Chapter 3

Multimodal analyses of audio-visual information: Some methods and issues in prosody research

Barbara Gili Fivela

University of Salento

The chapter aims to discuss some methods which have been adopted to perform multimodal analyses of audio-visual speech materials, focusing on linguistic distinctions conveyed by prosody. Attention is paid firstly to the production and, secondly, to the perception of speech prosody in its audio and visual dimensions. As for visual information, the paper discusses both articulatory gestures directly involved in the production of speech (e.g., lip gestures) and information that may be more traditionally considered, and referred to, as speech accompanying gestures (head movements and facial expressions). In any case, the main characteristics of the various methods are described thanks to specific examples found in the scientific literature, focusing mainly on Italian and some other Romance languages. The final goal is to highlight the advantages and disadvantages related to the specific methodological choices, clarifying the key aspects in order to make the reader able to choose among the various methods and offering the relevant references for a deeper understanding.

1 Introduction

In his 1995 work, David Crystal defines prosody as

> a term used in SUPRASEGMENTAL PHONETICS and PHONOLOGY to refer collectively to variations in PITCH, LOUDNESS, TEMPO and RHYTHM (Crystal 1995; capitals in the original).

Barbara Gili Fivela. 2018. Multimodal analyses of audio-visual information: Some methods and issues in prosody research. In Ingo Feldhausen, Jan Fliessbach & Maria del Mar Vanrell (eds.), *Methods in prosody: A Romance language perspective*, 83–122. Berlin: Language Science Press. DOI:10.5281/zenodo.1441339

The term "prosody" is indeed used to refer to the modulation of the aforementioned parameters with reference to units higher than phonemes in the prosodic hierarchy, e.g. syllables and phrases. In this respect, intonation, stress, tone, and, for some linguists, rhythm as well, may be regarded as prosodic features (Beccaria 1994). As the above-mentioned definition highlights, prosody is usually seen as a matter of phonetics and phonology, that is a matter of sounds, related to spoken communication. In this perspective, prosody has often been investigated as if it was unimodal, involving sound only (to give a few examples, see the contributions related both to acoustics and perception of speech since the sixties, e.g., Lehiste 1975; Lehiste & Wang 1977). However, prosody may be clearly expressed by means of the visual channel. In sign language; for instance, prosody does not pertain to the sound domain as it is expressed by facial expressions, head and body movements as well as gesture duration and tension (e.g. Nespor & Sandler 1999; Wilbur 2000; Sandler 2005). In similar cases, therefore, prosody pertains to the visual rather than the audio domain.

Even though the tendency may be to treat prosody as if it was unimodal, some investigations more easily and naturally acknowledge the multimodal character of prosody; the fact that, in spoken communication, it usually relates to both audio and visual information. For instance, Cavé et al. (1996) recorded ten subjects while answering to yes/no questions and found out that a rising-falling eyebrow movement was associated with a fundamental frequency (henceforth, F0) rise in 71% of cases, suggesting a linguistically-driven relation between eyebrow movement and intonation. Other studies have shown that linguistic information is expressed by both visual and audio information. In fact, visual information has been reported to be used to highlight prominent words in an utterance (Krahmer & Swerts 2007; Swerts & Krahmer 2008) or to give positive or negative feedback (Barkhuysen et al. 2005), and visual expressions were found to signal the end of a sentence or a speaker turn (Barkhuysen et al. 2008). Noteworthy, in various of the studies which take into account the multimodal nature of prosody, a debated issue relates to the relevance of visual vs. audio information in conveying prosody. Indeed, according to some works, audio information appears to play a crucial and major role in comparison to visual information (e.g., House 2002; Dijkstra et al. 2006; Dohen & Loevenbruck 2009; Srinivasan & Massaro 2003), while in other works the relevance of audio and visual cues seems to be more balanced, and one cue appears to be somehow related to the other one (e.g. Crespo-Sendra et al. 2013) – for details, see §3.1.

In line with a traditional view of what may be of strict interest to linguistic research, investigating prosody as if it were unimodal may be sufficient enough to shed light on the linguistic message conveyed. Indeed, felicitous communication may be just unimodal in those contexts in which either the audio or the visual signal is the only source of information (e.g., in conversations on the phone or via sign language). In general, the information in one channel is sufficient to interpret the message (e.g., it is fully included in the verbal signal with no clear added value of multimodal analyses). Nevertheless, multimodality is often exploited and it is also very powerful in communication. Actually in some cases, both unimodal and multimodal communication may take place simultaneously, as two different communication channels may be differently used at the same time, with a relevant impact on the message conveyed. For instance, Gili Fivela & Bazzanella (2014: 118–119) discuss an example in which two local contexts[1] are created, with the message (and prosody too) being conveyed in a unimodal way in one context and in a multimodal way in the other, with the result of inducing two completely different interpretations. In particular, the authors show that, in the case of a person who speaks with someone on the phone (someone who has access only to the verbal signal in a non-face-to-face conversation) and has someone else standing in front of him/her (someone who has access to both audio and visual information in a face-to-face conversation), the speaker may actually convey verbally a message to the interlocutor on the phone while, at the same time, denying the content of the message to the person standing in front, by means of visual information available only to him/her. In a similar situation, depending on the source of information available to the interlocutor (audio only, or audio-visual), then, the interpretation of the utterance changes as its "truth value" is modified. In the example given by the authors, a woman is talking on the phone with an interlocutor to whom she wants to express politeness and a positive message, while showing to another interlocutor standing in front of her, by means of mimicry and gestures, that the politeness and the content of the message expressed through the phone is false. Thus, a speaker conveys two completely different meanings, being aware of the different information available in the uni- and in the multimodal communication.

Indeed, in

> the process of understanding we do not only refer to what is said, but we also resort to a network of paralinguistic and extralinguistic means, as those expressed by changes in prosody (which intervenes with a crucial role […]), gesture, gaze, smiles, laughter, and kinetic devices, such as nodding. These

[1] Akman & Bazzanella (2003) propose the existence of both a global and a local context, the former corresponding to an a priori component (including, e.g., the participants' sociolinguistic data, their respective (and mutual) knowledge/beliefs), the latter being constructed during the interaction and concerning linguistic (that is, knowledge of the preceding and following discourse), gestural, and action levels.

> verbal and nonverbal means can function in an integrative or opposing way, both in assuming or negating the truth of the propositional content, and in upgrading or mitigating the related illocutionary force (Gili Fivela & Bazzanella 2014: 100).

Therefore, the multimodality of communication (which moreover is often available even in computer mediated communication, e.g., via Skype) cannot be denied. As a matter of fact the integration of both audio and visual information is considered here to be crucial in order to obtain a complete overview of what plays a role in both message production and interpretation. For this reason, in the following sections of this chapter the attention is focused on some methods which have been adopted in the literature on prosody to perform multimodal investigations of speech material and are related to linguistic distinctions.

However, before focusing on the core of the paper, several issues should be clarified. Firstly, when referring to multimodal communication, the intent is, quite straightforwardly, to refer to the integration of verbal and visual communication, that is a communication that takes place thanks to both the verbal signal and what we do to produce it, and the visual signal, that is what we do while producing it, which does not correspond (only) to sounds.[2] In this respect, the speech sounds and their acoustic characteristics (as well as the articulatory gestures to produce them) are clearly considered as part of the verbal channel. However, articulatory gestures necessary to produce at least some sounds, that is those for which external articulators offer information (e.g. bilabials vs. non-bilabials, rounded vowels and consonants produced or affected by lip protrusion), are visible through the visual channel, although they offer information that is tightly related to the production of the verbal signal. Finally, facial expressions, head movements and body gestures in general surely constitute a part of the visual signal that is less directly related to the mechanics of speech production and, in a sense, for this reason represent a specific added value to multimodal communication (adding on to the message interpretation as in the above-mentioned example). This differentiation within the visual information available will be considered in the following sections, where, though, the attention will be restricted to gestures involving the face and head (thus not all body gestures will be considered, e.g., no hand gestures).[3]

[2] In principle, this includes the information related to the visual context as a whole, including, but not being limited to, the speaker expressions and gestures.

[3] Given this wide view on what is relevant in the visual channel (from lip gestures needed to articulate speech to head gestures accompanying linguistic meanings), there is not one single definition of gesture that fits the discussion. Rather, the reader is referred to the definition(s) of gesture relevant within the various frameworks referred to in the parts of the paper.

Secondly, a distinction between analyses and information has to be made. Indeed, in this paper the attention is also oriented towards different types of multimodal analyses, those being the methods we use to investigate speech and prosody (e.g., intonation) as conveyed by more than one modality. In this respect "multimodal" simply indicates that more than one channel is taken into account in the analysis. However, multimodal information (differently from analysis) corresponds to the integration of information stemming from different channels or the way the coding/decoding of information is affected as it happens through/ is conveyed by different channels. Consistently, multimodal analyses and multimodal information do not always match, as it is possible to perform, for instance, multimodal analyses of sound and speech gestures that convey either unimodal or multimodal information. As for the former, it brings to mind investigations on prosody and inner articulator gestures, such as that of the tongue, in which the analysis is multimodal (it relates to sound, e.g., intonation, and visual information, e.g., eyebrow movements or even lip gestures), but the information offered to the interlocutor is unimodal as conveyed/included in one channel only (sound); as for the latter, examples are those concerning, say, prosody and facial expressions or prosody and even outer articulatory gestures, in which both the analysis and the information is multimodal (it relates to both audio and visual information).

Given these premises, in the following sections attention is concentrated on methods used for performing multimodal analyses on prosody in speech material conveying multimodal, audio-visual information, and in particular referring to linguistic distinctions. Methods will be described thanks to examples found in literature mainly on Italian and some other Romance languages. The main goal is to highlight and discuss advantages and disadvantages related to the adopted methodologies, clarifying the key aspects to allow the reader to choose from the various methods and suggesting the relevant references for their deeper understanding. The studies described also exemplify research questions which have been addressed by means of the various methods while, at the same time, offering material for discussion on advantages and disadvantages. Such discussion centers on both practical issues and on the impact of methodological choices on theoretical considerations and models that can be referred to. Attention is devoted firstly to the production of speech prosody together with articulatory gestures, head movements and facial expressions (§2) and, secondly, to the interplay of speech prosody and visual cues in perception (§3). Finally, concluding remarks complete the paper (§4).

2 Production

2.1 Introduction

Multimodal analyses of speech prosody may mainly regard the analysis of verbal speech signal including an examination of either articulatory gestures directly involved in the production of the verbal signal or gestures which accompany the production of speech. These, while not being directly physiologically related to the production of speech sounds, are, however, linked to the message conveyed. As for articulatory gestures directly involved in the production of the verbal signal, think of gestures involving the lips, as external articulators, or even tongue movements (in the latter, though, gestures may be part of a multimodal analysis but are not considered as part of a message which is interpreted multimodally). Regarding gestures that accompany the production of speech, consider eyebrow movements, as well as head position, which are not physiologically necessary for speech production, but may be related to it and therefore may offer information to the interlocutor. The way materials are collected and analyzed varies and depends on the type of data investigators want to focus on.

One of the most important choices in studying multimodal communication regards/relates to the way to elicit material to be investigated, exactly as it happens in unimodal investigations. In fact, the elicitation method influences the speech style that will be focused on and, at least to a certain extent, the data that will be collected both in quantitative and qualitative terms.[4] In investigating linguistic prosody in speech production within a multimodal perspective, the choice often regards very controlled speech styles, obtained by eliciting isolated sentences or sentences in context (e.g., short dialogues inducing the intended pragmatic interpretation on the target utterance/word), including target words or pseudowords. In fact, methods to elicit more spontaneous-like speech styles are not significantly considered in the literature on multimodal analyses of multimodal communication, even though the scientific community has been quite recently taking them into account at least for investigating unimodal communication (e.g., recordings of semi-spontaneous speech, such as Map Task (Brown et al. 1983; Anderson et al. 1991), spot-the-difference dialogues (Savy & Cutugno 2009; Pean et al. 1993) or even possibly more spontaneous speech, such as that obtained by means of the Discourse Completion Task (Blum-Kulka et al. 1989, Vanrell et al., this volume) or dialogues (e.g., Geng et al. 2013).

[4] To get an idea of the amount of change in prosodic characteristics that depend on the speech style, think, for instance, about results of comparisons of read and spontaneous speech: the latter shows more syllables produced per second and, on average, a wider F0 range (Blaauw 1995) as well as a high number of rising boundaries (Ayers 1994; Blaauw 1995).

In any case, the choice of speech style is heavily influenced by the type of data to be collected, as will be discussed in the following section with particular reference to tracking and imaging data.

2.2 Methods for data collection and analysis: some examples

In works which adopt a multimodal perspective, data collection usually involves the recording of verbal signals simultaneously with tracking or imaging data.

Tracking data are those collected by recording the position in time of specific markers. They are usually collected by means of either optotracking systems, exploiting cameras that record the infrared 3D signal reflected by markers glued on the speaker face (e.g., eyebrows, lips), or systems recording the position in time of electrodes that are placed within an appropriate electromagnetic field. In the latter, recording takes place by means of systems such as magnetometers or electromagnetic articulographs that work thanks to electrodes that may be glued both on the speaker's face and in the speaker's mouth (e.g., eyebrows, lips, tongue). Independently of the system adopted, the procedure consists of gluing markers/electrodes on the articulators to track, using three stable positions (usually behind the ears and either on the nose or, if possible, on the upper incisors) for head position normalization. In all cases, the corpus recorded is usually highly controlled, the number of repetitions recorded for each item and speaker is quite high (e.g., 7 to 10), while the number of subjects is limited (it was even one in earlier studies, but is increasing and now may reach even 10, at least in the case quite recent recording systems are adopted). Various research questions on prosody have been answered by collecting such kind of data.

For instance, Avesani et al. (2007; 2009) investigate the accent-induced articulatory strengthening, focusing on the kinematics of lip movements in the production of syllables which are variably prominent, being unstressed, stressed and nuclearly accented. They collect articulatory data by means of ELITE, an automatic optotracking movement analyzer, which allows 3D kinematic data acquisition and synchronous recording of the acoustic signal. Markers considered for the analysis are those glued on the lower and upper lip, and on both the tip of the nose and the earlobes for head position normalization. Eight repetitions are recorded of nonce-words (CVCV(C)CV, where C = [b, m]; V=[a, i]) produced by two female speakers (of two varieties of Italian). The target words are inserted in declarative sentences in short dialogues to elicit the intended interpretation, so that the penultimate syllable of the nonce-words can be unstressed, stressed or nuclearly accented in a contrastively focused constituent.

To make a slightly different example that relates to intonation, in Stella et al. (2014) articulatory differences in the alignment of the L+H* pitch accent with the lip gestures are investigated, focusing on the syllable [ma] in three different languages, that is Italian, Spanish and Catalan. In such languages, in fact, the L+H* pitch accent conveys different pragmatic functions, as it expresses a narrow-contrastive focus in the latter two languages while it is non-focal in the former; the goal is then to check if there are differences in the phasing of the acoustic rise (L+H*, produced by a rising laryngeal gesture) with the lip gestures to produce [ma]. In this work, an AG500 articulograph is used to track speech gestures (simultaneously with the acoustic signal registration), by gluing 4 sensors on the tongue, 2 on upper and lower lips (see Figure 1), 2 on upper and lower incisors and 2 behind the ears, for head movement normalization. The authors record 8 speakers in total, and ask them to produce 10 repetitions of a corpus composed by pseudowords such as [mi.ˈma.mi] and [mi.ˈma.mi.ma]. Target words are inserted in dialogues consisting of two question-answer exchanges, built in such a way that the answers including declarative sentences with non-focal or contrastive correction focus in a prenuclear position.

Under imaging data, a set of quite different techniques may be included, ranging from the video recording of speakers (e.g., her/his head, half of her/his body; Ekman & Friesen 1978) to the collection of, say, tongue imaging data during speech production (by means of ultrasound systems; Stone 2005). Of course, investigating what happens inside the mouth, as already mentioned, may be more

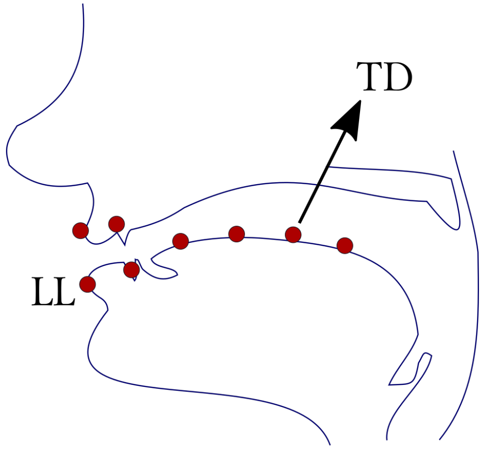

Figure 1: Sensors glued on the articulators; LL and TD stand for Lower Lip and Tongue Dorsum, respectively.

useful for a multimodal analysis of speech production rather than for an analysis of multimodal information in speech communication and, therefore, will not be further discussed in this paper. However, when considering the collection of speech data together with the video of, say, the speaker's head, the method for data collection is quite straightforward and consists of recording audio and video simultaneously, obtaining time aligned audio and video materials. Digital cameras are used for audio-video recordings, with the highest frame rate per second possible. In the case of these methods too, various research questions may be addressed.

For instance, Crespo-Sendra et al. (2013) collect audio-video material in order to investigate (also perceptually) the interaction between intonation and facial gestures in the expression of information-seeking and incredulity yes/no questions in Catalan and Dutch. They perform audio-video recordings of 5 subjects for each language thanks to a digital camera that records (25 frames per second) the upper body and face of subjects. Participants are asked to read (for ten times) "in an expressive fashion" the target sentences inserted in dialogues inducing the two intended interpretations of yes/no questions.

To offer another example, Gili Fivela (2015) also exploits the use of a digital camera to acquire both audio and video signals (the upper body and face – see Figure 2) while 5 subjects read aloud, for at least five times, short discourses aimed to simulate a Discourse Completion Task (Blum-Kulka et al. 1989, Vanrell et al., this volume).[5] In this work, facial expressions and head movements are investigated across sentence modalities, considering statements, wh-questions and exclamations with the aim to check if visual information could be more important for the message interpretation when it represented non-neutral, thus marked, information (e.g., in wh-questions conveying surprise or in exclamations vs. neutral statements).

Let's now turn to discussing methods adopted for analysis. Data analysis, as expected, changes depending on data collection, though, when the goal is to perform multimodal analysis of multimodal information related to prosody, the analysis usually concerns both the verbal and the visual modalities.

Methods used for analysing speech sounds are assessed, due to the long tradition of studies focused on both qualitative and quantitative aspects. In fact,

[5] The simulation consisted in having the subjects memorize the target sentence that was proposed within various contexts used to induce the different interpretations. This procedure was needed to create a communication context that was as natural as possible, though the sentence structure and composition could not be left to the speaker's choice. Such high control on the productions was necessary as, at a later stage of investigation, various combinations of audio and video signals had to be matched (see below).

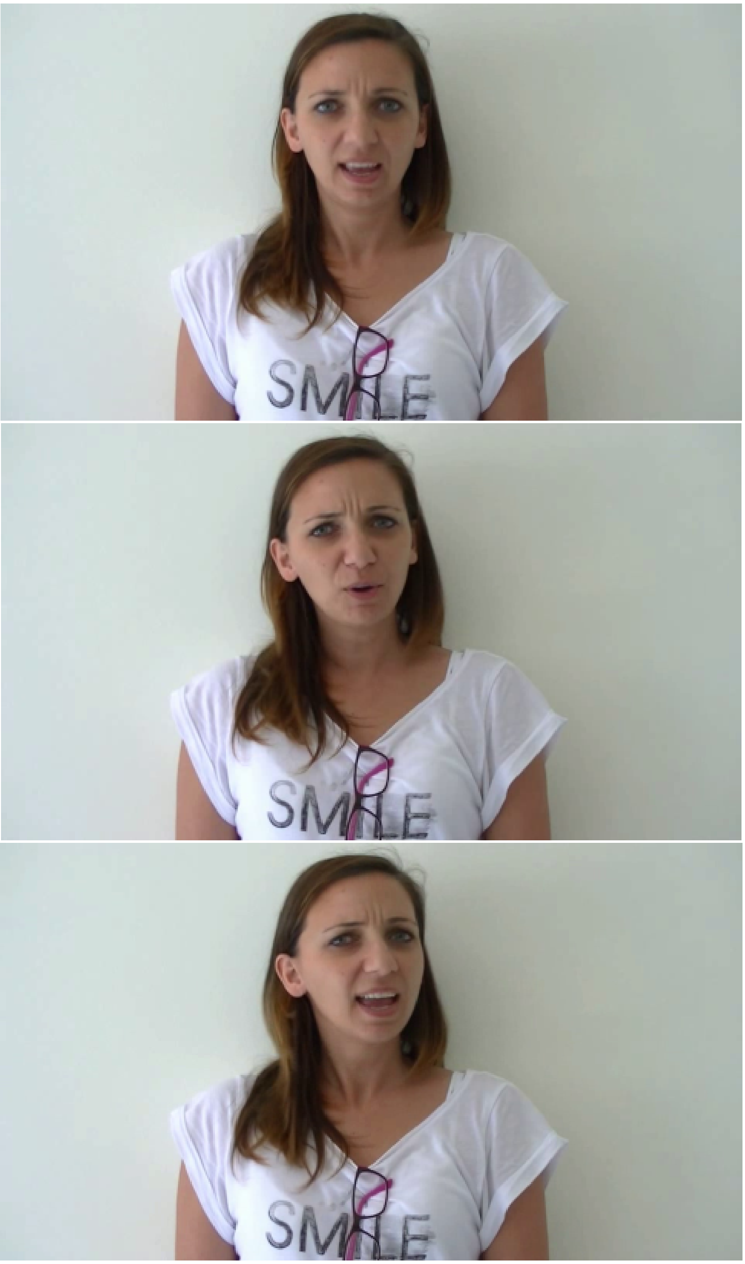

Figure 2: Examples of snapshots taken from the recording of a surprised wh-question (from the corpus used in Gili Fivela 2015); changes in visual information is clearly detectable.

since the 60s, speech prosody has been studied by performing acoustic measurements of various parameters, such as duration, F0, intensity. However, especially when linguistic information is at issue, the analysis usually involves qualitative evaluations too, that stem from examination aiming to highlight the existence of linguistic categories out of the variation of phonetic parameters. In the case of prosody, and particularly in the case of intonation, this is true, for instance, for all the works whose goal is the identification of phonological categories within the Autosegmental-Metrical theory (Pierrehumbert 1980; Ladd 1996; Beckman 1997). Methods adopted for these purposes are not particularly new to the field of Romance linguistics or even general linguistics and therefore will not be discussed in detail here.

As for visual information in the production of clearly linguistic information (e.g., prosodic focus, sentence modality) the situation is more heterogeneous. On the one hand, a relatively long tradition of studies has systematically investigated speech production and the synchronization of acoustic signal and articulatory movements/gestures. These works have focused on speech articulatory gestures as a whole, rather than on visual information, and were inspired for instance by Browman and Goldstein's proposal within the task dynamics (Browman et al. 1984; Browman & Goldstein 1985; for prosody, e.g. Edwards et al. 1991; Beckman et al. 1992 and following works on jaw movements related to prosodic structure Byrd & Saltzman 1998; 2003; see also Gili Fivela 2008 and Avesani et al. 2007, as mentioned above; for intonation, D'Imperio et al. 2007; Prieto et al. 2007; Mücke et al. 2009). Though the focus of such works is not on visual information, lip movement is quite often focused on, which may also be seen as a relevant part of visual information related to speech and speech prosody. The analysis usually regards the vertical or the horizontal movement of markers/electrodes whose position was previously recorded. In some cases, analyses relate to the position of specific electrodes (e.g. the one glued on the lower lip), while in other cases it may already be related to derived measures (e.g., a track corresponding to the lip aperture signal — i.e. to the distance between the positions recorded for the two lips — is directly taken into account). In any case, relevant landmarks are identified in the labelling phase (e.g., onset and offset of gestures on the position track, at the zero-crossings in the corresponding velocity signals; the velocity peak of gestures on the velocity track) and measures are taken of their temporal (ms) and spatial (mm) characteristics. These measures allow then to calculate other, derived, measures, such as gesture duration and displacement or gesture stiffness (as the ratio between peak velocity and displacement). Statistical analysis is then performed on these measures and usually also related to more traditional acoustic measures performed on the very same recordings (as audio was simultaneously recorded).

The usual method adopted in similar investigations may be exemplified by taking into account one of the foundational works concerning prosody. Byrd & Saltzman (1998) analyze kinematic data by three subjects to check, among other things,

> whether multiple levels of prosodic boundaries can be distinguished in the spatio-temporal patterning of articulation (Byrd & Saltzman 1998: 173).

They basically look at the articulatory correlates of final lengthening, a phenomenon which had already been found acoustically by the end of prosodic constituents (Oller 1973; Wightman et al. 1992). In order to achieve their goal, the authors record a CV sequence within which five different prosodic boundaries were realized. By means of a magnetometer system (EMMA, by Perkell et al. 1992), the authors record the horizontal and vertical position of two electrodes glued on the upper lip and the lower lip, and, after the recordings, calculate the Lip Aperture signal as corresponding to the Euclidean distance between the two lips. By means of a software dedicated to signal processing (HADES, Rubin 1995) they automatically mark (at the zero-crossings in the corresponding velocity signals) the onset/offset of the lip closing/opening movement for each of the consonants in the target sequence. On the basis of the movement onset, peak, offset, and movement peak velocity the authors calculate a number of dependent variables, such as the duration of the pre-boundary opening and post-boundary closing movement and of the transboundary interval (that is, the duration of pre-boundary opening and post-boundary closing; Byrd & Saltzman 1998: 179). Articulatory data, together with data concerning the acoustic characteristics of the sequences under investigation allow the authors to show, for instance, that three levels of prosodic boundaries may be statistically distinguished by the temporal and spatial characteristics of lip gestures which are adjacent to the boundaries (e.g., by the lengthening of the pre-boundary opening movements and mainly by the lengthening of post-boundary closing movements).

Along similar lines, though the kinematic data were acquired by means of an optotrack system, Avesani et al. (2007; 2009) analyze data on accent-induced articulatory strengthening, as already mentioned. They label and measure acoustic data by means of Praat (Boersma & Weenink 2017) while articulatory data are analyzed by means of *Interface* (Tisato et al. 2005). To offer more details on the analysis phase, it is worth to recall that, firstly, the authors check each utterance for the realization of pitch accents on the target syllables; secondly, they center their attention on the lip aperture and take spatial (mm) and temporal (ms) measures of the onset, the target and the peak velocity of both the opening and the closing gestures. They then calculate various dependent measures, such as gesture duration, its displacement, peak velocity, time-to-peak velocity (which is the duration of the acceleration phase) and gesture stiffness (as the ratio between peak velocity and displacement). By statistically analyzing acoustic and articulatory data the authors show, among other things, that in the production of one speaker:

> compared to stressed, unstressed syllables show shorter acoustic and articulatory duration, smaller displacement, equal peak velocity, shorter TTP and higher stiffness for both opening and closing gestures (Avesani et al. 2007: 983).

The authors observe that, for the same speaker, a dynamic mechanism of linear rescaling seems to take place when accented syllables rather than stressed syllables were taken into account. However, for the other speaker different dynamics seem to characterize gestures when considering different levels of prominence, that is linear rescaling does not seem to necessarily take place.

Other studies adopt very similar methods to investigate intonation, and in particular the intergestural coordination between laryngeal and supralaryngeal gestures. In fact, tonal alignment may be investigated as a matter of coordination between gestures to produce F0 modulation and gestures to produce segments, syllables or other units, that is as a coupling of tonal and oral gestures (D'Imperio 2002; Ladd 2008; Prieto & Torreira 2007). For instance, the investigation on Italian, Catalan and Spanish by Stella et al. (2014) mentioned above may help in exemplifying the methods exploited in the analysis. To perform their analysis the authors label tonal targets and segmental boundaries by means of a Praat script and label the articulatory data by means of a MatLab graphical user interface-based software for multimodal articulatory data inspection and analysis, MAYDAY (Sigona et al. 2015). The Praat TextGrids, containing segmental and tonal labels, are imported in MAYDAY, where the articulatory data are then semi-automatically labelled, marking the onset, offset and peak velocity of each opening and closing gesture realized to produce the CV target sequence in the target words. Latencies between tonal targets and articulatory landmarks are then computed by means of a MatLab script. The method described so far may be considered to be quite traditional (apart from the choice in the software and scripts to label and measure the articulatory data, which pretty much varies depending on the laboratory and research group). What is worth mentioning as for the methodological choices is that, rather than just analyzing measures by means of statistics to identify significant differences, in this work a MatLab script is also implemented and used to obtain graphical plots of the alignment patterns in order to visually inspect the timing relations between tonal and oral gestures. The graphical plot of the alignment patterns is based on the mean temporal values for the 10 repetitions analyzed for each item (that is for each speaker, each syllable and word stress type). As shown in Figure 3, the alignment plot is formed by 4 tiers (Segments, Tones, Lower Lip and Tongue Dorsum), showing the temporal values of articulatory, segmental and tonal landmarks normalized at the onset of the target word. Visual inspection of the alignment patterns and two-way ANOVA allow the authors to highlight the quite stable alignment of tonal targets with articulatory landmarks.

To conclude, methods used in the analysis of the interplay between gestures for producing speech and acoustic features of speech are quite well known, though there are not many works adopting such methods to deal with prosody in Romance languages.

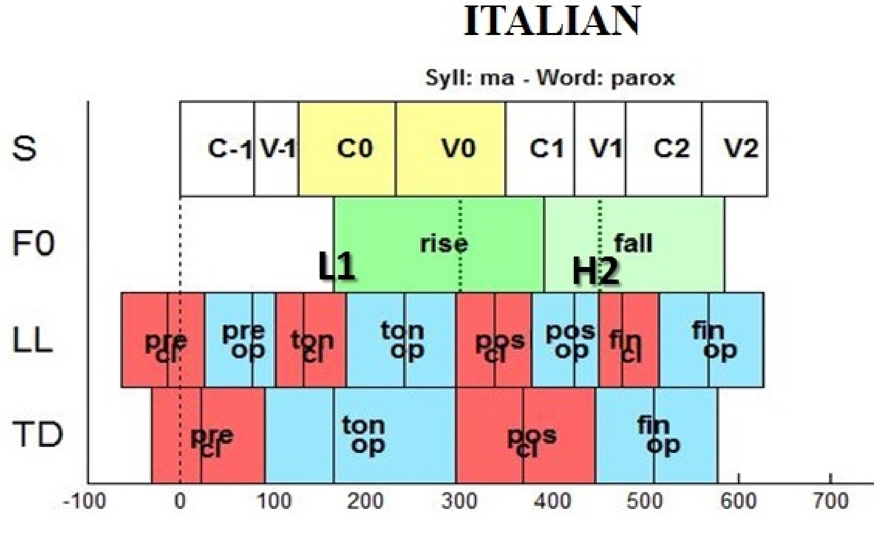

Figure 3: Patterns of alignment of L+H* productions in Italian (see also Stella et al. 2014); the 4 tiers in the alignment plot (from top to bottom, Segments, Tonal events, Lower Lip and Tongue Dorsum) show the temporal values of articulatory, segmental and tonal landmarks normalized at the onset of the target word.

On the other hand, interest in the production of visual information which is not directly related to the articulation of speech units (that is, it is not physiologically related to the production of speech sounds, but rather to the message conveyed) is even more rare in investigations related to the linguistic message. In fact, methods for investigating visual information in communication are assessed, but intensive studies on the role of such information in the coding and decoding of linguistic prosody and message have definitely not been a top priority (see §1).

In the early 80s a system to code facial expressions and head movements was proposed by Ekman (Ekman & Friesen 1978; Ekman et al. 2002) and it is still used nowadays. The system is called the *Facial Action Coding System* (FACS) and it identifies *Action Units* (AUs), corresponding to the activation of one or more muscles producing a change in the facial appearance. AUs are identified by numbers (letters+numbers in some cases) and names (e.g., AU 4 – *Brow Lowerer*): the former are basically arbitrary and their association to names helps in learning the coding system, even though the actual coding by experienced coders refer to numbers rather than names. The coding is basically performed by observing movements of the skin, specific parts of the face (to start with the coder's own face in the learning phase) and the head. Indeed, these movements allow to identify the appropriate AU that took place and to code it appropriately together with a score of its intensity. Indeed, AUs and their combinations can be described also in terms of intensity levels (from A, that is "trace", to E, that corresponds to "maximum"). For instance, the images reported in Figure 4, starting from top to bottom may be labelled as AU4 *Brow Lowerer*, AU4+7 *Brow Lowerer + Lid Tightener* and AU2 - *Outer Brow Raiser*; their intensity level may be labelled as *C- Marked or Pronounced.*

The coding system has been used even quite recently to label visual information related to prosodic information that clearly plays a linguistic role. For instance, in both the previously mentioned works by Crespo-Sendra et al. (2013) and Gili Fivela (2015), after the recording of audio-video material (to be then used in perception experiments) a coding of the main patterns of facial expressions observed during the target utterances (usually on the nuclear accented word) was given. For instance, Crespo-Sendra et al. (2013: 6) report that

> for information-seeking interpretations the most common facial expression consisted of a combination of action units AU1 + 2 (Inner and Outer Brow Raisers) and head movement M59 (Head Down and Head Up). For incredulity question interpretations, the most common pattern was a combination of AU4 (Brow Lowered), M59 + 58 (Head Down and Head Back) and squinting of the eyes. (Crespo-Sendra et al. 2013: 6)

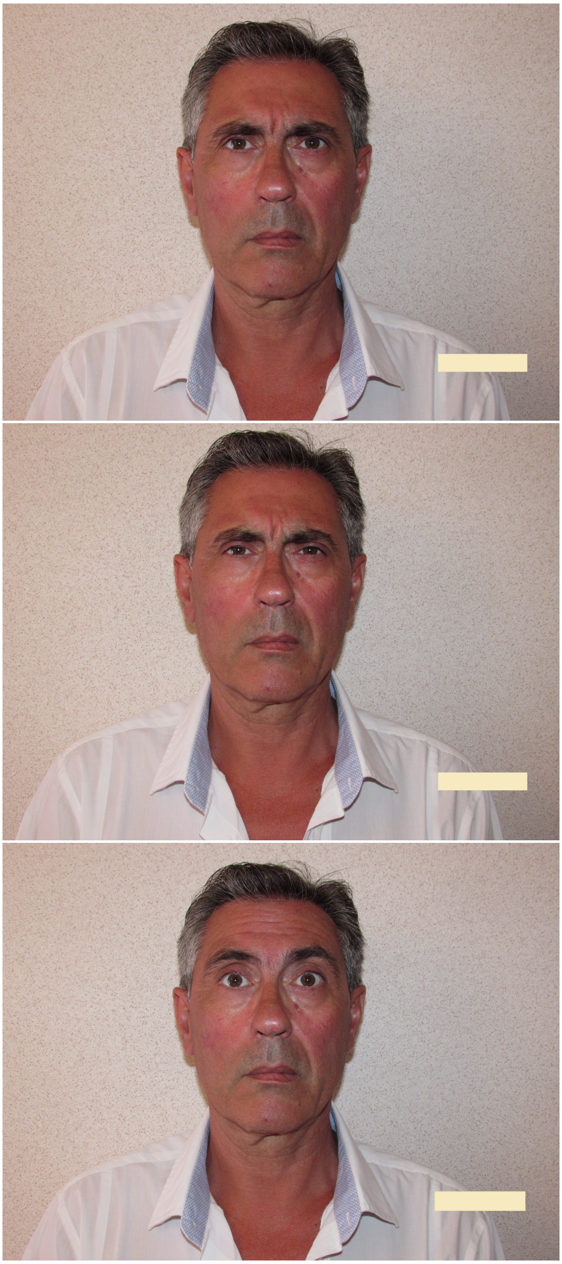

Figure 4: Examples of coding of Action Units involving Brows: AU4 Brow Lowerer (first picture) or AU4+7 Brow Lowerer + Lid Tightener (second picture) and AU2 - Outer Brow Raiser (third picture).

On the contrary, in her work on statements, wh-questions conveying surprise and exclamations, Gili Fivela (2015) analyzes all the recorded audio-video sequences and labels the following main patterns in terms of FACS :

- for statements, AU0 - *Neutral face* and AU M69 – *Head and/or Eyes Look at Other* or M59 - *Head Shake Up and Down*
- for wh-questions (surprise and positive attitude), AU4+7 - *Brow Lowerer + Lid Tightener* or AU2 - *Outer Brow Raiser* and M60 - *Head Shake Side to Side*
- in exclamations (positive attitude), AU2 - *Outer Brow Raiser* or AU4 - *Brow Lowerer* and M59 - *Head Shake Up and Down.*

Results show a clear difference in eyebrow and lid gestures when comparing neutral statements and other sentence modalities, while wh-questions expressing surprise and exclamations (both underlying positive attitude) show eyebrow and lid gestures which are not always easily distinguishable (e.g., eyebrows rising or lowering usually take place in both questions and exclamations produced with a positive attitude, possibly with *Lid tightening* in questions). On the other hand, results show that head movements seem to be similar in statements and exclamations *(with head shaking up or down),* while wh-questions differ more clearly (because usually accompanied by *head shaking side to side*).

Therefore, the analysis of visual information is performed with reference to a well-known coding system, with the aim of finding a correlation between prosodic sound features and visual prosody in expressing linguistic information. However, take note that the FACS coding system would require more than one coder/ transcriber to analyze the data and, moreover, more than one coder/transcriber which was officially trained in using the FACS coding system (for which inter-transcriber agreement thresholds are known too). To the author's knowledge, this methodological procedure, in particular as for the official training, has not been really followed in works on linguistic prosody, probably due to practical reasons. However, as for the number of transcribers, the situation brings to mind what is required for the coding of intonation patterns within the ToBI system or with the coding of Map-Task dialogues, for which having more than one transcriber would be methodologically correct and for which, not surprisingly, inter-transcriber agreement thresholds have also been proposed in the literature (Silverman et al. 1992; Beckman 1997; Isard & Carletta 1995).

2.3 Advantages and disadvantages of the methods presented

As for the use of optotracking systems, magnetometers or electromagnetic articulographs, the advantages, of course, relate to the chance of observing the articulatory correlates of prosody (e.g., lengthening or strengthening phenomena) together with the timing of other articulatory and acoustic events (e.g. the timing of F0 peaks). This allows investigators to propose and refine models of gestural dynamics related to linguistically relevant prosodic events and, the other way around, to consider linguistically relevant prosodic events as related to gestural dynamics too.

For instance, thanks to the articulatory investigation Byrd & Saltzman (1998) showed that various levels of prosodic boundaries are distinguished by the temporal and spatial characteristics of articulatory gestures adjacent to the boundaries. In such a model, which is further developed in a following paper (Byrd & Saltzman 2003), they propose that this lengthening is related to a specific prosodic gesture that regulates the duration of gestures at prosodic boundaries. Mentioning another work described above, it is the sum of acoustic and articulatory data that allows Avesani et al. (2007) to illustrate that syllable and vowel prominence are somehow directly proportional to the length, the velocity and the displacement of lip closing gestures. Furthermore, these data allow the authors to relate their results to a mass-spring gestural model (Browman & Goldstein 1995; Fowler 1995; Saltzman 1995; Saltzman & Munhall 1989), arguing that, at least for one out of the two speakers considered, results can be accounted for by a single mechanism of linear rescaling. Thanks to both articulatory and acoustic data, Stella et al. (2014) show that in the three languages they considered the investigated acoustic tonal targets have a quite stable alignment with articulatory gestures. Moreover, they argue that results may be related to the *Coupled Oscillator Model* of speech production (Goldstein et al. 2009), and that the rising tonal gesture would be in anti-phase relation with both the consonantal and the vocalic gesture.

Disadvantages relate to the heaviness of data collection, which has an impact on both the number of speakers who are usually recorded and the complexity of data analysis; these aspects, of course, affect the whole experimental design, as they orient the choice of speech style and corpus to be recorded. These first observations especially apply to data collection by means of articulographs, though the number of speakers considered in the studies has been gradually increasing over years. Moreover, the time required on average for the set-up, needed for gluing the sensors onto the subjects, both outside and inside the mouth, and the quite frequent event of a detachment of the glued sensor during data collection, make the whole recording phase time-consuming and challenging for both subjects and experimenter. In this respect, recording by means of diverse systems has various consequences as big differences may regard different versions of the same system (e.g., AG500 is far more instable than the new AG501 – Stella et al. 2012; 2013 – and therefore it is obviously more difficult to record bigger corpora by means of the former than the latter). Whatever the system is, it is obviously true that recording articulatory data is more challenging than acquiring acoustic data only (sometimes even because of the difficulties in recruiting subjects willing to have electrodes glued, say, in their mouth).

Moreover, data collected in articulatory studies usually correspond to very controlled speech material. Though, especially now that more stable machines are available, dialogical speech involving two speakers is collected by means of two systems at the same time (e.g., Geng et al. 2013), in most cases, data acquisition consists in exploiting one system and collecting very controlled, read speech data (e.g., see experiments described in the previous section). In fact, a single coil/articulator trajectory is extremely sensitive to the segmental environment not only in the sense that it may be modified by co-articulation, as expected and as observed for segments in acoustic data. The relevant point here is that it is the single coil/articulator to be investigated (e.g. lower lip, tongue dorsum) and in order to be able to observe its trajectories it is necessary to ensure the presence of significant (detectable) gestures from and to not adjacent segments. In this respect, segmental contexts that would not be problematic in acoustic investigations (where reaching of the set of expected articulatory targets in the sequence may ensure the necessary acoustic information) are in fact very problematic in articulatory investigations. For instance, given a pseudoword such as [mimi], thus including bilabial consonants produced by means of lip and jaw gestures, the tongue dorsum position for [i] will only slightly vary, and the risk is to be unable to unambiguously detect a significant tongue gesture in the [i]-to-[i] cycle; thus, an [a]-to-[i] cycle would be preferable such as in [mami] or [mima]. Of course, these constraints do not have an impact only on investigations of segments. As exemplified in the experiments described in the previous sections, they are relevant in prosodic investigations too, where reference to segments is usually needed and target words are usually chosen in order to satisfy the above-mentioned basic requirements.

Finally, labelling, measurements and data analysis relate to various sensors (e.g., tongue dorsum for vowel articulation and lower lip for bilabial consonant) and axes (e.g. the z-axis for vertical movement and the x-axis for front-back horizontal movement). This means that a specific software is needed for labelling and measurements (e.g., HADES mentioned in relation to Byrd & Saltzman’s 1998 paper, *Interface* mentioned in relation to Avesani et al. 2007; 2009 works, or MAYDAY, which was developed at CRIL for dealing with both kinematic data and ultrasound images, Sigona et al. 2015); moreover, it means that the amount of data is not easy to manage, especially when it also has to be related to data on acoustic prosody, e.g. to F0 changes.

Turning to the imaging techniques, they also imply advantages and disadvantages. Gaining a wider view of the interaction between audio and visual information may shed light on relevant factors that affect linguistic meaning and that are usually not taken into account in linguistic investigations. For instance, Gili Fivela (2015) reports a clear difference in eyebrow and lid gestures when comparing neutral statements and the other sentence modalities considered, while underlining a similarity in the head movements observed during statements and exclamations. These observations may actually be of help when considering the phonological coding of intonational events. For instance, wh-questions are quite often found to be phonologically identical to statements (e.g., various contributions in Frota & Prieto 2015, starting from the paper on Italian, i.e. Gili Fivela et al. 2015). Multimodal investigations may show that the adoption of the same intonational pattern in statements and wh-questions could be problematic because of a number of other cues that speakers may use to distinguish statements from questions, among which visual cues could be considered besides, say, lexical and syntactic ones. That is, in the long run, a wider perspective in investigating speech may offer hints on the impact of visual information on the variation observed in speech in general, as for both pattern choice and phonetic implementation.

As for the disadvantages, the use of cameras for acquiring both audio and video may imply a loss in the acoustic signal quality. The choice is then taken to be more appropriate when no accurate and extensive acoustic measurements or manipulations are performed. Another possible disadvantage may relate to criteria for subject selection, as some of the subjects, especially those who may already have troubles in immerging themselves in the given context during audio-recordings, may be even more clumsy if they know that video-recording is going on.

3 Perception

3.1 Introduction

Investigations regarding the integration of audio-visual information in the perception of prosody have been strongly influenced by works on the McGurk effect, that is, on the integration of visual and auditory information which are not always consistent.

In their 1976 work, McGurk & MacDonald asked their subjects to judge stimuli corresponding to the production of syllables [ba], [da], [ga], playing through a talking head both stimuli in which either audio or video was available and stimuli in which both audio and video were available, though they were not always congruent (that is, for instance, both the audio and the visual information corresponded to the production of [ba] or the audio corresponded to [ba] while the video showed the lip movement for [ga]). Of course, attention was paid to the realization of stimuli that seemed as natural as possible, and indeed, as the authors stated,

> Dubbing was carried out so as to ensure, within the temporal constraints of telerecording equipment, that there was auditory-visual coincidence of the release of the consonant in the first syllable of each utterance (McGurk & MacDonald 1976: 746.)

In particular, results of perception of stimuli in which the audio corresponded to [ba] while the lip movement was that corresponding to [ga], showed that listeners reported hearing [da]; moreover, when subjects were presented the audio for [ga] on to the lip movement for [ba], apart from [da], they mainly reported hearing [gabga], [bagba], [baga] or [gaba]. The authors argued that in the [ba]-audio/[ga]-video condition the acoustics for [ba] had features shared with [da] but not with [ga], that the visual information was consistent with both [ga] and [da] and that, therefore, subjects were sensitive to the common information in both modalities.

The influence of visual information on the perception of audio information reported by McGurk & MacDonald represented a milestone in the investigation of multimodal perception, with clear methodological and theoretical impacts. Such impacts are considered in the following sections as for their influence on the investigation of prosody in more recent studies, that basically started from the end of the 90s (see Lansing & McConkie 1999 on the identification of statements vs. questions on the basis of visual cues in the upper facial regions and the observation that the recognition of prosodic information from visual cues alone was more difficult than that of auditory cues).

Before addressing the specific methods adopted in the case of multimodal analyses, it is worth recalling that the focus here is the perception of linguistic information conveyed by prosody: methods to unimodally investigate this issue are quite well-known and are also used for multimodal investigations. For instance, identification and discrimination tests are used in checking for the existence of categorical perception which, on the basis of the perception of segments, and consonants in particular (Liberman et al. 1957; see the contradicting results for vowels by Fry et al. 1962), has been often taken to be a property of phonological (linguistic) units. In particular, specific characteristics are often expected in the identification of linguistically relevant sounds in that, in a traditional categorical perception paradigm such as the one proposed for consonants by Liberman and colleagues in 1957, given a continuum of stimuli, when subjects are asked to identify the linguistic category they belong to, results are expected to be S-shaped, with a sharp switch from the perception of a category to the perception of another one; when subjects are asked to discriminate the same stimuli, that is they are given pairs or triplets of those stimuli and asked to judge whether some of them are equal or not, they are expected to be more sensitive to differences across categories, that is difference between intermediate stimuli in an S-shaped plot.[6] The existence of categorical perception has been investigated with respect to intonation categories too, adopting the same methods and formulating the same hypothesis, but reaching quite contradicting results which are more in line with those obtained for the perception of vowels (e.g., see the contradictory results reported by Vanrell 2006; Schneider et al. 2006 and Niebuhr & Kohler 2004, and the discussion in the latter). A discussion of methods to unimodally investigate prosody and intonation, and for instance to design identification and discrimination tests is out of the scope of the present paper (but see, for instance, Gussenhoven 1999; 2004; Gili Fivela 2008; Prieto 2012) However, it is worth remembering that at least identification tests are often used in investigating multimodal perception too (see the next section) and that a possible distinction drawn among the various methods used to investigate the perception of prosody may be useful to understand the criteria of selection of methods considered here. In particular, as proposed by Gili Fivela (2008), it is possible to distinguish methods for collecting subject's metalinguistic judgements and procedures for directly recording speaker's response and action taking. Among the former, methods are included requiring judgements on perceptual equivalence of stimuli, on successful imita-

[6] Of course these expectations are in line with a quantal theory of speech (Stevens 1972; 1989; see also Stevens & Keyser 2010), according to which categories correspond to quantal regions, clearly different from each other and whose members show acoustic and auditory characteristics which are quite stable, despite changes in articulatory settings. However, it is worth recalling that, following works on natural categories and their corresponding semantic categories (Rosch 1975: 193; see also e.g. Berlin & Paul 1969) showing that members of a category do not necessarily share an equal degree of membership, some works on segmental phonological categories (Kuhl 1991) and on intonation categories too (Schneider et al. 2006; 2009; Gili Fivela 2012) addressed issues concerning the presence of prototypes or best examples, assuming the existence of non-homogenous categories - including prototypes - and the possibility to perceive differences in meaning or shades of meaning within a category (Gili Fivela 2012).

tion, asking for prominence judgements, for semantic differences, semantic scaling, goodness rating, matching, as well as in identification and discrimination in categorical perception paradigms; among the latter, methods are included asking for imitating stimuli, collecting eye tracking data, asking to perform games using audio stimuli and, in general, methods including reaction time measurements (for discussion, Gili Fivela 2008).

For space limits, in what follows only some methods relying on subjects' metalinguistic judgements are basically referred to (e.g., investigations on neurophysiological correlates of multimodal perception are not discussed).

3.2 Methods for data collection and analysis: some examples

Data collection usually involves base stimuli including both audio and video, though this information may either be natural (audio and video taken from recording of speaker's production) or synthetic (audio corresponding to synthetic speech and video relating to computer-animated heads, that is talking heads). Synthetic stimuli are necessarily used when continua are investigated and need to be judged by speakers. In these cases, both audio and video continua (typically representing the shift between two categories) may be created and synchronized with each other or one continuum may be created, e.g., the audio one, and synchronized with a sort of neutral condition on the other channel, e.g., regarding visual information. In other cases, audio-video natural recordings are used, and the manipulation usually aims at crossing audio and visual conditions, rather than at realizing continua of changes. In these cases, the audio and video signals in the original recordings are separated via software, offering audio files and video clips that can be used as stimuli in audio-only and video-only tasks and that can also be crossed to obtain incongruent audio-visual stimuli, usually by creating all the possible combinations of audio and video cues.

A check on the relevance of audio and visual information is often carried out in investigations, with audio-only and video-only stimuli included in perception tests. However, in line with the traditional testing of the McGurk effect, the experimental procedure often also includes an explicit check for the audio-visual integration, by means of audio-visual stimuli obtained by matching congruent and non-congruent audio and video information.

In all cases, audio-visual stimuli are created by paying specific attention to the audio-visual information synchronization, to create stimuli that are as natural as possible and that are free of artefacts. Short pre-tests may indeed be used to check for the quality of stimuli. Stimuli are presented to subjects in random order, and usually in different blocks, and subjects are typically asked to perform an identification test and to judge whether stimuli are instances of one or another category. Reaction times in answering are measured in some investigations, and subjects may also be asked to score by means of Likert scales how confident they were when answering or how much they liked the specific item with respect to the category it was judged to belong to. Answers to (not manipulated) original recordings can be taken as control for every single subject or a control task may be included in the design, to check for subject comprehension of the task and stimuli.

For instance, House (2002) investigates intonational cues and visual facial cues to the interrogative and statement mode in Swedish. In a first experiment, he manipulates the acoustic information only, creating two sets of six stimuli in which the focal accent peak is shifted and two F0 ranges for the focal accent are considered. As for the visual information, no head, eye or eyebrow movement is visible in the talking head presented to 11 subjects. On the basis of audio-visual stimuli, created by paying specific attention to the audio-visual synchronization, subjects are asked to judge whether the speaker intended to produce a statement or ask a question, and to mark on a 1-to-5 scale how much confident they are in their choice. However, in a second experiment, involving 27 subjects, the author uses the same audio stimuli, pairing them, in two different sets, with the movement configurations conveying either an interrogative (slow up-down head nod and eyebrow lowering) or a declarative mode (a smile throughout the whole utterance, a short up and down head nod and eye narrowing). The author can then demonstrate that the addition of the facial cues reinforces the information given by declarative intonation and inhibits that by the interrogative intonation: basically, the interrogative face introduces more confusion to the perception of the stimuli and, subjects are less confident than when judging audio with no changing visual information.

Srinivasan & Massaro (2003) analyze the perception of echoic questions and statements in English, presenting subjects with an auditory continuum that was crossed with a visual continuum, using synthetic speech and a talking head. In a first experiment, the authors present subjects with statement/question pairs in order to identify the pair which was best discriminated and used the acoustic and visual parameters of that pair as prototypical in order to synthesize the stimuli to be used for investigating audio-visual integration. They used *Wavesurfer* (Sjolander & Beskow 1999) for investigating the acoustic parameters and a speech software tool called *MarkupGUI* (Wouters et al. 1999) to modify the acoustic (pitch contour, amplitude, duration) and visual (eyebrow, head tilt) parameters. In the second experiment, sixteen subjects evaluate stimuli (4 sentence pairs, auditorily, visually and bimodaly), judging each of the conditions 16 times in two sessions (8 times per session). Finally, in a third experiment, the visual and prosodic cues previously exploited are considered as useful to create synthetic versions of an ideal statement and an ideal question and are then varied independently of one another. This way a five step series is created so that it

> becomes more question-like with changing pitch contour (of the entire sentence), and increasing amplitude and duration (of the final syllable). The visual continuum becomes more question-like with increasing eyebrow raise and head tilt (Srinivasan & Massaro 2003: 9).

Forty-three subjects judged the stimuli (8 repetitions) realized by means of the 'Baldi' synthetic talking head and the Festival synthetic speech. The authors report strong individual differences in the perception of auditory or visual cues and in general a stronger relevance of auditory cues (results were replicated in a follow-up experiment in which either the visual cues were doubled in magnitude or the auditory cues were more ambiguous, narrowing the range of variation in the statement-question continuum)

Turning to Romance languages, more recently, Borràs-Comes et al. (2011) describe two perception experiments in which stimuli, represented by manipulated speech and/or video signal, are used to test the integration of audio and visual information and, in particular, the interaction of intonational and gestural information in the distinction between counter-expectational questions and narrow focus statements; a second goal is to identify the facial gestures conveying the counter-expectation interpretation. To reach the first goal, the authors use an acoustic continuum representing the shift, in Catalan, from a typical narrow focus statement to a typical counter-expectational question (which are both realized with a rising pitch accent followed by a low boundary tone); as for visual information, a continuum of facial gestures is created by means of a 3D animated character, tuning its movements in order to represent different levels of activation of an incredulity expression. In a second experiment, subjects judge stimuli composed by video information only, corresponding to animated sequences in which the same 3D character conveys incredulity in 4 different levels of activation by means of the three main gestures involved, that is brow furrowing, eyelid closure and backward head movement, in all possible combinations. In both experiments stimuli are presented in random order by means of the software E-prime to eighteen Catalan listeners, who judge 5 blocks of stimuli. The subjects have to express their preference as for the interpretation of the utterances and the software also collects their response times apart from the response frequencies. As for the interaction of audio and visual information, results described by Borràs-Comes et al. (2011) show that the impact of intonation decreases as the visual counter-expectation interpretation information is clearer. The relevance of both audio and visual information is shown by reaction time measurements, as intonation has a great impact on them but it also interacts with gestures. However, as the second experiment shows, brow furrowing is crucial in distinguishing counter-expectation questions from narrow focus statements when dependent on facial gesture information, but subjects also rely on the other visual features (that appear to be given a specific degree of importance: brow furrowing > backward head movement > eyelid closure).

In terms of methods adopted to collected perception data, audio-video recordings are also used in the literature, rather than talking heads, together with a manipulation solely aimed at crossing audio and visual conditions rather than at realizing continua of changes.

To propose some examples, Crespo-Sendra et al. (2013), as already mentioned, record audio-visual material in order to create stimuli for a perception experiment. The final aim is to compare the interaction between intonation and facial gestures in the expression of information-seeking and incredulity yes/no questions in Catalan and Dutch. The authors check for the audio-visual integration by means of audio-video stimuli, and for the contribution of both audio and video by means of audio-only and video-only stimuli. The audio and video signals in the original recordings are separated (by means of the software Adobe Premiere), the audio files and the video clips are then used as stimuli in the audio-only and video-only tasks respectively. As for the audio-video task, original recordings are used as congruent stimuli, while non-congruent stimuli are obtained by manipulating the audio-video signals (with the above-mentioned software). Manipulation consists in matching, for each speaker, all the possible combinations of audio and video cues for the various interpretations (e.g., neutral face-incredulous intonation and incredulous face-neutral intonation). Once a pre-test of the material ensures their naturalness and lack of artefacts, the tests can take place (each preceded by a training phase). Crespo-Sendra et al. (2013) ask their subjects to perform the video-only and audio-only test in a different order, and both before the audio-visual task, which is also preceded by a short documentary projection to avoid possible learning effects. In addition, they have a short final control task to confirm that the 10 audio-only and video-only stimuli (by a new speaker) are unambiguously interpreted by participants. All tasks are run by means of E-Prime. Thus, given a stimulus, subjects have to choose between a neutral and an incredulous information seeking question. The authors find that, in both languages, visual cues have a stronger impact than auditory cues to induce correct identification of incredulity in questions. However, languages differ as for the weight given to the cues. Indeed, as audio-video stimuli show, Catalan listeners give more weight to facial cues than Dutch listeners.

As a final example, Gili Fivela (2015), as mentioned in §2.2, investigates facial expressions across sentence modalities, considering wh-questions, statements and exclamations in Italian. Similarly to Crespo-Sendra et al. (2013), the author checks for the audio-visual integration by means of audio-only, video-only and audio-visual stimuli, including both congruent and incongruent stimuli. The procedure followed is very similar, apart from the fact that the separation of audio and video channel is performed by means of a public domain software, VirtualDub, a simple break is taken before the audio-visual task and the answers to (not manipulated) original recordings are taken as control for every single subject (no final control task is included in the design). The entire experiment is run by means of the software Presentation and subjects are asked both to choose between three options, which is a statement, a question and an exclamation, and to rate on a 7-point Likert scale the negative-positive attitude of the speaker. The analysis of subject answers in favour of the three given options shows a fairly articulated picture and the lack of a systematic positive influence of video over audio or vice versa. In particular, video information related to neutral statements does not interfere with audio information; on the contrary, video information regarding questions, and, though to a lesser extent, that related to exclamations affect the interpretation of the audio information on neutral statements.

3.3 Advantages and disadvantages of the methods presented

The main advantage of the methods used to investigate the perception of multimodal information is considered here to be, of course, the chance of observing both the audio and the visual correlates of prosody, and the possibility of understanding how they are integrated. Moreover, some methodological choices allow to do so even in the case of artificial continua of variation. All in all, these methods allow to investigate the communication of prosody as the multimodal phenomenon it usually is. However, results reported in the literature so far are quite composite and much work still needs to be done to really understand the issue.

For instance, it was the investigation of audio-visual integration that allowed House (2002) to show that the addition of facial cues reinforced the information offered by declarative intonation only, while it inhibited the information related to interrogative intonation (as the interrogative face introduced confusion to the perception of the stimuli). Along quite similar lines, by investigating both audio and visual information Borràs-Comes et al. (2011) could show that the impact of intonation decreased as the visual (counter-expectation interpretation) information was clearer, while Srinivasan & Massaro (2003) could report a stronger relevance of auditory cues, apart from strong individual differences in the perception of auditory or visual cues.

Additionally, similar investigations can specifically emphasize the relationship between the quantity and quality of information in audio (in terms of phonetic and phonological information available) and in video and their role in audio-visual integration. For instance, Crespo-Sendra et al. (2013) found that, in both Catalan and Dutch, visual cues have a stronger impact than auditory cues to induce correct identification of incredulity in questions, though Catalan listeners give more importance to visual cues than Dutch listeners, probably because of the more subtle distinction due to acoustic information in Catalan (pitch range difference) with respect to the information available in Dutch (where a different sequence of tonal events, that is a different set of phonological categories, characterize the contours of the utterances under investigation). Nevertheless, as Gili Fivela (2015) argues, the picture on the audio-visual integration of information is quite articulated, and this may explain the lack of consistent results on a systematic positive influence of video over audio or vice versa. In particular, results on Italian show that visual information on surprised questions and exclamations affect the interpretation of audio information on neutral statements, but not the other way around, independently of the information available on the audio channel (i.e. on the phonological pattern which was implemented). Thus, marked facial expressions and head movements (in her work associated to questions and exclamations) seem to affect the interpretation of utterances which are not associated to marked information on the same channel (in her work, neutral statements), rather than to affect information which is ambiguous in the other channel, that is sound.

Not surprisingly, then, these investigations possibly support different theories of speech perception, such as the single channel model (SCM), the weighted averaging model (WTAV) and the fuzzy logical model of perception (FLMP) (for a discussion, see Massaro 1989; Massaro & Cohen 1993; Srinivasan & Massaro 2003).[7] A discussion of the models is out of the scope of the present paper. How-

[7] Briefly, according to the SCM only one of the auditory and visual channels of information is functional on any given bimodal input, that is, SCM is a non-integration model according to which a single channel of information can be processed at any one time. However, according

ever, it is worth mentioning here that Srinivasan & Massaro (2003: 20) found that the FLMP was not significantly better than the WTAV/SCM models, while Borràs-Comes et al. (2011) do not assume a clear position as for the model (WTAV or FLMP) that is better supported by their data, though they suggest that their results could be consistent with FLMP (especially for the relevance of both audio and visual information shown by reaction time results). Along similar lines, Crespo-Sendra et al. (2013) argue that their results agree with the FLMP, as an ambiguous or weaker cue in one modality seems to enhance the role of the other modality. However, Gili Fivela (2015) observes that her results seem to support the idea that it is not only the relation of information available in the channels that plays a role (e.g. the visual information and the phonological pattern implemented and conveyed by means of the audio channel), but also the balancing of information within the same channel. Indeed, the visual information in questions (and partly in exclamations) affects the audio interpretation of statements, but not the other way around (visual in statements does not equally affect audio in questions). Thus, investigating the perception of multimodal prosodic information is still needed to really answer the question concerning the role of audio and visual information and the way they are integrated. Luckily, this can be done also by resorting to quite a high number of subjects for each perception experiment, which makes results more solid and generalizable.

As for disadvantages related to the methods described here, it is important to underline that they correspond to difficulties rather than to real disadvantages. As a matter of fact, one main difficulty is detected in data collection, mainly because of the need to ensure naturalness in the stimuli used for perception experiments. This aspect brings us back to difficulties in collecting the speech material to be used to create stimuli, that is in eliciting as spontaneous and as natural sounding speech as possible (see §2.3). However, the naturalness of stimuli to be used in perception experiments also strongly depends on the manipulation procedures applied to cross the audio and visual information. In this respect, the details given by McGurk & MacDonald (1976: 746)[8], already put the issue in the correct light, emphasizing the importance of the temporal alignment of auditory and visual information. Even if the concern is not directly the segmental information, as in the original McGurk & MacDonald investigation, this is an important matter any time a manipulation is necessary to match information conveyed by

to the other two models, different sources of information may be processed. According to the WTAV, they "are averaged according to the weight assigned to each modality" (Srinivasan & Massaro 2003: 10), while according to the FLMP the influence of one modality is going to be greater when the other is weaker and more ambiguous.

[8] See citation reported above.

different channels, for instance, any time incongruous stimuli are created. In fact, the naturalness of stimuli represents one of the most important factors to warrant the reliability of collected perceptual data. It may be important to keep the issue in mind even before the creation of incongruous stimuli, that is when the originals are segmented. Indeed, as Gili Fivela (2015: 211) observes, generating files of very similar duration (and, in particular, audio-video composed by the same number of frames) and in which the utterance starts after a given time-interval from the starting point of the file may be of great help in facilitating the best match when modifying the pairing of the two channels in order to generate the various audio-video combinations. Of course, the utterance duration itself within the file may be another issue as, even warranting the same starting point in the production of speech and visual information, a problem may relate to the matching of the utterance length and this may require some extra manipulation. Moreover, particularly when considering visual information and prosody or, more specifically, intonation, explicit attention has to be devoted to the alignment of visual and audio information when pitch accents are realized, as the peaking of visual information aligned with pitch accents is reported in the literature (e.g., Cassell et al. 1994; Loehr 2004; Swerts & Krahmer 2008). So the manipulation phase is very delicate and a final check on the naturalness of stimuli is needed to warrant the results of perception data collection.

4 Conclusions

The paper offers an overview of the methods used in the literature on prosody and intonation to perform multimodal analyses of audio-visual material conveying linguistic information in speech. Importantly, as for visual information, the paper discusses both articulatory gestures directly involved in the production of speech (e.g., lip gestures) and information that may be more traditionally considered and referred to as speech accompanying gestures (focusing on head movements and visual expressions).

Methods adopted to investigate speech production and perception are considered, by mainly describing experimental designs of works focusing mainly on Italian and some other Romance languages. The quite detailed description of methods offered in sections 2.2 and 3.2 aims at emphasizing the key aspects allowing the reader to choose among the various methods and aims at offering the relevant references for their deeper understanding. Additionally, it represents a necessary, preliminary step to discuss advantages and disadvantages related to the different methodological choices, both by highlighting very practical issues or drawbacks related to them and by stressing their impact in terms of theoretical issues and models they are used to refer to.

In very general terms, visual information as a whole may be taken to belong to the extralinguistic context the speakers resort to in order to understand messages and optimize them in production. However, some specific visual information clearly participates in conveying strictly linguistic information, such as sentence modality (see §2.2). The relevance of such visual cues with respect to the audio ones is still to be understood (e.g., see §3.3). However, the importance of resorting to both audio and visual cues is quite clear when thinking of most communication going on in everyday life. Moreover, it is clear also in specific situations. For instance, it is possible to create different local contexts in which the "truth value" of an utterance changes, by exploiting the flow of information in the channels or modalities available to the speakers (that is audio-only or audio-video, as discussed by Gili Fivela & Bazzanella 2014 and recalled at the beginning of the paper – see §1).

The examples discussed and the possible specific suggestions given in the paper are in line with the idea that multimodal analyses of multimodal, audio-visual information may be useful in order not only to understand the relation between the various sources of information we usually exploit in communication per se, but also to shed a possible new light on the variability otherwise observed in acoustic and articulatory investigation of speech material. The visual information may indeed represent an extra factor to be considered, besides those usually focused on in linguistic investigations, such as the lexical and syntactic make-up of utterances. Indeed, it may shed light on the variation observed in speech as for both pattern choice and phonetic implementation. Along this line of reasoning, it is plausible that the relevance of visual information could also play a role in relation to the differences observed in the perception of members of the same categories. In this respect, the existence of prototypes and non-prototypes, also mentioned in relation to the perception of intonation categories (e.g., Schneider & Möbius 2005; Schneider et al. 2006; 2009; Gili Fivela 2012; see Footnote 2 in §3.1), could also turn out to be relevant to the issue. Indeed, a non-prototypical member of a category, judged because of its acoustic characteristics, may actually be judged differently once that visual information is also considered. This would be in line with the possibility to resort to intra-category variability to express shades of meanings by means of the modulation of acoustic and, possibly, visual information too.

References

Akman, Varol & Carla Bazzanella. 2003. The complexity of context. *Journal of Pragmatics* 35(3). 321–329.

Anderson, Anne H., Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth M. Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, Catherine Sotillo, Henry S. Thompson & Regina Weinert. 1991. The HCRC Map Task Corpus. *Language and Speech* 34. 351–366.

Avesani, Cinzia, Mario Vayra & Claudio Zmarich. 2007. On the articulatory bases of prominence in Italian. In *Proceedings of ICPhS*, 981–984. Saarbrücken, Germany.

Avesani, Cinzia, Mario Vayra & Claudio Zmarich. 2009. Coordinazione vocale-consonante e prominenza accentuale in italiano. La sfida della Articulatory Phonology. In G. Ferrari, R. Benatti & M. Mosca (eds.), *Linguistica e modelli tecnologici di ricerca* (Pubblicazioni della Società di Linguistica Italiana), 353–386. Roma: Bulzoni.

Ayers, Gayle. 1994. Discourse functions of pitch range in spontaneous and read speech. *OSU Working Papers in Linguistics* (44). 1–49.

Barkhuysen, Pashiera, Emiel Krahmer & Marc Swerts. 2005. Problem detection in human-machine interactions based on facial expressions of users. *Speech Communication* 45. 343–359.

Barkhuysen, Pashiera, Emiel Krahmer & Marc Swerts. 2008. The interplay between the auditory and visual modality for end-of-utterance detection. *Journal of the Acoustical Society of America* 123(1). 354–65.

Beccaria, Gian Luigi. 1994. *Dizionario di linguistica e di filologia, metrica, retorica.* Torino: Einaudi.

Beckman, Mary E. 1997. A Typology of Spontaneous Speech. In Yoshinori Sagisaka, Nick Campbell & Norio Higuchi (eds.), *Computing Prosody. Computational Models for Processing Spontaneous Speech*, 7–26. New York: Springer.

Beckman, Mary E., Jan Edwards & Janet Fletcher. 1992. Prosodic structure and tempo in a sonority model of articulatory dynamics. In G.J. Docherty & D.R. Ladd (eds.), *Papers in Laboratory Phonology II: Gesture, Segment, Prosody*, 68–86. Cambridge: Cambridge University Press.

Berlin, Brent & Kay Paul. 1969. *Basic Color Terms: Their Universality and Evolution.* Berkeley: University of California Press.

Blaauw, Eleonora. 1995. *On the perceptual classification of spontaneous and read speech.* Utrecht, Netherlands: OTS (Institute for Language & Speech) dissertation.

Blum-Kulka, Shoshana, Juliane House & Gabriele Kasper. 1989. Investigating crosscultural pragmatics: An introductory overview. In Shoshana Blum-Kulka, Juliane House & Gabriele Kasper (eds.), *Cross-cultural Pragmatics. Requests and Apologies*, 1–34. Norwood (NJ): Ablex.

Boersma, Paul & David Weenink. 2017. *Praat: Doing phonetics by computer [Computer program]*. Version 6.0.30. http://www.praat.org/.

Borràs-Comes, Joan, Cecilia Pugliesi & Pilar Prieto. 2011. Audiovisual competition in the perception of counter-expectational questions. In Giampiero Salvi, Jonas Beskow, Olov Engwall & Samer Al Moubayed (eds.), *Proceedings of the 11th International Conference on Auditory-Visual Speech Processing*, 43–46. Stockholm: Volterra.

Browman, Catherine & Louis Goldstein. 1985. Dynamic modeling of phonetic structure. In V. Fromkin (ed.), *Phonetic Linguistics*, 35–53. New York: Academic Press.

Browman, Catherine & Louis Goldstein. 1995. Dynamics and Articulatory Phonology. In R. Port & T. Van Gelder (eds.), *Mind in Motion: Explorations in the Dynamics of Cognition*, 175–193. Cambridge, MA: The MIT Press.

Browman, Catherine, Louis Goldstein, J. A. Scott Kelso, Philip Rubin & Elliot Saltzman. 1984. Articulatory synthesis from underlying dynamics. *Journal of the Acoustical Society of America* 75. 22.

Brown, Gillian, Anne Anderson, George Yule & Richard Shillcock. 1983. *Teaching talk*. Cambridge: Cambridge University Press.

Byrd, Dany & Elliot Saltzman. 1998. Intragestural dynamics of multiple phrasal boundaries. *Journal of Phonetics* 26. 173–199.

Byrd, Dany & Elliot Saltzman. 2003. The elastic phrase: Modeling the dynamics of boundary-adjacent lengthening. *Journal of Phonetics* 31. 149–180.

Cassell, Justine, Catherine Pelachaud, Norm Badler, Mark Steedman, Brett Achorn, Tripp Becket, Brett Douville, Scott Prevost & Matthew Stone. 1994. Modeling the interaction between speech and gesture. In *Proceedings of the 16th Annual Conference of the Cognitive Science Society.*

Cavé, Christian, Isabelle Guaïtella, Roxane Bertrand, Serge Santi, Françoise Harlay & Robert Espesser. 1996. About the relationship between eyebrow movements and F0 variations. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, 2175–2178. Philadelphia, PA: IEEE.

Crespo-Sendra, Verònica, Costantijn Kaland, Marc Swerts & Pilar Prieto. 2013. Perceiving incredulity. The role of intonation and facial gestures. *Journal of Pragmatics* 47. 1–13.

Crystal, David. 1995. *A Dictionary of Linguistics and Phonetics.* Oxford (UK): Blackwell Publishing.

D'Imperio, Mariapaola. 2002. Language-specific and universal constraints on tonal alignment: The nature of targets and "anchors". In Bernard Bel & Isabelle Marlien (eds.), *Proceedings of the 1st International Conference on Speech Prosody*, 101–106. Aix-en-Provence: Laboratoire Parole et Langage.

D'Imperio, Mariapaola, Robert Espesser, Hélène Loevenbruck, Caroline Menezes, Noël Nguyen & Pauline Welby. 2007. Are tones aligned with articulatory events? Evidence from Italian and French. In Jennifer Cole & José Ignacio Hualde (eds.), *Laboratory phonology 9*, vol. 4-3 (Phonology and phonetics), 577–608. Berlin: Mouton de Gruyter.

Dijkstra, Christel, Emiel Krahmer & Marc Swerts. 2006. Manipulating Uncertainty. The contribution of different audiovisual prosodic cues to the perception of confidence. In Rüdiger Hoffmann & Hansjörg Mixdorff (eds.), *Proceedings of the 3rd International Conference on Speech Prosody*, 1–4. Dresden.

Dohen, Marion & Hélène Loevenbruck. 2009. Interaction of audition and vision for the perception of prosodic contrastive focus. *Language and Speech* 52. 177–206.

Edwards, J., Mary E. Beckman & Janet Fletcher. 1991. The articulatory kinematics of final lengthening. *Journal of the Acoustical Society of America* 89(1). 369–382.

Ekman, Paul & Wallace Friesen. 1978. *Facial Action Coding System: A Technique for the Measurement of Facial Movement.* Palo Alto: Consulting Psychologists Press.

Ekman, Paul, Wallace Friesen & Joseph C. Hager. 2002. *Facial Action Coding System: The Manual on CD ROM. A Human Face.* Salt Lake City.

Fowler, Carol. 1995. Acoustic and kinematic correlates of contrastive stress accent in spoken English. In Fredericka Bell-Berti & Raphael J. Lawrence (eds.), *Producing Speech: Contemporary Issues*, 355–373. New York: AIP Press.

Frota, Sónia & Pilar Prieto. 2015. Intonation in Romance: Systemic similarities and differences. In Sónia Frota & Pilar Prieto (eds.), *Intonation in Romance*, 392–418. Oxford: Oxford University Press.

Fry, Dennis B., Arthur S. Abramson, Peter D. Eimas & Alvin M. Liberman. 1962. The identification and discrimination of synthetic vowels. *Language and Speech* 5. 171–189.

Geng, Christian, Alice Turk, James M. Scobbie, Cedric Macmartin, Philip Hoole, Philip Richmond, Alan Wrench, Marianne Pouplier, Ellen G. Bard, Ziggy Campbell, Catherine Dickie, Eddie Dubourg, William Hardcastle, Evia

Kainada, Simon King, Robin Lickley, Satsuki Nakai, Steve Renals, Kevin White & Ronny Wiegand. 2013. Recording speech articulation in dialogue: Evaluating a synchronized double electromagnetic articulography setup. *Journal of Phonetics* 41(6). 421–431.

Gili Fivela, Barbara. 2008. *Intonation in Production and Perception: The Case of Pisa Italian.* Alessandria: Edizioni dell'Orso. Memorie del Laboratorio di Linguistica della Scuola Normale Superiore di Pisa.

Gili Fivela, Barbara. 2012. Meanings, shades of meanings and prototypes of intonational categories. In Gorka Elordieta & Pilar Prieto i Vives (eds.), *Prosody and meaning* (Interface explorations), 197–237. Berlin & Boston: De Gruyter Mouton.

Gili Fivela, Barbara. 2015. L'integrazione di informazioni multimodali: prosodia ed espressioni del volto nella percezione del parlato. In Elena Pistolesi, Rosa Pugliese & Barbara Gili Fivela (eds.), *Parole, gesti, interpretazioni: Studi linguistici per Carla Bazzanella*, 107–127. Roma: Aracne.

Gili Fivela, Barbara, Cinzia Avesani, Marco Barone, Giuliano Bocci, Claudia Crocco, Mariapaola D'Imperio, Rosella Giordano, Giovanna Marotta, Michelina Savino & Patrizia Sorianello. 2015. Varieties of Italian and their intonational phonology. In Sónia Frota & Pilar Prieto (eds.), *Intonation in Romance*, 140–197. Oxford University Press.

Gili Fivela, Barbara & Carla Bazzanella. 2014. The relevance of prosody and context to the interplay between intensity and politeness: An exploratory study on Italian. *Journal of Politeness Research* 10(1). 97–126.

Goldstein, Louis, Hosung Nam, Elliot Saltzman & Ioana Chitoran. 2009. Coupled Oscillator Planning Model of Speech Timing and Syllable Structure. In H. Fujisaki, J. Shen & G. Fant (eds.), *Frontiers in Phonetics and Speech Science*, 239–250. Beijing: The Commercial Press.

Gussenhoven, Carlos. 1999. Discreteness and gradience in intonational contrasts. *Language and Speech* 42. 281–305.

Gussenhoven, Carlos. 2004. *The phonology of tone and intonation.* Cambridge: Cambridge University Press.

House, David. 2002. Intonation and visual cues in the perception of interrogative mode in Swedish. In *Proceedings of ICSLP 2002*, 1957–1960.

Isard, Amy & Jean Carletta. 1995. Replicability of transaction and action coding in the Map Task corpus. In *Proceedings of AAAI Spring Symposium on Empirical Methods in Discourse Interpretation.* Palo Alto, CA.

Krahmer, Emiel & Marc Swerts. 2007. The effect of visual beats on prosodic prominence: acoustic analyses, auditory perception, and visual perception. *Journal of Memory and Language* 57. 396–414.

Kuhl, Patricia. 1991. Human adults and human infants show a 'perceptual magnet effect' for the prototypes of speech categories, monkeys do not. *Perception and Psychophysics* 50. 93–107.

Ladd, D. Robert. 1996. *Intonational phonology.* Vol. 79 (Cambridge studies in linguistics). Cambridge: Macmillan.

Ladd, D. Robert. 2008. *Intonational phonology.* 2nd edition. Cambridge: Cambridge University Press.

Lansing, C. R. & George W. McConkie. 1999. Attention to facial regions in the segmental and prosodic visual speech percept ion tasks. *Journal of Speech, Language, and Hearing Research.* 526–539.

Lehiste, Ilse. 1975. The phonetic structure of paragraphs. In Antonie Cohen & Sieb Nooteboom (eds.), *Structure and Process in Speech Perception*, 195–206. Springer-Verlag.

Lehiste, Ilse & William Wang. 1977. Perception of sentence and paragraph boundaries with and without semantic information. In Wolfgang Dressler & Oskar Pfeiffer (eds.), *Phonologica*, 277–283.

Liberman, Alvin M., Katherine S. Harris, Howard S. Hoffman & Belver C. Griffith. 1957. The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology* 54(5). 358–368.

Loehr, Dan. 2004. *Gesture and Intonation.* Washington, DC: Georgetown University dissertation.

Massaro, Dominic W. 1989. Testing between the TRACE model and the fuzzy logical model of speech perception. *Cognitive Psychology* 21(3). 398–421.

Massaro, Dominic W. & Michael Cohen. 1993. The paradigm and the fuzzy logical model of perception are alive and well. *Journal of Experimental Psychology* 122(1). 115–124.

McGurk, Harry & John MacDonald. 1976. Hearing lips and seeing voices: A new illusion. *Nature* 264. 746–748.

Mücke, Doris, Martine Grice, Johannes Becker & Anne Hermes. 2009. Sources of variation in tonal alignment: Evidence from acoustic and kinematic data. *Journal of Phonetics* 37. 321–338.

Nespor, Marina & Wendy Sandler. 1999. Prosody in Israeli sign language. *Language and Speech* 42. 143–176.

Niebuhr, Oliver & Klaus Kohler. 2004. Perception and cognitive processing of tonal alignment in German. In *Proceedings of the International Symposium on Tonal Aspects of Languages: Emphasis on Tone Languages*, 155–158. Beijing.

Oller, D. K. 1973. The effect of the position in utterance on speech segment duration in English. *Journal of the Acoustical Society of America* 54. 1235–1247.

Pean, Vincent, Sheila M. Williams & Maxine Eskenazy. 1993. The design and recording of ICY, a corpus for the study of intraspeaker variability and the characterisation of speaking styles. In *Proceedings of Eurospeech 1993*, 627–630. Berlin, Germany.

Perkell, Joseph S., Marc H. Cohen, Mario A. Svirsky, Melanie L. Matthies, Iñaki Garabieta & Michel T. T. Jackson. 1992. Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements. *Journal of the Acoustical Society of America* 92. 3078–3096.

Pierrehumbert, Janet B. 1980. *The phonology and phonetics of English intonation*. Bloomington: MIT dissertation.

Prieto, Pilar. 2012. Experimental methods and paradigms for prosodic analysis. In Abigail C. Cohn, Cécile Fougeron & Marie K. Huffman (eds.), *The Oxford Handbook of Laboratory Phonology* (Oxford Handbooks in Linguistics), 528–538. Oxford: Oxford University Press.

Prieto, Pilar, Doris Mücke, J. Becker & Martine Grice. 2007. Coordination patterns between pitch movements and oral gestures in Catalan. In *Proceedings of ICPhS*, 989–992. Saarbrücken, Germany.

Prieto, Pilar & Francisco Torreira. 2007. The segmental anchoring hypothesis revisited: Syllable structure and speech rate effects on peak timing in Spanish. *Journal of Phonetics* 35. 473–500.

Rosch, Eleanor H. 1975. Cognitive representations of semantic categories. *Journal of Experimental Psychology: General* 104(3). 192–233.

Rubin, Philip E. 1995. HADES: A case study of the development of a signal analysis system. *Applied speech technology.* 501–520. Boca Raton, FL, CRC Press.

Saltzman, Elliot. 1995. Dynamics and Coordinate Systems in skilled sensorimotor activity. In T. van Gelder & R. Port (eds.), *Mind as Motion: Explorations in the Dynamics of Cognition*, 150–173. Cambridge, MA: MIT Press.

Saltzman, Elliot & Kevin G. Munhall. 1989. A dynamical approach to gestural patterning in speech production. *Ecological Psychology* 1. 333–382.

Sandler, Wendy. 2005. Prosodic constituency and intonation in a sign language. In *Linguistische Berichte*, vol. 13, 59–86.

Savy, Renata & Francesco Cutugno. 2009. CLIPS: Diatopic, diamesic and diaphasic variations in spoken Italian. In Michaela Mahlberg, Victorina González-

Díaz & Catherine Smith (eds.), *On-line Proceedings of 5th Corpus Linguistics Conference* (paper 213). http://ucrel.lancs.ac.uk/publications/cl2009/213_FullPaper.doc.

Schneider, Katrin, Grzegorz Dogil & Bernd Möbius. 2009. German boundary tones show Categorical Perception and perceptual magnet effect when presented in different contexts. In *Proceedings of the 10th Annual Conference of the International Speech Communication Association 2009 (INTERSPEECH 2009)*, 2519–2522. Brighton, UK: Curran Associates, Inc.

Schneider, Katrin, Britta Lintfert, Grzegorz Dogil & Bernd Möbius. 2006. Phonetic grounding of prosodic categories. In Stefan Sudhoff, Denisa Lenertové, RolandMeyer, Sandra Pappert, Petra Augurzky, Ina Mleinek, Nicole Richter & Johannes Schliesser (eds.), *Methods in Empirical Prosody Research*, 335–362. Berlin: De Gruyter.

Schneider, Katrin & Bernd Möbius. 2005. Perceptual magnet effect in German boundary tones. In *Proceedings of the 6th Annual Conference of the International Speech Communication Association 2005 (INTERSPEECH 2005)*, 41–44. Lisbon, Portugal: Curran Associates, Inc.

Sigona, Francesco, Antonio Stella, Mirko Grimaldi & Barbara Gili Fivela. 2015. MAYDAY: A software for multimodal articulatory data analysis. In Antonio Romano & I. Meandri Rivoira (eds.), *Aspetti prosodici e testuali del raccontare: dalla letteratura orale al parlato dei media, Atti del 10° convegno AISV*, 173–184. Torino: Edizioni dell'Orso.

Silverman, Kim, Mary E. Beckman, John F. Pitrelli, Mari Ostendorf, C. Wightman, Patti Price, Janet B. Pierrehumbert & Julia Hirschberg. 1992. TOBI: A standard for labeling English prosody. In Bruce L. Berwing, Terrance M. Nearey & John J. Ohala (eds.), *Proceedings of the 1992 International Conference on Spoken Language Processing*, 867–870.

Sjolander, K. & J. Beskow. 1999. *WaveSurfer: An open source speech tool.* Stockholm, Sweden: Center for Speech Technology (CTT) at KTH. https://www.speech.kth.se/wavesurfer/wsurf_icslp00.pdf.

Srinivasan, Ravindra. J. & Dominic W. Massaro. 2003. Perceiving prosody from the face and voice: Distinguishing statements from echoic questions in English. *Language and Speech* 46(1). 1–22.

Stella, Antonio, Maria del Mar Vanrell, Massimiliano Iraci, Pilar Prieto & Barbara Gili Fivela. 2014. Intergestural coordination between tonal and oral gestures in Catalan, Italian. In Susanne Fuchs, Martine Grice, Anne Hermes, Leonardo Lancia & Doris Mücke (eds.), *Proceedings of the 10th International Seminar on Speech Production*, 421–424. Cologne, Germany.

Stella, Massimo, Paolo Bernardini, Francesco Sigona, Antonio Stella, Mirko Grimaldi & Barbara Gili Fivela. 2012. Numerical instabilities and three-dimensional electromagnetic articulography. *Journal of Acoustic. Soc. of Am.* 132(6). 3941–3949.

Stella, Massimo, Antonio Stella, Francesco Sigona, Paolo Bernardini, Mirko Grimaldi & Barbara Gili Fivela. 2013. Electromagnetic Articulography with AG500 and AG501. In *Proceedings of the 14th Annual Conference of the International Speech Communication Association 2013 (INTERSPEECH 2013)*, 1316–1320. Lyon, France.

Stevens, Kenneth N. 1972. The Quantal Nature of Speech: Evidence from Articulatory-acoustic Data. In David Jr. Edward & Peter B. Denes (eds.), *Human Communication: A Unified View*, 51–66. New York: McGraw-Hill.

Stevens, Kenneth N. 1989. On the quantal nature of speech. *On the quantal nature of speech* 17. 3–45.

Stevens, Kenneth N. & Samuel J. Keyser. 2010. Quantal theory, enhancement andoverlap. *Journal of Phonetics* 38(1). 10–19.

Stone, Maureen. 2005. A Guide to Analyzing Tongue Motion from Ultrasound Images. *Clinical Linguistics and Phonetics* 19(6-7). 455–502.

Swerts, Marc & Emiel Krahmer. 2008. Facial expressions and prosodic prominence: comparing modalities and facial areas. *Journal of Phonetics* 36(2). 219–238.

Tisato, Graziano, Piero Cosi, Carlo Drioli & Fabio Tesser. 2005. Interface. Strumenti interattivi per l'animazione delle teste parlanti. In Piero Cosi (ed.), *Misura dei parametri, Etti del I convegno nazionale dell'AISV, Padova*, 817–846. Brescia: EDK Editore.

Vanrell, Maria del Mar. 2006. A scaling contrast in Majorcan Catalan interrogatives. In Rüdiger Hoffmann & Hansjörg Mixdorff (eds.), *Proceedings of the 3rd International Conference on Speech Prosody*, 807–810. Dresden.

Vanrell, Maria del Mar, Ingo Feldhausen & Lluïsa Astruc. 2018. The Discourse Completion Task in Romance prosody research: status quo and outlook. In Ingo Feldhausen, Jan Fliessbach & Maria del Mar Vanrell (eds.), *Methods in prosody: A Romance language perspective* (Studies in Laboratory Phonology), 191–228. Berlin: Language Science Press.

Wightman, Colin, Stefanie Shattuck-Hufnagel, Mari Ostendorf & Patti Price. 1992. Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of the Acoustical Society of America* 91(3). 1707–1717.

Wilbur, Ronnie B. 2000. Phonological and prosodic layering of non-manuals in American sign language. In Karen Emmorey & Harlan Lane (eds.), *The Signs of Language Revisited*, 215–247. Mahwah, NJ: Lawrence Erlbaum Associates.

Wouters, Johan, Brian Rundle & Michael Macon. 1999. Authoring tools for speech synthesis using the sable markup standard. In *Proceedings of Eurospeech*, 963–966.