presented to Gösta Bruce*, (pp. 89–127). Dordrecht: Kluwer Academic Publishers.
- Terken, J., & Hirschberg, J. (1994). Deaccentuation of words representing ‘given’ information: Effects of persistence of grammatical function and surface position. *Language and Speech*, 37(2), 125–145.
- Terken, J., & Nootboom, S. G. (1987). Opposite effects of accentuation and deaccentuation on verification latencies for given and new information. *Language and Cognitive Processes*, 2(3-4), 145–163.
- Thompson, W. F., Schellenberg, E. G., & Husain, G. (2004). Decoding speech prosody: Do music lessons help? *Emotion*, 4(1), 46–64.

Bibliography

- Toledo, G. A., Fernández Planas, A. M., Romera, L., Ortega, A., & Matas, J. (2001). Tiempo y tono en español peninsular. In M. Díaz García (Ed.) *Actas del II Congreso Nacional de Fonética Experimental*, (pp. 318–323). Sevilla: Universidad de Sevilla.
- Tomasello, M., & Call, J. (1997). *Primate Cognition*. Oxford: Oxford University Press.
- Torreira, F., Simonet, M., & Hualde, J. I. (2014). Quasi-neutralization of stress contrasts in Spanish. In *Proceedings of the 7th International Conference on Speech Prosody (SP2014)*, (pp. 197–201). Dublin, Ireland.
- Truckenbrodt, H. (1999). On the relation between syntactic phrases and phonological phrases. *Linguistic Inquiry*, 30(2), 219–255.
- Tuite, K. (1993). The production of gesture. *Semiotica*, 93(1-2), 83–105.
- Turk, A., & Sawusch, J. (1996). The processing of duration and intensity cues to prominence. *Journal of the Acoustical Society of America*, 99(6), 3782–90.
- Tylor, E. B. (1865). *Researches into the Early History of Mankind and the Development of Civilization*. London: John Murray.
- Vainio, M., & Jarvikivi, J. (2006). Tonal features, intensity, and word order in the perception of prominence. *Journal of Phonetics*, 34(3), 319–342.
- Vallduví, E. (1992). *The informational component*. Ph.D. thesis, University of Pennsylvania.
- Vallduví, E., & Engdahl, E. (1996). The linguistic realization of information packaging. *Linguistics*, 34(3), 459–520.
- van Bergem, D. R. (1993). Acoustic vowel reduction as a function of sentence accent, word stress, and word class. *Speech Communication*, 12(1), 1–23.
- van der Hulst, H. (1985). *Syllable Structure and Stress in Dutch*. Dordrecht: Foris.

Bibliography

- van Kuijk, D., & Boves, L. (1999). Acoustic characteristics of lexical stress in continuous telephone speech. *Speech Communication*, 27(2), 95–111.
- van Maastricht, L., Krahmer, E., & Swerts, M. (2016). Prominence patterns in a second language: Intonational transfer from Dutch to Spanish and vice versa. *Language Learning*, 66(1), 124–158.
- van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences of the United States of America*, 102(4), 1181–1186.
- Vanrell, M., & Fernández Soriano, O. (2013). Variation at the interfaces in Ibero-Romance. Catalan and Spanish prosody and word order. *Catalan Journal of Linguistics*, 12, 253–282.
- Venditti, J., Maekawa, K., & Beckman, M. E. (2008). Korean intonational phonology and prosodic transcription. In S. Miyagawa, & M. Saito (Eds.) *Handbook of Japanese Linguistics*, (pp. 456–512). Oxford: Oxford University Press.
- Vico, G. (1725/1744). *Scienza nuova [New science]*. Naples: Stamperia Muziana.
- Vogel, I., Athanasopoulou, A., & Pincus, N. (2016). Prominence, Contrast and the Functional Load Hypothesis: an acoustic investigation. In J. Heinz, R. Goedemans, & H. van der Hulst (Eds.) *Dimensions of Phonological Stress*, (pp. 123–167). Cambridge: Cambridge University Press.
- Wagner, P. (2005). Great expectations – introspective vs. perceptual prominence ratings and their acoustic correlates. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH - 2005*, (pp. 2381–2384). Lisbon, Portugal.
- Wagner, P., Malisz, Z., & Kopp, S. (2014). Gesture and speech in interaction: An overview. *Speech Communication*, 57, 209–232.

Bibliography

- Wagner, P., Origlia, A., Avesani, C., Christodoulides, G., Cutugno, F., D'Imperio, M., Escudero Mancebo, D., Gili Fivela, B., Lacheret, A., Ludusan, B., Moniz, H., Ní Chasaide, A., Niebuhr, O., Rousier-Vercruyssen, L., Simon, A. C., Simko, J., Tesser, F., & Vainio, M. (2015). Different parts of the same elephant: A roadmap to disentangle and connect different perspectives on prosodic prominence. In *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS15)*. Glasgow, UK: University of Glasgow.
- Warton, D. I., & Hui, F. K. (2011). The arcsine is asinine: the analysis of proportions in ecology. *Ecology*, *92*(1), 3–10.
- Watson, C. S., Qiu, W. W., Chamberlain, M. M., & Li, X. (1996). Auditory and visual speech perception: confirmation of a modality-independent source of individual differences in speech recognition. *Journal of the Acoustical Society of America*, *100*(2), 1153–1162.
- Watson, D., & Gibson, E. (2004). The relationship between intonational phrasing and syntactic structure in language production. *Language and Cognitive Processes*, *19*(6), 713–755.
- Watson, D. G., Arnold, J. E., & Tanenhaus, M. K. (2008). Tic Tac TOE: Effects of predictability and importance on acoustic prominence in language production. *Cognition*, *106*(3), 1548–1557.
- Wennerstrom, A. (2011). Rich pitch. The humorous effects of deaccent and L+H* pitch accent. *Pragmatics and Cognition*, *19*(2), 310–332.
- Whittingham, M. J., Stephens, P. A., Bradbury, R. B., & Freckleton, R. P. (2006). Why do we still use stepwise modelling in ecology and behaviour? *Journal of Animal Ecology*, *75*(5), 1182–1189.
- Wightman, C. W., & Ostendorf, M. (1994). Automatic labeling of prosodic patterns. *IEEE Transactions on Speech and Audio Processing*, *2*(4), 469–481.
- Wundt, W. M. (1904-1920). *Völkerpsychologie : eine Untersuchung der Entwicklungsgesetze von Sprache, Mythos und Sitte* [Social psychology. An investigation of the laws of

Bibliography

- evolution of language, myth, and custom*]. Leipzig: Engelmann, vols. 1-5; Kröner, vols. 5-10.
- Xu, Y. (1999). Effects of tone and focus on the formation and alignment of contours. *Phonetica*, 27, 55–105.
- Yasinnik, Y., Renwick, M., & Shattuck-Hufnagel, S. (2004). The timing of speech-accompanying gestures with respect to prosody. In *Proceedings of From Sound to Sense: 50+ years of discoveries in speech communication*, (pp. 97–102). Boston, USA.
- Zeger, S. L., & Karim, M. R. (1991). Generalized linear models with random effects; a Gibbs sampling approach. *Journal of the American Statistical Association*, 86(413), 79–86.
- Zendel, B. R., Tremblay, C. D., Belleville, S., & Peretz, I. (2015). The impact of musicianship on the cortical mechanisms related to separating speech from background noise. *Journal of Cognitive Neuroscience*, 27(5), 1044–1059.
- Zimmermann, M. (2007). Contrastive focus. In C. Féry, G. Fanselow, & M. Krifka (Eds.) *Interdisciplinary Studies on Information Structure 6. The notions of information structure*, (pp. 147–159). Potsdam: Universitätsverlag Potsdam.
- Zioga, I., Di Bernardi Luft, C., & Bhattacharya, J. (2016). Musical training shapes neural responses to melodic and prosodic expectation. *Brain Research*, 1650, 267–282.
- Zubizarreta, M. L. (1998). *Prosody, focus, and word order*. Cambridge, MA: MIT Press.