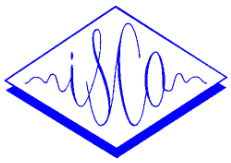

ESCA Workshop on
Audio-Visual Speech Processing
Rhodes, Greece
September 26-27, 2997

A METHODOLOGY TO QUANTIFY THE CONTRIBUTION OF VISUAL AND PROSODIC INFORMATION TO THE PROCESS OF SPEECH COMPREHENSION

Loredana Cerrato, Federico Albano Leoni**, Andrea Paoloni**
*Fondazione Ugo Bordoni, Via B. Castiglione , 59 00142 Roma Italy
Tel. +39 6 54803351 Fax +39 6 54804405 email Pao@fub.it
*CIRASS, Università di Napoli,
Via Porta di Massa 1, Napoli, Italy
Tel. +39 81 5420280 Fax +39 81 5420370 e-mail fealbano@unina.it

ABSTRACT

We will report in this paper the results of a series of comprehension tests designed in order to investigate the contribution of visual and prosodic information to the process of comprehension of conversational speech.
Particular attention will be paid to the methodology used.
Our research follows the new interest in studying the integration of various sources of information in the process of speech perception and approaches the problem from the point of view of general comprehension.

1 INTRODUCTION

It has been known since the 1950s by Sumby and Pollack [1] that listeners show substantial increases in intelligibility of speech when they view the speakers' face; more recently Remez et alii [2], and Summerfield [3] argued for a theory of speech processing which takes into account the integration of various sources of information (auditory, visual, tactile).
It has also been shown that a multimodal speech signal (auditory + visual) is extremely robust and informative and it provides information that perceivers are able to exploit during perceptual analysis [4, 5]. Moreover Pisoni et alii [6] showed that the addition of visual information in the stimulus display about the speaker's articulation, affects the efficiency of initial encoding operations at the time of perception, and also results in a more detailed and robust presentations of the stimulus event in memory.
Most of the experiments reported in literature tried to evaluate the support of phonological information visually conveyed by articulatory movement [7] using,as speech material, corpora made of syllables or short words.
Our experiment aims at a confirmation of the results obtained so far, this time using conversational speech material rather than a built up corpus.
Our experiment follows the original paradigm proposed by Lindblom in his H & H Theory which states not only that speakers are able to adapt their production along a continuum which ranges from hyper-speech to hypo-speech according to the communicative situation context, but also that listeners adjust their perceptual framework to fit expectations about the incoming speech signal, often anticipating it from their knowledge of the speaker, the context of the conversation and their general knowledge. As a consequence, listeners are able to understand speech also in noisy conditions as it contains many redundant cues (among which are visual cues), which can support perception in adverse conditions.[8]

2 MATERIAL AND METHODS

The audio-visual sample used in the experiment consisted of 6 minutes of conversation among some Italian people about a particular topic, held during a TV talk show. We chose real audio/video conversational speech material rather than a built up corpus because we believe that real spontaneous speech is representative of actual speech and that nobody goes round uttering phonemes or words in isolation!
36 university students (9 male and 27 female) aged around 20, served as subjects for our experiment, which consisted of three phases:
in the first phase we presented 12 subjects (with no sight and hearing pathologies) with the multimodal speech signal (auditory + visual);
in the second phase we submitted another group of 12 subjects to the uni-modal speech sample (auditory) without the integration of the visual information; gestures, postures and facial expression, on which the speakers put part of the communicative intentions, are in fact, cut out in this type of transmission.
In the last phase we submitted another group of 12 to the sample material, which this time was presented under the form of an orthographic transcription without punctuation, without prosodic information and of course without any speech or visual signals. We decided for a literally transcription of the oral text with the aim of producing a text without any suprasegmental feature (i.e. intonation, pauses, rhythm, quantity, phonation types, speed of elocution and gestures, postures and facial expressions). Suprasegmental features undertake an important role in speech perception and understanding: the basic functions of prosody are to segment and to

highlight; cues in rhythm and intonation patterns notify the listener of major syntactic boundaries, which help the listener mentally process speech units smaller than the entire sentence; the alternation of stressed and unstressed syllables identifies the words that the speaker considers important to understand the speech message and also helps in word comprehension via placement of lexical stress. Moreover prosody signals other aspects of syntactic structure: in many languages a question requesting a yes/no answer from a listener ends with an intonation rise. Intonation can also signal whether the clause is main or a subordinate, and highlight he functions of words.
Cues to the state of the speaker; attitudes and emotions are primarily signalled through the features related to phonation types (falsetto, breathy voice, creaky voice). [9]
After the presentation of the sample material the subjects of the three groups were required to answer a questionnaire relative to the presented situation.

2.1 The questionnaire

The questionnaire we created was modelled on the instruction of those used in the didactic of languages and aimed at the assessment of the ability of the subjects to decode the conversation.
Decoding a conversation involves a series of abilities [10, 11]:
1) the ability to reconstruct the communicative situation in which the speech act takes place,
2) the ability to understand the topic of the conversation (i.e.what is talked about/discussed),
3) the ability to understand the roles exhibited by the interlocutors in the conversation,
4) the ability to understand (and remember) the different points of view expressed by the different interlocutors about the topic of conversation.
A translation of the questionnaire is reported at the end of the paper.

The questionnaire, which was formulated in order to be adequate for the three phases of the test, was structured in three main parts:

Part A consists of questions with a multiple choice answer, which aimed at the assessment of the ability of the subjects to reconstruct the communicative situation in which the conversation took place (A1, A2,) and to understand the topic of the conversation (A3),.
Part B consists of a matching task, which aimed at the assessment of the ability of the subjects to assign the correct role to each speaker in the conversation.
Part C consists of questions with a true/false answer, which aimed at the assessment of the ability to understand (and remember) the different points of view and expressed by the different interlocutors about the topic of conversation. A translation of the questionnaire follows:

A.Answer the following questions choosing the most appropriate answer

1)In your opinion in which situation is the conversation held?
a) among friends
b) between lawyer and clients
c) during a talk show

2) In your opinion where does the conversation take place?
a) in a private house
b) in a legal office
c) in a tv studio

3) In your opinion what is the main topic of conversation?
a) Pacciani case,
b) chronic diseases,
c) delay,
d) friendship
e) work

B Assign to each of the speakers on of the following roles:

defendant	Maria,
defender	Lucia
moderator	Olga
spectator	Franca
spectator	U
spectator	S
injured party	D

C. Judge the following statements true or false

Lucia and Olga are sisters ❑

Lucia and Olga had an argument ❑

Olga arrived late at a work appointment with Lucia ❑

Olga is a very punctual person ❑

Olga arrived late to annoy Lucia ❑

Maria speaks about her dog ❑

Uomo agrees that delay is an evidence of lack of respect for people waiting ❑

Franca states that also Pacciani is a latecomer ❑

Lucia is a very serious and precise person❑

Olga went to the Mass at midday ❑

Franca speaks Polish ❑

Uomo agrees with latecomers ❑

Franca is kleptomaniac ❑

Lucia speaks very little ❑

Olga is sorry ❑

If Olga goes out an hour earlier she gets to the appointment late as well ❑

in appendix.

3 ANALYSIS OF THE RESULTS

The questionnaire was scored assigning one point to each correct answer for each part of the questionnaire. The higher the score, the better the comprehension of the text.
The results, expressed in percentage, are plotted on a histogram shown in fig. n.1.

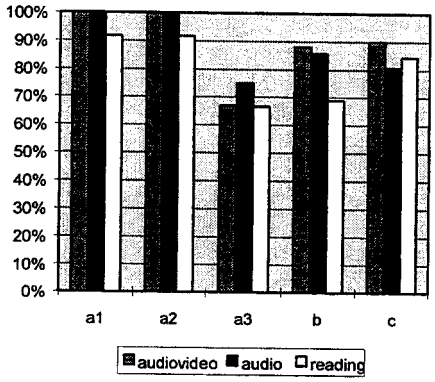

Fig. 1 Percentage of correct answer given to each series of questions (a1, a2, a3, b, c) for the three tasks (audio-video, audio, reading).

The graph reads that:
-the ability of the subject to reconstruct the communicative situation in which the conversation takes place is basically 100% for all the three tests (only in the reading group one of the subjects didn't give the correct answers);
-the ability of the subjects to understand the topic of conversation shows values smaller than the ones for the other abilities, but the wrong answers are all the same (all the subjects who gave a wrong answer chose the topic "friendship" instead of the correct topic "delay") and this may depend on the fact that the defendant and the injured party claimed to be were close friends.
-The ability of the subjects to assign the right role to each of the speaker shows results which are consistent with general expectation in that the score decreases with the impoverishment of the channel of transmissions.
-The ability of the subjects to understand and to remember the text shows some surprising results:
while the percentage of correct answers decreases passing from bimodal perception to unimodal perception, there is a slight increase in the score for the reading group, where instead we expected to get a further decrease of the percentage because of the lack of prosodic information. This result, which is apparently inconsistent with the original hypothesis, shows its consistency when we look at the values of the mean of correct answers and its standard deviation shown in table n1.

	mean of correct answers on 16	standard deviation
audio-video	14.41	1.08
audio	13.08	2.02
reading	13.5	2.57

Table 1 Mean of correct answers to the questionnaire and standard deviation.

If on one side the value of the mean of correct answers for the reading group is slightly higher than that for the audio-group, on the other side we can see that the standard deviation is higher for the reading group compared to the audio group and this is because among the readers the percentage of correct answers spans from 50% to 100%.

4. DISCUSSION

Our results lead to a confirmation of the original hypothesis that the comprehension of conversational speech decreases, passing from bimodal perception to unimodal perception. Making a correct estimate of the contribution of visual information to the process of comprehension of conversational speech is quite hard, though, as the problems that pertain to the comprehension of conversational speech are quite different from those concerning the perception of phonemes; as they imply the intervention of the cognitive system of the listener and the interaction of different sources of information.
To explain the slight improvement in the score of the reading task, for which, on the contrary, we were expecting to see a further decrease of the percentage of correct answers to the questionnaire, we might try to explain the results, bearing in mind that a written text cannot be considered a transcription of the oral text because the structural and formal characteristics of written texts are different from those of oral texts. Secondly, the speech signal and the graphic signal are supposedly perceived in two different ways: the spatial separation of the words in the written text favours a linguistic analysis in terms of discrete units, avoiding all the coarticulation phenomena typical of oral production; moreover, print emphasises the characteristics of persistency and stability of the written text as opposed to the dinamicity and temporariness of the phonic substance.

Finally, reading a message rather that listening to it, favours the effects of feedback and the so-called effect of garden path, which helps the reader to reconstruct the original prosodic pattern and to better process and memorise the message.
We might then conclude that the methodology we tested is suitable to verify and roughly quantify the contribution of visual information to the process of speech comprehension, while the reading task is not befitting to evaluate the contribution of prosodic information to the process of speech comprehension as it is impossible to consider the written text as a simple transcription of the oral text deprived of the suprasegmental information. In order to investigate the contribution of prosodic information to the process of comprehension of speech, it might be advisable, in next experiments,to generate a synthesis of the original text depriving it of all the suprasegmental features.

5 REFERENCES

[1] Sumby & Pollack 1954 Visual contribution to speech intelligibility in noise JASA 26, pp.212-215..

[2] Remez R.E., Rubin P.E., Pisoni B.B., Carrel T.D., 1981, Speech perception without traditional speech cues" Sci. 212, 947-950

[3] Summerfield A.Q., 1981, Some preliminaries to a comprehensive account of audio-visual speech perception, in B. Dodd & R. Campbell (Eds.) Hearing by eye: The psychology of lip-reading, Lawrence Erlbaum , London, pp.3-51.

[4] Saldana H, Pisoni D., Audio-Visual speech perception without speech cues, in ICSLP 1996 Philadelphia

[5] Le Goff B, Guiard-Marigny, Benoît C. 1995,Read my lips..and my jaw! How intelligible are the components of a speaker's face? In Proceedings Esca Eurospeech 1995 pp.291-294

[6] Pisoni, D.B., Saldana H.,M., Sheffert S.,M., Multi-modal encoding of speech in memory: a first report, in ICSLP 1996 Philadelphia

[7] Magno Caldognetto E., Vagges K., Il riconoscimento visivo dei movimenti articolatori da parte di soggetti normali e ipoacusici, in Scritti in Onore di Croatto L.- Centro di Studi pe le Ricerche di Fonetica del C.N.R Padova, 1990, Litocenter Padova.

[8] Lindblom B., 1996 "*Role of artiulation in speech perception: clues from production*", JASA, 99, 3 p.1683-1692

[9] Bertinetto P.M., Magno Caldognetto E. *Ritmo e Intonazione,* in Sobrero A.(a c. di) Introduzione all'italiano contemporaneo. Le strutture, Laterza Bari 1993

[10] Fraunfelder U.,H., Tyler L.,1987, "*The process of spoken recognition: an introduction*", Cognition 25, p.1-20.

[11] Hymes D., 1972, On Communicative Competence, in Pride J. B, Folmes J (eds) Socolinguistics, Harmondsworth, Penguin London.