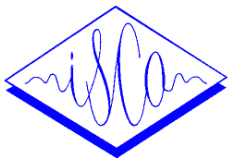

ESCA Workshop on
Audio-Visual Speech Processing
Rhodes, Greece
September 26-27, 2997

MICRO- AND MACRO-BIMODALITY

Emanuela Magno Caldognetto* and Isabella Poggi**

* Centro di Studio per le Ricerche di Fonetica del CNR
Via G. Anghinoni, 10 - 35121 Padova, ITALY
FAX: +39/49/8274416 - E-mail: Magno@csrf.pd.cnr.it

** Università Roma 3
Dipartimento di Scienza del linguaggio
Via del Castro Pretorio, 20 - Roma, ITALY
FAX +39/6/4957333 - E-mail: poggi@uniroma3.it

ABSTRACT

In audio-visual communication, acoustic and optical modalities are simultaneously at work and their respective signals are intertwined in complex ways. Two kinds of bimodality are distinguished: micro-bimodality and macro-bimodality, the relationship of these optical and acoustic signals with the optical signals performed by hand gestures, facial expression and body movements. Some differences and similarities are described between the two kinds of bimodality on the production and perception side, and a model of context is proposed to account for how acoustic and optical modalities are arranged in the planning and understanding of bimodal communication.

1. INTRODUCTION

In face to face interaction, as well in audio visual media, communication is bimodal in that different kinds of information get across audio and visual modalities. But we may distinguish two senses of bimodality. First, a **micro-bimodality** (from now on, MIB, see [1]), the relationship between the optical signals delivered by the visible articulatory movements (lips, jaw, teeth, tongue) and the co-produced acoustic signal, determined by the phonetic events that perform the phonological intent. Second, a **macro-bimodality** (MAB, see [2]), the relationship among these optical and acoustic signals and the optical signals performed by all other bodily signals simultaneously produced: hand gestures, facial expression and body movements.
The two kinds of bimodalities work on different time spans, and the kinds of information they provide differ in their semiotic status, function, internal organization. The interaction among these kinds of information in both levels of bimodality is highly complex, and for production and comprehension one has to account for both integration and segregation of signals. Moreover, the study of these facts raises issues about cortical localization, neuro-anathomical and neuro-physiological specialization, modal vs. amodal internal representation of information [3], [4].

Though the two levels of bimodality are traditionally studied in quite distant research fields, opposition among description methods may be reduced by proposing a common notion of GESTURE: in MIB an articulatory gesture (according to the articulatory phonology, see [5]) is a distinctive dynamic unit that may substitute for discrete segments of linear phonology; in MAB, a hand gesture is a unitary sequence of hand and/or arm movements (onset, stroke, recovery [6]). (We now confine ourselves to hand gestures, but body movements and facial expression could be defined in a like way).
One more analogy can be found in present studies on MIB and MAB, where for both it was independently adopted a method called the SCORE, a description that allows to compare parallel signals performed in different modalities, while visualizing their temporal relationships. A particular kind of "score" is used in analyzing the relationship among articulatory gestures [5]; another one is used to show the interaction of the acoustic output with hand, face and body gestures [7].
We mentioned some general analogies between MIB and MAB; now we come to more specific differences and resemblances.

2. THE SAME VERSUS DIFFERENT EVENTS

A first difference is that signals in MIB are produced out of one and the same action, while in MAB they are different actions that may even be independent of each other. The goal of moving one's lips, tongue, teeth is (in normal conditions) to produce acoustic signals, and the fact that mouth movements may also be seen might be viewed as a side effect produced while producing sounds.
In macro-bimodal production, instead, voice and gesture are performed by two separate motor programs planned on a hierarchically higher level. The general goal is to convey a global meaning, and the same or different parts or aspects of that meaning are distributed across modalities. In this view, acoustic and optic signals perform parts of a common task.

3. TYPES OF INFORMATION CONVEYED

Three types of information are transmitted during communication.

1. **Speaker's identity**: information on idiosyncratic traits of the Speaker: the anatomo-physiological features of sex and age, and, perhaps, personality: those characters that make part of one's identity and that one cannot voluntarily change.

2. **Speaker's Mind**: information on the Speaker's goals, beliefs, evaluations, emotions [8]; these corrispond to the kind of information traditionally classified as paralinguistic, but we do not agree in distinguishing them from linguistic information, if the difference considered is of not being systematic and rule-governed. In the last thirty years, Pragmatics and Cognitive Science have shown that the Speaker's goals and beliefs are part of "linguistic proper"; and also emotion may be and is in fact expressed in a systematic way by both verbal and non verbal, phonic and kinetic communication.

3. **Content**: Information on the World: objects, persons, places, events, both concrete and abstract.

Speaker's identity information is one that leaks from the Speaker's communication without, sometimes against, Speaker's intention. Content and Speaker's Mind information make part of what the Speaker specifically intends to communicate. Now, the Types of information n.1. and 2. may be equally provided in MIB by voice and lip shape, and in MAB by voice and hands, face, body. I can tell whether the Speaker is a female, and how old is she (Information on Speaker's identity) not only when hearing and seeing a whole crunch of vocal and gestural communication (MAB), but also by hearing a single sound or seeing a single movement of her mouth (MIB). This holds also for information on the Speaker's emotions or communicative goals: from how she utters a vowel or from her smiling lips I can tell if she is ironic.But whereas also MIB can provide information on the Speaker's mind and identity, it is only in MAB that we get proper Content information, information on the World.

4. INTER-MODALITY RELATIONSHIPS

At both the micro- and the macro- level, signals in different modalities may bear particular relationships to each other. In MAB, the relationships among different signals - say, a word and a gesture - are quite complex: the global meaning depends on the merging of information provided by different signals at a higher level.

In the "score" of multimodal [7], a procedure to analyze the complex intertwining of messages in MAB, five production modalities are taken into account: **verbal prosodic, gestural, facial**, and **bodily**. Each signal of each modality is physically described as to its acoustic or visual features, then its meaning is glossed and classified according to a semantic typology, and to function (semantic relationship) of simultaneous signals to one another **repetition**, (or it conveys the same meaning); **addition**, (it adds complementary congruent information); **substitution**, (gives information not provided in the other modality); **contradiction**, (it provides contrasting information); finally two signals may be **independent** from one another, as they take part in different communicative plans.

The global planning of MAB in real discourse may be very complex. For example, the additional meaning conveyed by a gesture in some cases concerns the **discourse topics**, as when the speaker gesturally depicts a triangular shape while talking of "a large balcony"; but in other cases it is of a metatextual kind, concerning the **discourse structure**: as when the Speaker counts on his fingers while listing the points of his argument.

Let us now see whether analogues of all these functions may be found in MIB.

In Italian, a case of **repetition** between acoustic and optical signal is the production of a labiodental phone as [f]. The Hearer may both hear the consonant noise and see the Speaker's teeth on lower lip: both signals tell a labiodental is being produced, thus giving rise to an "audio-visual synergism"[9], [10]. Bilabial stop [p] is a case of **additive** function, because of an asynchronism between visible articulatory and acoustic signal. While in the acoustic signal there is a silent phase, one common to all stops ([p], [t], [k]), visual information specifies which stop it is. One more example of **additive** function: suppose a Speaker is protruding lips in uttering a [u], but with protrusion longer than usual: I can get the additional information that the Speaker is or wants to look quite snobbish.

Lip-reading is a typical case where the acoustic signal is **substituted** by the optical one.

The only natural case of **contradictory** function seems to be ventriloquism; ill-made film dubbing is a technological case, while the McGurk effect [11], [12], [13] is a weel-known experimental example.

Finally, a case of **independence** might be yawning or chewing, where a parasite mouth movement is superimposed to phonetic articulation, with a possible destorting effect.

5. PLANNING PRODUCTION IN MAB AND MIB

Given these similarities and differences between the two kinds of bimodality on a descriptive level, what may be the implications for models of production and comprehension? A first question is how are the two kinds of bimodalities generated. Does the Speaker first generate an amodal signal whose information is then distributed and specified through the different modalities; and, if this is the case, which might be the criterion of such a distribution: how to decide what to communicate through acoustic and what through optical signals? In our hypothesis, first it comes the macro planning on how to convey a global meaning, and then, if and where a decision is made to use (only or also) acoustic com-

munication in the MAB, planning on the micro bimodality occurs.

6. A MODEL OF CONTEXT

As for any action, communicative behavior is determined by both our goals and the world conditions at hand. Starting from this principle, this is our hypothesis of how context determines Speaker's planning.

Speaker has a global idea to communicate, a **communicative Goal**, that is a goal of communicating something to a particular Hearer; within Speaker's communicative goal we may distinguish Content Information (what the Speaker is speaking of) and Information on the Speaker's mind (one's goal, why is one communicating that, one's beliefs, evaluations and emotions on what one is saying). Now, in order to decide how to communicate both mental and referential, both concrete and abstract information, Speaker has to take into account one's **communicative possibilities**.

These include on the one side **Speaker's internal capacities**: his/her linguistic competence (for instance being a foreigner, not mastering a language completely) and possible pathologic conditions of a transitory or permanent kind (say, slips of the tongue or aphasia).

On the other side we may include **external conditions**, the environment physical and cognitive features. Among the **physical constraints**, Speaker must wonder whether communication is face to face or at distance, whether Hearer is in the same spatio-temporal situation, and what are the available modalities: only acoustic (e.g. on the phone), only visual (through a glass or across a road), or both.

The **cognitive constraints** depend on the Model the Speaker has of the Hearer, that includes: Hearer's **linguistic competence** (whether deaf or foreigner for instance) and **inference capacity** (a child, or adult or so); and the **social relationship** between Speaker and Hearer (dimensions of power vs solidarity, how familiar they are to each other and so on).

Let us see which aspects in MAB planning depend on which of the constraints listed before. On the basis of one's *communicative goals*, Speaker decides which relative weights to attribute to different parts of one's communicative act: one may decide for instance what counts as topic and what as comment in a sentence ("rhematic factors", in terms of [14]); or which parts of an argument are most important to one's persuasive goals. That is why one may decide to stress a word or repeat a sentence, or to convey the same concept by both a word and a gesture. On the basis of *internal capacities*, Speaker may substitute a word by an iconic gesture when speaking a foreign language, or for a "tip of the tongue" case, or if s/he is aphasic [7]. On the basis of *physical environmental constraints*, Speaker may decide to lean on only one modality or to distribute meanings across acoustic and optical modalities. In this case, distribution may be once more determined by goal constraints. Finally, on the basis of the Hearer Model, particularly *Hearer's understanding capacities*, one decides to be redundant or to use a more iconic language (say, gesture vs. word). This is why one may tend to use more gesture and body communication in talking to a foreigner or a child. Another decision node is the level of *familiarity between Hearer and Speaker*, that may induce not to use gestures, if sanctioned as too informal, or to use just some very formal or solemn ones.

When decisions are made on a macro level, the micro adaptive variability is triggered (see the "Hyper&hypo" speech theory: [15]). Here, as we said, the optical signal is but a side-issue of the acoustic production, and not directly decided for itself, even if it may take advantage of how the acoustic signal is produced: as I utter words with slow and amplified lip movements to provide a clearer signal to a foreigner, the effect of a better lip-reading is not directly decided for itself but is anyway determined by clearer pronounciation. Yet, sometimes, the relationship main action - side issue is the other way around: when on the basis of the macro planning Speaker decides that, because of noisy environment or of hearing impairment, optical signals are greatly of use to Hearer, the primary goal may be scanning words to produce a better visible signal, and the acoustic rendition of words may be changed accordingly.

7. PROBLEMS ON THE HEARER SIDE

On the side of perception and comprehension in MIB and MAB, the issue of dominant modality is often raised in perception that for MIB explored hypotheses both of dominant modality and of dynamic amodal representation [12], [13], [16], [17], [18]. Here, far from proposing a model of comprehension, we make two alternative hypotheses as to how modalities might be arranged in comprehension. At the macro-level, a first hypothesis is that one modality (say, the acoustic one) is privileged in that, by default, it is considered most relevant by Hearer, while the other gets relevant only as the first does not provide complete or plausible information. Such a hypothesis, though, could only account for cases of **substitutive** or **repetitive** function (just in case of lacking or uncertain information, should I resort to signals in other modalities) but not for cases of **additive** or **contradictory** function. Sometimes we do feel that Speaker, while verbally saying X, in fact looks like saying not-X. It is more plausible, then, that informations in all modalities are caught (at most optical signals might be caught at a lower level of awareness). Once processed, they are integrated to form a global meaning, but at the same time they are segregated, in that Hearer may keep track of which modality they passed through, compare signals in different modalities, and assess their respective function for further processing.

Thus, we may suppose that when a signal has a function of **addition** or **substitution** for another one, we take up

the different pieces of information, by whatever modality they are conveyed, and make up a global meaning. In case of repetitive function, we may in addition wonder why should the Speaker be so redundant: does s/he think we are a bit dull, or is the message so important that Speaker does not want to risk a misunderstanding?
Finally, in case of **contradictory** signals we might guess where is the truth. Suppose in an interview on intercultural relationships the interviewed shows very tolerant in words, but his spatial behavior abounds in distance signals; we should decide which modality to believe, and only by inference from contextual or world knowledge might we find an answer.
As for MIB, if, as we said, the optical signal is but a side effect of the acoustic one, not one planned independently, and if the only cases of contradiction are experimental or very rare natural cases, then we should make use of the optical signal only in case the acoustic is degraded or lacking: that is, only as it has a substitutive function, like in lip-reading. More generally for MIB, decision to pay attention also or primarily to optical signals might be triggered contextually: if in a noisy room, I look at Speaker's lips very attentively.

8. CONCLUSION

In audio-visual communication, acoustic and optical modalities are symultaneously at work and their respective signals are intertwined in complex ways. We distinguished two kinds of bimodality, MIB and MAB, describing some differences and similarities between them on the production and perception side, and proposing a model of context that can account for how acoustic and optical modalities are arranged in the planning and understanding of bimodal communication.

REFERENCES

[1] D.G.Stork and M.E.Hennecke (Eds.), *Speechreading by humans and machines*, NATO ASI Series, Springer-Verlag ,Berlin Heidelberg, 1996.

[2] L.Messing (Ed.), *Proceedings of the Workshop on the Integration of Gesture and Language in Speech*, Applied Science and Engineering Laboratories, Newark and Wilmington, Del.

[3] R. Campbell, *Seeing Brain Reading Speech: a Review and Speculations*, in Stork & Hennecke, 115-134.

[4] De Gelder B., Bertelson P.and Vroomen J. (1996), "Aspects of modality in audio-visual processes", in Stork & Hennecke, 179-191.

[5] Browman C.P. and Goldstein L. (1992), "Articulatory phonology:an overview", *Phonetica* 49, 155-180.

[6] A. Kendon, *Gesticulation and Speech: two Aspects of the Process of Utterance*, in M. Ritchie Key (Ed.), "The Relationship of Verbal and Non-verbal Communication", Mouton Publ., The Hague, 1980, 207-227.

[7] Poggi I. and Magno Caldognetto E. (1996), "A score for the analysis of gestures in multimodal communication", in L.Messing (Ed.), 235-244.

[8] Poggi I. (1996), "Mind Markers". Poster presented at the 5th International Pragmatics Conference, Mexico City, July 4-9, 1996.

[9] Erber N.P. (1975), "Auditory-visual perception of speech", *Journal of Speech and Hearing Disorders* 40, 481-492.

[10] Benoit C., Guiard-Marigny T, Le Goff B. and Adjoudany A. (1996), "Which components of the face do humans and machines best speechread?", in D.G.Stork and M.E.Hennecke, 315-328.

[11] McGurk H. and MacDonald J.W. (1976), "Hearing lips and seeing voices", *Nature* 264,746-748.

[12] Summerfield A. Q. (1987), "Some preliminaries to a comprehensive account of audio-visual speech perception", in B.Dodd and R. Campbell (Eds.), *Hearing by eye:the psychology of lip-reading*, Lawrence Erlbaum Ass.Publ., Hillsdale,3-51.

[13] Massaro D. W. (1987), "Speech perception by ear and eye", in B.Dodd and R. Campbell (Eds.), *Hearing by eye:the psychology of lip-reading*, Lawrence Erlbaum Ass.Publ., Hillsdale, 53-83.

[14] Cassell J. and Prevost S. (1996), "Distribution of semantic features across speech and gesture by humans and machines", in L.Messing.

[15] B. Lindblom, *Adaptive Variability and Absolute Constancy in Speech Signals: two Themes in the Quest for Phonetic Invariance*, Proc. of the Xith ICPhS (August 1-7, 1987), Tallin, 1987, Vol. 3, 9-18.

[16] J. Robert-Ribes, M. Piquemal, J. Schwartz, P. Escudier, *Exploiting Sensor Fusion Architectures and Stimuli Complementarity in AV Speech Secognition*, in Storck and Hennecke.

[17] A. M. Liberman and I. G. Mattingly, *The Motor Theory of Speech Perception Revised*, Cognition 21, 1985, 1-36.

[18] C. A. Fowler, *An Event Approach to the Study of Speech Perception from a Direct-realistic Perspective*, Journal of Phonetics 4, 1986, 3-28.