Speech Through the Ear, the Eye, the Mouth and the Hand

Marion Dohen

Dohen M., 2009. Speech Through the Ear, the Eye, the Mouth and the Hand. In Esposito A., Hussain A., Marinaro M. (Eds), *Multimodal Signals: Cognitive and Algorithmic Issues,* pp. 24-39. Springer: Berlin/Heidelberg.

Speech Through the Ear, the Eye, the Mouth and the Hand

Marion Dohen
Speech and Cognition Department, ICP, GIPSA-lab
961 rue de la Houille Blanche – Domaine Universitaire
38402 Saint Martin d'Hères Cedex, France
Marion.Dohen@gipsa-lab.inpg.fr

Abstract. This chapter aims at describing how speech is multimodal not only in its perception but also in its production. It first focuses on multimodal perception of speech segments and speech prosody. It describes how multimodal perception is linked to speech production and explains why we consider speech perception as a sensory-motor process. It then analyses some aspects of hand-mouth coordination in spoken communication. Input from evolutionary perspectives, ontogenesis and behavioral studies on infants and adults are combined to detail how and why this coordination is crucial in speech.

Keywords: multimodal speech, auditory-visual perception, sensori-motor, prosody, hand-mouth coordination, speech and language development, pointing

1 Introduction

Speech is multimodal and is produced with the mouth, the vocal tract, the hands and the entire body and perceived not only with our ears but also with our eyes (and even our hands in the Tadoma method for example). The aim of this chapter is to draw a general working framework for the analysis of speech in a multimodal perspective.

It will focus on two main points. The first one will be multimodal perception of speech segments and prosody which will be related to their production in order to explain how and why speech perception is a sensory-motor process. The second part of the chapter will explore some aspects of hand-mouth coordination in adult spoken communication. It will describe why this coordination is so important for understanding cognitive communicative processes. A conclusion will then summarize all the ideas developed throughout and describe their implications for studying speech in a multisensory perspective. The framework presented here is based on an analysis

and synthesis of previous works which have led to the conception of speech as a multisensory process in nature. It aims at putting forward the importance of conceiving speech as multisensory to analyze and describe the cognitive processes involved.

2 Multimodal Perception of Speech Segments and Speech Prosody in Relationship with their Production

> "Speech is rather a set of movements made audible than a set of sounds produced by movements."
>
> Raymond H. Stetson [1]

As R.H. Stetson put it, speech is truly multisensory and does not consist of sounds which are produced, somehow, just to be heard.

2.1 Speech is Also Visual

It is possible to perceive and understand speech with the eyes: 40 to 60% of the phonemes and 10 to 20% of the words (up to 60%) can be recovered through lipreading (data from [2]). This ability is highly inter-speaker dependent: the deaf are often better than the hearing for example. As shown in many studies [3-12], the use of vision for speech perception is obvious when the acoustic modality is degraded by noise. Many other situations however put forward the advantage of seeing a speaker while he/she is speaking. For example, when spelling something over the phone, you often have to disambiguate: M as in 'made', N as in 'nanny'…. Reisberg and colleagues [13] showed that vision also helped for perceiving speech in a foreign language, speech produced by a non-native speaker or semantically complex utterances.

This could lead to think that vision simply provides redundant information: when some of the auditory information is missing, it can be recovered by vision. As shown by Summerfied [14], perceptual confusions between consonants are different and complementary in the visual and the auditory modalities. Another clear and very famous example of the fact that the visual information is not just redundant is the McGurk effect [15]. An audio [ba] dubbed onto a visual [ga] results in a [da] percept. This effect is very robust: it works in many languages and, even when people are aware of it, it can not be suppressed [15]. All these observations suggest that the visual and auditory information are fused rather than the visual information being superimposed on the auditory one.

If the auditory and the visual information are not redundant, it suggests that speaking is producing gestures aiming at being both heard *and* seen. This is further backed by the fact that there is a marked preference for bilabials (consonants articulated with an initial contact of the superior and inferior lips), which are visually salient, at the beginning of babbling [16]. This preference is reinforced in deaf children [17] and unsalient in blind children [18]. Moreover, the [m]/[n] contrast which exists in almost all the languages in the world [19], is not very audible but very

visible. Moreover, the fact that we use vision for speech perception does not seem to be a learnt skill. Infants 18 to 20 weeks-old, can indeed identify a speaking face corresponding to a speech signal out of two movies of speaking faces (one congruent to the audio and another incongruent) displayed at the same time [20]. Stetson's statement could therefore be adjusted to:

> Speech is (…) a set of movements made audible *and visible*.

2.2. Auditory-Visual Fusion: How and When?

If the auditory and visual information are fused in speech perception, one can wonder which cognitive processes underlie this fusion: how and when does it happen? After analyzing the fusion models proposed in the literature, Robert-Ribes [21], Schwartz and colleagues [22] presented four main potential fusion architectures summarized in Fig. 1:

- **Direct Identification** (DI): the auditory and visual channels are directly compiled.
- **Separate Identification** (SI): the phonetic classification is operated separately on both channels and fusion occurs after this separate identification. Fusion is therefore relatively late and decisional.
- **Recoding in the dominating modality** (RD): the auditory modality is considered to be dominant and the visual channel is recoded under a compatible format to that of the auditory representations. This is an early fusion process.
- **Recoding in the motor modality** (RM): the main articulatory characteristics are estimated using the auditory and visual information. These are then fed to a classification process. This corresponds to an early fusion.

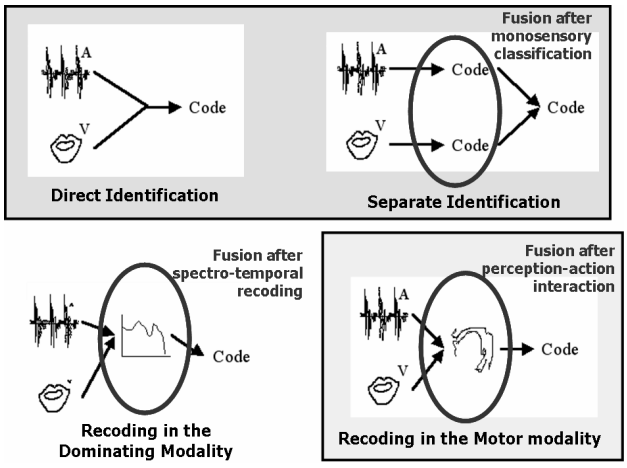

Fig. 1. The four main types of auditory-visual fusion models as presented in [21-22].

As stated in [2], the models which have been mostly tested in speech recognition applications are the DI and SI models because they are easier to implement. Some studies of brain activations however suggest that the interactions between modalities may occur very early (*e.g.* [23]). Moreover, as suggested in [2], a number of behavioral studies provide valuable information on the fusion process. Grant and Seitz [24] suggested the fact that, not only does vision help us better understand speech in noise, it also improves auditory detection of spoken sentences in noise. Actually, when we see a speaker, we perceive him/her as speaking louder. In the line of this observation, Schwartz and colleagues [25] designed an experiment to further test whether there are early interactions between audition and vision. They tested the intelligibility in noise of sequences which are not distinguishable by lip-reading. They found that visual only perception was not significantly better than chance which confirmed that nothing could be recovered from lip-reading. They however found that adding vision significantly improved auditory perception (AV>A). They interpreted this as potentially corresponding to reinforcement of the voicing feature by vision. The beginning of the labial gesture starts about 240ms before the vowel target is reached. In French, voicing is preceded by a prevoicing of about 100ms. Visual perception of the beginning of the labial gesture would therefore announce to the listener that he/she should lend an ear to detect the potential prevoicing. This temporal marker could reduce by a few dBs the detection threshold. The authors also conducted a second experiment in which the visual information of the speaker was replaced with visual information corresponding to a volume bar moving just as the speaker's lips. In this case, they found no improvement of perception when vision was added to audition suggesting that the effect described above is "speech specific". Put together, these behavioral data further suggest that the auditory-visual interactions are early which is incompatible with the DI and SI models.

Another question raised in [2] is that of the control of the fusion process: is there a context bias? For example, as shown by Tiippana and colleagues [26], visual attention can modulate audiovisual speech perception. If visual attention is distracted, the McGurk effect is indeed weaker. There also are strong inter-individual variations (see *e.g.* [27]) as well as inter-linguistic differences (see *e.g.* [28]).

The data presented above suggest that the RD and RM models appear to be more likely to reflect the cognitive processes underlying auditory-visual fusion. The RM model states that the auditory and visual channels would be combined and recoded into motor representations. This suggests that there would be a link between speech perception and speech production.

2.3. Evidence for Perceptuo-Motor Links in the Human Brain

The first thing we should talk about here is mirror neurons: not because it is fashionable but because this system could be important for speech… The mirror-neuron system was found in monkeys (area F5 of the monkey premotor cortex) by Rizzolati and colleagues (for reviews, see: [29,30]). These neurons are neurons which fire during both the production and the perception of goal-directed actions as well as

during watching a similar action made by another individual (*e.g.* pick up a peanut on a tray). Kohler, Keysers and colleagues [31,32] suggested the existence of audio-visual mirror-neurons. These particular neurons fire when the monkey hears the sound produced by an action (*e.g.* experimenter breaking a peanut), when it sees the action being produced, when it sees and hears the action being produced and when it produces the action itself. Moreover, these neurons seem to be specialized: each neuron responds to one particular type of action (*e.g.* peanut breaking as opposed to ring grasping). Mirror neurons play a role in orofacial actions [33]: for example, the same neuron fires when the monkey grasps food with the mouth and when the experimenter grasps food with the mouth. This is also the case for "communicative" orofacial actions such as lip-smacking/lip protrusion [33-35]. All these observations were however made for monkeys: is there any evidence for a mirror neuron system in humans? It is not possible to record the activity of a single neuron in humans. A number of neurophysiological (EEG, MEG, TMS) and neuroimaging (fMRI) studies (for reviews see: [29,30,36]) have however showed activations of motor regions involved in performing specific actions in the perception of these actions performed by others. Rizzolatti and colleagues [37,36] have proposed that the mirror neuron system would play a fundamental role in speech processing by providing a neurophysiological mechanism that creates parity between the speaker and the listener and allows for direct comprehension. A number of neurobiological models of speech perception [38-45] actually consider that motor areas linked to speech production (the so-called dorsal stream) intervene in speech perception either always or only under certain circumstances (speech acquisition for example).

Some studies using single-pulse transcranial magnetic stimulation (TMS) have shown a motor resonance in speech perception. This technique consists in exciting a specific motor region (in this case linked to speech production) in order for the hypothesized motor response, which is weak, to go over a threshold and produce a motor evoked potential (MEP) which can be recorded using electromyography (EMG). Using this technique, Fadiga and colleagues [46] showed that, during the perception of a speech sequence, there is an increase in MEPs recorded from the listeners' tongue muscles when the segmental content of speech involves tongue movements compared to when it does not. Another study using this technique [47] showed that there is an increase in MEPs recorded from the listeners' lip muscles during auditory alone perception of speech compared to the perception of nonverbal audio signals and during the visual perception of speaking lips compared to that of moving eyes and brows. Some neuroimaging studies (fMRI: functional Magnetic Resonance Imaging) have shown that the dorsal stream is activated in speech perception [48,49].

All these observations suggest that speech perception is tightly linked to speech production and that it can be considered as a sensori-motor process.

2.4. What About Suprasegmentals?

All the observations described above were made for segmental perception of speech, that is, perception of which phonemes/words are produced by a speaker. Prosody (intonation, rhythm and phrasing) is however crucial in spoken communication as

illustrated by a nice example from L. Truss[1]. The two sentences below correspond to exactly the same segmental content, but, pronounced differently, they have completely different meanings (in this example, punctuation is the written equivalent of prosody):

A woman, without her man, is nothing.

A woman: without her, man is nothing.

For a long time, prosody was uniquely considered as acoustic/auditory. However, recent studies conducted in our lab, as well as studies conducted elsewhere (*e.g.* [50-61]) have showed that prosody also has potentially visible correlates (articulatory or other facial correlates). We analyzed contrastive informational focus as described by Di Cristo for French [62]. It is used to put forward a constituent pragmatically and contrastingly to another as in the example below (capital letters signal focus):

Did Bob eat the apple?

No, SARAH ate the apple.

It can be realized using prosody (as in the example). The acoustic (intonational, durational and phrasing) correlates of prosodic focus have been widely studied [63-69]. Several production studies [70-72] in which we measured the articulatory and facial correlates of contrastive prosodic focus in French showed that there are visible correlates to contrastive focus in French. Two main visible articulatory strategies (inter-individual differences) were found:

- Absolute strategy: the speakers hyper-articulate and lengthen the focused constituent;
- Differential strategy: the speakers both slightly hyper-articulate the focused constituent and hypo-articulate the following constituent(s).

Both these strategies result in a visible contrast within the utterance between what is focused and what is not. In the line of other studies ([73-76,59] also see [77-78] for audiovisual perception studies using animated talking heads), we found that prosodic contrastive focus was detectable from the visual modality alone and that the cues used for perception at least partly corresponded to those identified in the production studies [70,71].

In another study [79], we designed an experiment to avoid the ceiling effect: auditory only perception of prosodic focus is close to a 100% and a potential advantage of adding vision cannot be measured. The speech in noise paradigm is not adequate here since voicing is a robust feature to noise [80]. We used whispered speech for which there is no F0. We found that auditory only perception was degraded and that adding vision clearly improved prosodic focus detection for whispered speech (see Fig. 2a). Reaction time (RT) measurements showed that adding vision reduced processing time. A further analysis of the data suggested that audition and vision are actually integrated for the perception of prosodic contrastive focus in French.

[1] Truss, L.: Eats, Shoots & Leaves – The Zero Tolerance Approach to Punctuation. Profile Books Ltd, London (2003)

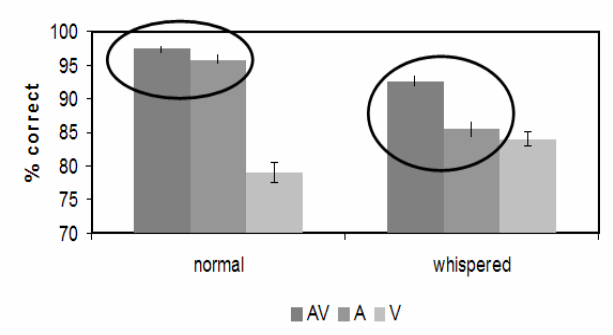

Fig. 2. Percentages of correct answers for each modality (AV, A, V) for loud and whispered speech.

These results suggest that, not only is segmental perception of speech multimodal, but supra-segmental perception of speech is as well.

3. Some Aspects of Hand-Mouth Coordination in Speech Communication

Speech is vocal and gestural by nature. As shown in the previous section, articulatory gestures are produced aiming at being both heard and seen. Articulatory gestures are however not the only gestures we produce when we speak. We move our entire body and especially our hands [81-84] and this is meaningful as well.

3.1. Gesture and Speech in an Evolutionary Perspective: Towards a Crucial Role of Hand-Mouth Coordination in Language Evolution

From an evolutionary perspective, theories defending vocal origins of language are opposed to those defending gestural origins of language. Vocal origin theories argue that speech and language would derive from animal calls such as alarm calls. In the framework of the Frame then Content theory and deriving phylogenesis (as the study of the evolution of a species) from ontogenesis (as the study of the development of a human being from birth to adulthood), MacNeilage [85,86] argues that the motor control of the articulators used in chewing would have led to speech and language. This suggests an orofacial (related to the face and mouth) basis for language. On the other hand, theories of gestural origins of language argue that the hand was used for communication before the mouth. Corbalis [87] argues that non-human primates (great apes) produce *unvoluntary* vocalizations whereas they *control and understand* manual gestures which are moreover often produced in dyads unlike calls. In his sense, this would have led to a gestural form of communication which would have led to a more complex control of the articulators and to vocal communication. What Arbib [88,89] and Holden [90] argue is that there would have been a co-evolution of

manual and speech gestural systems towards language and communication. This is in line with ontogenesis observations by Iversen and Thelen [91] who showed that the motor control of the vocal tract is longer to acquire than that of the hand but that the two systems develop in cooperation. This view underlines the fundamental link there would be between hand gestures and articulation and argues for a hand-mouth gestural basis for language. This suggests that hand-mouth coordination would have played a key role in communication evolution.

3.2. Gestures and Language Development: The Special Role of Pointing

As demonstrated, among others, by the researchers from Goldin-Meadow's team, gestures play a crucial role in language development (*e.g.* [92,93]). A key stage in language development is the combination of a word and a gesture carrying complementary (non redundant) information [94]. An example of such a combination is a baby saying "monster" and producing a gesture miming "big" to mean "big monster". These combinations lead to the production of two-word utterances. [95] underlines the fact that the motor links between the hand and the mouth develop progressively from birth until they reach synchrony (around 16 to 18 months). This association between speech and hand gestures seems to be a motor rather than a perceptive association. Blind children indeed produce gestures as well [96].

As pointed out by Butterworth [97]: "pointing is the royal road toward language for babies". Pointing indeed appears to play a special role at many stages of language development (see *e.g.* [94,98,99]). The first gestures, produced around 9-11 months and predictive of first word onset, are deictic (gestures that show). These are not synchronized to vocal productions. Then babies combine words and pointing gestures carrying the same information as in pointing at a dog and saying "dog" at the same time. Words are then combined to pointing gestures with no information redundancy as in pointing at a dog and saying "eat" to mean "the dog is eating". This leads to the production of two-word utterances at around 18 months. These observations show that pointing plays a special role in speech development in infants.

3.3. Hand-Mouth Coordination in Adult Spoken Communication

Many studies have analyzed hand-mouth coordination in adults' speech. The two main approaches commonly used are "controlled" experiments and analysis of spontaneous productions. "Controlled" experiments make precise quantitative evaluation of the links between gestures and speech possible but involve using controlled material and recordings in an artificial environment such as a lab. Analyzing spontaneous productions enables the evaluation of the "natural" links between gestures and speech but are not replicable and do not allow for quantitative analyses. In this chapter, I will mainly focus on the description of results from "controlled" experiments. For these experiments, two main protocols have been used: dual and single task paradigms and rhythmic tasks. I will describe results from dual and single task studies. In these studies, several conditions are compared to each other: speech alone, speech and gesture and gesture alone.

Hollender [100] conducted several experiments using a dual-task paradigm in which participants were asked to name a letter which appeared on a computer screen and press a key on a keyboard. This showed that, when speech and gesture were combined, the vocal response was delayed compared to the speech alone condition potentially for the two systems to be synchronized. The author concluded that there was an adaptation of the systems due to limitations in processing capacities. Castiello and colleagues [101] also conducted a dual-task study in which the participants uttered the syllable /tah/ while grasping an object. The authors found that "the two-responses were generated independently of each other" (see also [102]). These studies tend to suggest that hand-mouth coordination may not be so strict. However a crucial issue should be addressed: if the hand and mouth are coordinated in adult spoken communication, what could this coordination be like?

McNeill [82] described hand-mouth coordination as a synchronization of the hand gesture to the word which it is related to semantically and co-expressively. This raises a question as for the studies described above: there was no "semantic" relationship between the gestures and the speech. In this sense and as stated in [103,104], it seems particularly interesting to study pointing. For pointing, there should be synchrony between the gesture (which points at something) and the part of speech that shows (determiner, focus…). Levelt and colleagues [105] conducted a dual-task study on pointing and speech in which participants uttered "this lamp" while they pointed at the corresponding lamp. They observed a delay in both the manual and vocal responses in the gesture and speech condition. This delay was greater for the vocal response. These findings were replicated by Feyereisen [106]. These authors concluded that there were different processes which were competing for the same resources. This suggests that the hand and mouth systems would use the same resources and that one system would therefore have to wait for the other to free the resources to be able to use them. This would explain the delay measured for vocal responses. In a slightly different view, this delay could simply be due to coordination requirements: the vocal and gestural responses would have to be synchronized at some point and when a gesture is produced at the same time as speech, speech would wait in order for the synchrony to be achieved. Another important point is that the studies described above analyzed voice onset time. No articulatory measurements were conducted. If the coordination is a hand-mouth motor coordination, it seems unsatisfactory to analyze voice onset as an assessment of the vocal motor response. Finally, in the studies described above, the "part of speech that shows" (mostly a determiner) was always at the beginning of the utterance which probably has an effect on coordination.

Rochet-Capellan and colleagues [104] conducted a study in which they tested the effect of stress on hand-jaw coordination. They used a deictic task in which subjects pointed at a target while naming it. The name of the target was a CVCV bisyllable (2 consonant-vowel syllables) in which stress position was varied (*e.g.* /pápa/ *vs.* /papá/) as well as consonantal context (*e.g.* /pata/ *vs.* /tapa/). The target was always located in the right visual field, either near or far. Twenty native Brazilian Portuguese speakers were tested. The finger and jaw motion were captured using an IRED tracking system (Optotrak). The hypothesis was that, when stress was on the first (resp. second) syllable, the pointing apex would be synchronized with the maximum jaw displacement corresponding to the first (resp. second) syllable. This was actually the

case when stress was on the first syllable. When stress was on the second syllable, the pointing apex (furthest position of the index finger corresponding to the extension of the arm in the pointing gesture) was located half way between the two jaw apices and the return of the pointing gesture was synchronized with the maximum jaw displacement corresponding to the second syllable. There was no effect of target position or consonant on the coordinative pattern. A finger adaptation was however observed: a longer plateau (delay between pointing apex and beginning of the return stroke) was measured when stress was on the second syllable. There was also a high subject variability and the authors suggested an adaptation of the jaw from the first to the second stress condition: the jaw response onset was 70ms earlier but the maximum jaw displacements were 100ms earlier. The main conclusion the authors made was that the jaw was synchronized to the finger pointing plateau (the part of the pointing gesture that shows). This study shows that there seems to be a tight coordination of the hand and jaw which serves communication purposes. This study also confirms that the coordination is rather between the hand and the articulatory gestures than between the hand and the sound. However it only analyzed hand-jaw coordination and one can wonder about coordination with other articulators.

3.4. The Co-Development of Speech and Sign

Ducey-Kaufman and colleagues [107,108] designed a theoretical framework to analyze the co-development of speech and sign in speech acquisition. This theory mainly puts forward a "developmental rendez-vous" between what the authors call the sign frame (cyclicity of hand gesture in pointing) and the speech frame (jaw oscillations in speech). This meeting point occurs at about 12 months before when the speech and sign frame develop in parallel. This development leads to the possibility of producing two syllables within one pointing gesture and puts forward the potentially crucial role of the foot (phonological unit larger than or equal in size to the syllable and corresponding to the smallest rhythmic group) in speech. They tested this with a longitudinal study on six babies from 6 to 18 months. They found a 400ms mean duration of the syllable and a 800ms mean duration of the pointing gestures. Rochet-Capellan and colleagues tested this on adults [109]. The task they used was to name and show a target at a GO signal twice in a row. The name of the target was either a 1, or a 2, or a 3, or a 4 syllable word (/pá/ *vs.* /papá/ *vs.* /papapá/ *vs.* /papapapá/). They recorded finger motion using an IRED tracking system (Optotrak). The predictions were that the delay between the two pointing apices would remain constant for target names consisting of 1 or 2 syllables. This delay would then increase from 2 to 3 syllables and remain constant from 3 to 4 syllables. This is what they observed. They also analyzed the durations of the jaw gestures (from gesture initiation to the last maximum displacement, corresponding to the last syllable). They found a ratio between the duration of the pointing gesture and that of the jaw gestures of 0.5 whatever the number of syllables. These preliminary observations could confirm the two-syllable for one pointing gesture hypothesis.

4. Conclusion

The aim of this chapter was to draw a general multimodal framework for studying speech. The first aspect presented was the fact that speech is not only auditory. We also perceive speech through the eyes and the role of vision is not only that of a backup channel. Auditory and visual information are actually fused for perceptual decision. A number of potential cognitive fusion architecture models were presented. Some data suggest that audiovisual fusion takes place at very early stages of speech perception. Studies of the cognitive and brain mechanisms show that motor representations are most probably used in speech perception. Speech perception is a sensori-motor process: speech consists of gestures produced to be heard *and* seen. The multisensory aspect of speech is not limited to segmental perception of speech. It appears that there are visible articulatory correlates to the production of prosodic information such as prosodic focus and that these are perceived visually. Moreover, when the acoustic prosodic information is degraded, it is possible to put forward an auditory-visual fusion which enhances speech perception.

The second main aspect of this paper is the fact that we do not communicate only with our mouths. Our entire body, and especially our hands, is involved in this process. An analysis of the evolutionary theories of language development suggests that hand-mouth coordination may have played a crucial role in language evolution. This is at least the case for language development in babies in which the pointing gesture plays a crucial role. In adults, several studies seem to point out that hand and mouth (articulatory gestures) are tightly coordinated for communication purposes (for pointing at least). This still has to be analyzed for "real" speech (not only bisyllables) and for other communicative gestures: is the coordination gesture specific? Is it communication specific? Is it linked to grammar and prosody? Another important point is the fact that it appears that, "naturally" one pointing gesture would embed two syllables (in infants and maybe adults).

Acknowledgments. First of all, I would like to thank Anna Esposito for giving me the opportunity to take part in the International School "E. R. Caianiello" XIII course on Multimodal Signals: Cognitive and Algorithmic Issues. I also thank the two anonymous reviewers who provided insightful comments on a previous version of this chapter. For all the input he provided as well as for his advice and support, I thank Jean-Luc Schwartz. I also thank Amélie Rochet-Capellan for the inspiration I found in her work as well as for some material she provided. I also thank Marc Sato, Hélène Lœvenbruck and Gérard Bailly.

References

1. Stetson, R.H.: Motor Phonetics: A Study of Speech Movements in Action. North-Holland, 2nd Edition, Amsterdam (1951)
2. Schwartz, J.-L.: La parole multisensorielle: Plaidoyer, problèmes, perspective. In: Actes des XXVes Journées d'Etude sur la Parole JEP 2004, pp. xi-xviii (2004)
3. Sumby, W.H., Pollack, I.: Visual Contribution to Speech Intelligibility in Noise. J. Acoust. Soc. Am. 26(2), 212--215 (1954)

4. Miller, G.A., Nicely, P.: An Analysis of Perceptual Confusions Among Some English Consonants. J. Acoust. Soc. Am. 27(2), 338--352 (1955)
5. Neely, K.K.: Effects of Visual Factors on the Intelligibility of Speech. J. Acoust. Soc. Am. 28(6), 1275--1277 (1956)
6. Erber, N.P.: Interaction of Audition and Vision in the Recognition of Oral Speech Stimuli. J. Speech Hear. Res. 12(2), 423--425 (1969)
7. Binnie, C.A., Montgomery, A.A., Jackson, P.L.: Auditory and Visual Contributions to the Perception of Consonants. J. Speech Hear. Res. 17(4), 619--630. (1974)
8. Erber, N.P.: Auditory-Visual Perception of Speech. J. Speech Hear. Dis. 40(4), 481--492 (1975)
9. Summerfield, A.Q.: Use of Visual Information for Phonetic Perception. Phonetica 36, 314--331 (1979)
10. MacLeod, A., Summerfield, A.Q.: Quantifying the Contribution of Vision to Speech Perception in Noise. Brit. J. Audiol. 21, 131--141 (1987)
11. Grant, K.W., Braida, L.D.: Evaluating the Articulation Index for Audiovisual Input. J. Acoust. Soc. Am. 89, 2952--2960 (1991)
12. Benoît, C., Mohamadi, T., Kandel, S.: Effects of Phonetic Context on Audio-Visual Intelligibility of French. J. Speech Hear. Res. 37, 1195--1203 (1994)
13. Reisberg, D., McLean, J., Goldfield, A.: Easy to Hear but Hard to Understand: A Lip-reading Advantage with Intact Auditory Stimuli. In: Dodd, B., Campbell, R. (eds.) Hearing by Eye: The Psychology of Lip-reading, pp. 97--114. Lawrence Erlbaum Associates, Hillsdale, NJ (1987)
14. Summerfield, Q.: Comprehensive Account of Audio-Visual Speech Perception. In: Dodd, B., Campbell, R. (eds.), Hearing by Eye: The Psychology of Lip-reading, pp. 3--51. Lawrence Erlbaum Associates, Hillsdale, NJ (1987)
15. McGurk, H., MacDonald, J.: Hearing Lips and Seeing Voices. Nature 264, 746--748 (1976)
16. Vihman, M.M., Macken, M.A., Miller, R., Simmons, H., Miller, J.: From Babbling to Speech: A Re-Assessment of the Continuity Issue. Language 61, 397--445 (1985)
17. Stoel-Gammon, C.: Prelinguistic Vocalizations of Hearing-Impaired and Normally Hearing Subjects: A Comparison of Consonantal Inventories. J. Speech Hear. Dis. 53, 302--315 (1988)
18. Mulford, R.: First Words of the Blind Child. In: Smith, M.D., Locke J.L. (eds.), The Emergent Lexicon: The Child's Development of a Linguistic Vocabulary, pp. 293--338. Academic Press, New-York (1988)
19. Boë, L.J., Vallée, N., Schwartz, J.L.: Les tendances des structures phonologiques : le poids de la forme sur la substance. In: Escudier, P., Schwartz, J.L. (eds.), La parole, des modèles cognitifs aux machines communicantes - I. Fondements, pp. 283-323. Hermes, Paris (2000)
20. Kuhl, P.K., Meltzoff, A.N.: The Bimodal Perception of Speech in Infancy. Science 218, 1138--1141 (1982)
21. Robert-Ribes, J.: Modèles d'intégration audiovisuelle de signaux linguistiques : de la perception humaine à la reconnaissance automatique des voyelles. PhD Thesis, Institut National Polytechnique de Grenoble, France (1995)
22. Schwartz, J.-L., Robert-Ribes, J., Escudier, P.: Ten Years After Summerfield: A Taxonomy of Models for Audiovisual Fusion in Speech Perception. In: Campbell, R., Dodd, B.J., Burnham, D. (eds.), Hearing by Eye II: Advances in the Psychology of Speechreading and Auditory-Visual Speech, pp. 85--108. Psychology Press, Hove (1998)
23. Calvert, G.A., Bullmore, E.T., Brammer, M.J., Campbell, R., Williams, S.C.R., McGuire, P.K., Woodruff, P.W.R., Iversen, S.D., David, A.S.: Activation of the Auditory Cortex During Silent Lipreading. Science 276, 593--596 (1997)
24. Grant, K.W., Seitz, P.-F.: The Use of Visible Speech Cues for Improving Auditory Detection of Spoken Sentences. J. Acoust. Soc. Am. 108(3), 1197--1208 (2000)

25. Schwartz, J.-L., Berthommier, F., Savariaux, C.: Seeing to Hear Better: Evidence for Early Audio-Visual Interactions in Speech Identification. Cognition 93, B69--B78 (2004)
26. Tiippana, K., Andersen, T.S., Sams, M.: Visual Attention Modulates Audiovisual Speech Perception. Eur. J. Cog. Psychol. 16(3), 457--472 (2004)
27. Schwartz, J.-L., Cathiard, M.: Modeling Audio-Visual Speech Perception: Back on Fusion Architectures and Fusion Control. In: Proceedings of Interspeech 2004, pp. 2017--2020 (2004)
28. Sekiyama, K., Tohkura, Y.: Inter-Language Differences in the Influence of Visual Cues in Speech Perception. Journal of Phonetics 21, 427--444 (1993)
29. Rizzolatti, G., Fogassi, L., Gallese, V.: Neurophysiological Mechanisms Underlying the Understanding and Imitation of Action. Nat. Rev. Neurosci. 2(9), 661--670 (2001)
30. Rizzolatti, G., Craighero L.: The Mirror-Neuron System. Annu. Rev. Neurosci. 27, 169--192 (2004)
31. Kohler, E., Keysers, C., Umilta, M.A., Fogassi, L., Gallese, V., Rizzolatti G.: Hearing Sounds, Understanding Actions: Action Representation in Mirror Neurons. Science 5582(297), 846--848 (2002)
32. Keysers, C., Kohler, E., Umilta, M.A., Nanetti, L., Fogassi, L., Gallese, V.: Audiovisual Mirror Neurons and Action Recognition. Exp. Brain Res. 153, 628--636 (2003)
33. Ferrari, P.F., Gallese, V., Rizzolatti, G., Fogassi, L.: Mirror Neurons Responding to the Observation of Ingestive and Communicative Mouth Actions in the Monkey Ventral Premotor Cortex. Eur. J. Neurosci., 17, 1703--1714 (2003)
34. Gil-da-Costa, R., Martin, A., Lopes, M.A., Munoz, M., Fritz, J.B., Braun, A.R.: Species-Specific Calls Activate Homologs of Broca's and Wernicke's Areas in the Macaque. Nature Rev. Neurosci. 9, 1064--1070 (2006)
35. Taglialatela, J.P., Russell, J.L., Schaeffer, J.A., Hopkins W.D.: Communicative Signaling Activates 'Broca's' Homolog in Chimpanzees. Curr. Biol. 18, 343--348 (2008)
36. Rizzolatti, G., Craighero, L.: Language and Mirror Neurons. In: Gaskell G. (ed.), Oxford Handbook of Psycholinguistics. Oxford University Press, Oxford (2007)
37. Rizzolatti, G., Buccino G.: The Mirror-Neuron System and its Role in Imitation and Language. In: Dehaene, S., Duhamel, G.R., Hauser, M., Rizzolatti, G. (eds.), From Monkey Brain to Human Brain, pp. 213-234. MIT Press, Cambridge, NY (2005)
38. Hickok, G., Poeppel, D.: Towards a Functional Neuroanatomy of Speech Perception. Trends Cogn. Sci. 4(4), 132--138 (2000)
39. Scott, S.K., Johnsrude, I.S.: The Neuroanatomical and Functional Organization of Speech Perception. Trends Neurosci. 26(2), 100--107 (2003)
40. Callan, D.E., Jones, J.A., Munhall, K., Kroos, C., Callan, A.M., Vatikiotis-Bateson, E.: Multisensory Integration Sites Identified by Perception of Spatial Wavelet Filtered Visual Speech Gesture Information. J. Cogn. Neurosci. 16(5), 805--816 (2004)
41. Hickok, G., Poeppel, D.: Dorsal and Ventral Streams: A Framework for Understanding Aspects of the Functional Anatomy of Language. Cognition 92, 67--99 (2004)
42. Scott, S.K., Wise, R.J.S.: The Functional Neuroanatomy of Prelexical Processing in Speech Perception. Cognition 92(1-2), 13--45 (2004)
43. Wilson, S.M., Iacoboni, M.: Neural Responses to Non-Native Phonemes Varying in Producibility: Evidence for the Sensorimotor Nature of Speech Perception. NeuroImage 33, 316--325 (2006)
44. Hickok, G., Poeppel D.: The Cortical Organization of Speech Processing. Nat. Rev. Neurosci. 8(5), 393--402 (2007)
45. Skipper, J.L., van Wassenhove, V., Nusbaum, H.C., Small, S.L.: Hearing Lips and Seeing Voices: How Cortical Areas Supporting Speech Production Mediate Audiovisual Speech Perception. Cereb. Cortex 17(10), 2387--2399 (2007)

46. Fadiga, L., Craighero, L., Buccino, G., Rizzolatti, G.: Speech Listening Specifically Modulates the Excitability of Tongue Muscles: A TMS Study. Eur. J. Neurosci. 15, 399--402 (2002)
47. Watkins, K.E., Strafella, A.P., Paus T.: Seeing and Hearing Speech Excites the Motor System Involved in Speech Production. Neuropsychologia 41, 989--994 (2003)
48. Callan, D.E., Jones, J.A., Munhall, K., Callan, A.M., Kroos, C., Vatikiotis-Bateson, E.: Neural Processes Underlying Perceptual Enhancement by Visual Speech Gestures. NeuroReport 14(17), 2213--2218 (2003)
49. Skipper, J.I., Nusbaum, H.C., Small, S.L.: Lending a Helping Hand to Hearing: Another Motor Theory of Speech Perception. In: Arbib, M.A. (ed.), Action to Language Via the Mirror Neuron System, pp. 250-285. Cambridge University Press, Cambridge, NY (2006)
50. Kelso, J.A.S., Vatikiotis-Bateson, E., Saltzman, E., Kay, B.A.: A Qualitative Dynamic Analysis of Reiterant Speech Production: Phase Portraits, Kinematics, and Dynamic Modeling. J. Acoust. Soc. Am. 77(1), 266--280 (1985)
51. Summers, W.V.: Effects of Stress and Final-Consonant Voicing on Vowel Production: Articulatory and Acoustic Analyses. J. Acout. Soc. Am. 82(3), 847--863 (1987)
52. Vatikiotis-Bateson, E., Kelso, J.A.S.: Rhythm Type and Articulatory Dynamics in English, French and Japanese. J. Phonetics 21, 231--265 (1993)
53. De Jong, K.: The Supraglottal Articulation of Prominence in English: Linguistic Stress as Localized Hyperarticulation. J. Acoust. Soc. Am. 97(1), 491--504 (1995)
54. Harrington, J., Fletcher, J., Roberts, C.: Coarticulation and the Accented/Unaccented Distinction: Evidence from Jaw Movement Data. J. Phonetics 23, 305--322 (1995)
55. Lœvenbruck, H.: An Investigation of Articulatory Correlates of the Accentual Phrase in French. In: Proceedings of the 14th ICPhS, vol. 1, pp. 667--670 (1999)
56. Erickson, D., Maekawa, K., Hashi, M., Dang, J.: Some Articulatory and Acoustic Changes Associated with Emphasis in Spoken English. In: Proceedings of ICSLP 2000, vol. 3, pp. 247--250 (2000)
57. Lœvenbruck, H. Effets articulatoires de l'emphase contrastive sur la Phrase Accentuelle en français. In: Actes des XXIIIes Journées d'Etude sur la Parole JEP 2000, pp. 165--168 (2000)
58. Erickson, D.: Articulation of Extreme Formant Patterns for Emphasized Vowels. Phonetica 59, 134--149 (2002)
59. Keating, P., Baroni, M., Mattys, S., Scarborough, R., Alwan, A., Auer, E.T., Bernstein, L.E.: Optical Phonetics and Visual Perception of Lexical and Phrasal Stress in English. In: Proceedings of ICPhS 2003, pp. 2071--2074 (2003)
60. Cho, T.: Prosodic Strengthening and Featural Enhancement: Evidence from Acoustic and Articulatory Realizations of /a,i/ in English. J. Acoust. Soc. Am. 117(6), 3867--3878 (2005)
61. Beskow, J., Granström, B., House, D. Visual Correlates to Prominence in Several Expressive Modes. In: Proceedings of Interspeech 2006 – ICSLP, pp. 1272--1275 (2006)
62. Di Cristo, A.: Vers une modélisation de l'accentuation du français (deuxième partie). J. French Lang. Studies 10, 27--44 (2000)
63. Dahan, D., Bernard, J.-M.: Interspeaker Variability in Emphatic Accent Production in French. Lang. Speech 39(4), 341--374 (1996)
64. Di Cristo, A.: Intonation in French. In: Hirst, D., Di Cristo, A. (eds.), Intonation Systems: A Survey of Twenty Languages, pp. 195--218. Cambridge University Press, Cambridge, NY (1998)
65. Di Cristo, A., Jankowski, L.: Prosodic Organisation and Phrasing after Focus in French. In: Proceedings of ICPhS 1999, pp. 1565--1568 (1999)
66. Rossi, M.: La focalisation. In: L'intonation, le système du français: description et modélisation, Chap. II-6, pp. 116--128. Ophrys, Paris (1999)

67. Jun, S.-A., Fougeron, C.: A Phonological Model of French Intonation. In: Botinis, A. (ed.), Intonation: Analysis, Modelling and Technology, pp. 209--242. Kluwer Academic Publishers, Dordrecht (2000)
68. Delais-Roussarie, E., Rialland, A., Doetjes, J., Marandin, J.-M.: The Prosody of Post Focus Sequences in French. In: Proceedings of Speech Prosody 2002, pp. 239--242 (2002)
69. Dohen, M., Lœvenbruck, H.: Pre-focal Rephrasing, Focal Enhancement and Post-focal Deaccentuation in French. In: Proceedings of the 8th ICSLP, pp. 1313--1316 (2004)
70. Dohen, M., Lœvenbruck, H., Cathiard, M.-A., Schwartz, J.-L.: Visual Perception of Contrastive Focus in Reiterant French Speech. Speech Comm. 44, 155--172 (2004)
71. Dohen, M., Lœvenbruck, H.: Audiovisual Production and Perception of Contrastive Focus in French: A Multispeaker Study. In: Proceedings of Interspeech 2005, pp. 2413--2416 (2005)
72. Dohen, M., Lœvenbruck, H., Hill, H.: Visual Correlates of Prosodic Contrastive Focus in French: Description and Inter-Speaker Variabilities. In: Proceedings of Speech Prosody 2006, pp. 221--224 (2006)
73. Thompson, D.M.: On the Detection of Emphasis in Spoken Sentences by Means of Visual, Tactual, and Visual-Tactual Cues. J. Gen. Psychol. 11, 160--172 (1934)
74. Risberg, A., Agelfors, E.: On the Identification of Intonation Contours by Hearing Impaired Listeners. Speech Transmission Laboratory - Quarterly Progress Report and Status Report, 19(2-3), 51--61 (1978)
75. Risberg, A., Lubker, J.: Prosody and Speechreading. Speech Transmission Laboratory Quarterly Progress Report and Status Report 19(4), 1--16 (1978)
76. Bernstein, L.E, Eberhardt, S.P., Demorest, M.E.: Single-Channel Vibrotactile Supplements to Visual Perception of Intonation and Stress. J. Acoust. Soc. Am. 85(1), 397--405 (1989)
77. Krahmer, E., Swerts, M.: Perceiving Focus. In: Lee, C.-M. (ed.), Topic and Focus: A Cross-Linguistic Perspective, pp. 121--137. Kluwer, Dordrecht (2006)
78. Granström, B., House, D.: Audiovisual Representation of Prosody in Expressive Speech Communication. Speech Comm. 46, 473--484 (2005)
79. Dohen, M., Lœvenbruck, H.: Interaction of Audition and Vision for the Perception of Prosodic Contrastive Focus. Lang. Speech (in press)
80. Miller, G.A., Nicely, P.: An Analysis of Perceptual Confusions Among Some English Consonants. J. Acoust. Soc. Am. 27(2), 338--352 (1955)
81. Kendon, A.: Gesticulation and Speech: Two Aspects of the Process of Utterance. In: Key, M.R. (ed.), The Relationship of Verbal and Nonverbal Communication, pp. 207--227. Mouton, The Hague (1980)
82. McNeill, D.: Hand and Mind. University of Chicago Press, Chicago (1992)
83. Kendon, A.: Gesture. Annu. Rev. Anthropol. 26, 109--128 (1997)
84. Kita, S. (ed.): Pointing: Where Language, Culture, and Cognition Meet. Lawrence Erlbaum Associates, Hillsdale, NJ (2003)
85. MacNeilage, P.F.: The Frame/Content Theory of Evolution of Speech Production. Behav. Brain Sci. 21(4), 499--511 (1998)
86. MacNeilage, P.F., Davis, B.: On the Origin of Internal Structure and Word Forms. Science 288, 527--531 (2000)
87. Corballis, M.C.: From Mouth to Hand: Gesture, Speech, and the Evolution of Right-Handedness. Behav. Brain Sci. 26(2), 199--260 (2003)
88. Arbib, M.A.: The Evolving Mirror System: A Neural Basis for Language Readiness. In: Christiansen, M., Kirby, S. (eds.), Language Evolution: The States of the Art, pp. 182--200. Oxford University Press, Oxford, UK (2003)
89. Arbib, M.A.: From Monkey-Like Action Recognition to Human Language: An Evolutionary Framework for Neurolinguistics. Behav. Brain Sci. 28(2), 105--124 (2005)
90. Holden, G.: The Origin of Speech. Science, 303, 1316--1319 (2004)

91.Iverson, J., Thelen, E.: The Hand Leads the Mouth in Ontogenesis Too. Behav. Brain Sci. 26(2), 225--226 (2003)
92.Iverson, J., Goldin-Meadow, S.: Gesture Paves the Way for Language Development. Psychol. Sci. 16, 368--371 (2005)
93.Ozcaliskan, S., Goldin-Meadow, S.: Gesture is at the Cutting Edge of Early Language Development. Cognition 96, 101--113 (2005)
94.Goldin-Meadow, S., Butcher, C.: Pointing Toward Two-Word Speech in Young Children. In: Kita, S. (ed.), Pointing: Where Language, Culture, and Cognition Meet, pp. 85--107. Lawrence Erlbaum Associates, Hillsdale, N.J. (2003)
95.Iverson, J., Thelen, E.: Hand, Mouth, and Brain: The Dynamic Emergence of Speech and Gesture. J. Conscious. Stud. 6, 19--40 (1999)
96.Iverson, J., Goldin-Meadow, S.: Why People Gesture as They Speak. Nature 396, 228 (1998)
97.Butterworth, G.: Pointing is the Royal Road to Language for Babies. In: Kita, S. (ed.), Pointing: Where Language, Culture, and Cognition Meet, pp. 9--33. Lawrence Erlbaum Associates, Hillsdale, NJ (2003)
98.Pizzuto, E., Capobianco, M., Devescovi, A.: Gestural-Vocal Deixis and Representational Skills in Early Language Development. Interaction Studies 6(2), 223--252 (2005)
99.Volterra, V., Caselli, M.C., Capirci, O., Pizzuto, E.: Gesture and the Emergence and Development of Language. In: Tomasello, M., Slobin, D. (eds.), Elizabeth Bates: A Festschrift, pp. 3--40. Lawrence Erlbaum Associates, Mahwah, NJ (2005)
100.Hollender, D.: Interference Between a Vocal and a Manual Response to the Same Stimulus. In: Stelmach, G., Requin, J. (eds.), Tutorials in Motor Behavior, pp. 421--432. North-Holland Amsterdam (1980)
101.Castiello, U., Paulignan, Y., Jeannerod, M.: Temporal Dissociation of Motor Responses and Subjective Awareness. Brain 114(6), 2639--2655 (1991)
102.Fagot, C., Pashler, H.: Making Two Responses to a Single Object: Implications for the Central Attentional Bottleneck. J. Exp. Psychol. 18, 1058--1079 (1992)
103.Rochet-Capellan, A.: De la substance à la forme: rôle des contraintes motrices orofaciales et brachiomanuelles de la parole dans l'émergence du langage. PhD Thesis, Cognitive Sciences, Institut National Polytechnique de Grenoble, France (2007)
104.Rochet-Capellan, A., Schwartz, J.-L., Laboissière, R., Galván, A.: The Speech Focus Position Effect on Jaw-Finger Coordination in a Pointing Task. J. Speech Lang. Hear. Res. (In Press)
105.Levelt, W.J.M., Richardson, G., Heij, W.L.: Pointing and Voicing in Deictic Expressions. J Mem. Lang. 24, 133--164 (1985)
106.Feyereisen, P.: The Competition Between Gesture and Speech Production in Dual-Task Paradigms. J. Mem. Lang. 36(1), 13--33 (1997)
107.Ducey-Kaufmann, V.: Le cadre de la parole et le cadre du signe : un rendez-vous développemental. PhD Thesis, Language Sciences, Stendhal University, Grenoble, France (2007)
108.Ducey-Kaufmann, V., Abry, C., Vilain, C.: When the Speech Frame Meets the Sign Frame in a Developmental Framework. In: Emergence of Language Abilities (2007)
109.Rochet-Capellan, A., Schwartz, J.-L., Laboissière, R., Galván, A.: Two CV Syllables for One Pointing Gesture as an Optimal Ratio for Jaw-Arm Coordination in a Deictic Task: A Preliminary Study. In: Proceedings of EuroCogSci07, pp. 608--613 (2007)