Integrating Audio and Visual Information for Modelling Communicative Behaviours Perceived as Different

Michelina Savino[(1)], Laura Scivetti, Mario Refice[(2)]
[(1)]Dept. of Psychology, University of Bari, ITALY
[(2)]DEE, Human-Machine Interaction Systems Lab, Polytechnics of Bari, ITALY
E-mail: m.savino@psico.uniba.it, lscivetti@yahoo.it, refice@poliba.it

Abstract

In human face-to-face interaction, participants can rely on a number of audio-visual information for interpreting interlocutors' communicative intentions, such information strongly contributing to the successfulness of communication. Modelling these typical human abilities represents a main objective in human communication research, including technological applications like human-machine interaction. In this pilot study we explore the possibility of using audio-visual parameters for describing/measuring the differences perceived in interlocutor's communicative behaviours. Preliminary results derived from the multimodal analysis of a single subject seem to indicate that measuring the distribution of some prosodic and hand gesture events which are temporally co-occurring contribute to the accounting of such perceived differences. Moreover, as far as gesture events are concerned, it has been observed that relevant information are not simply to be found in the occurences of single gestures, but mainly in some gesture modalities (for example, "single stroke" vs "multiple stroke" gestures, one-hand vs both-hands gestures, etc…).
In this paper we also introduce and describe a software package, ViSuite, we developed for multimodal processing and used for the work described in this paper.

1. Introduction

In current research on human communication, it is widely acknowledged the co-expressive nature of verbal and non-verbal channels (see seminal works of Kendon, 1972; 1980, McNeill 1992). Therefore, in human face-to-face interaction, participants can rely on a number of audio-visual information for interpreting interlocutors' communicative intentions, such information strongly contributing to the successfulness of communication. Such audio-visual cues are used by listeners not only for recognising specific interlocutor's communicative strategies, but also for identifying his/her emotional state. Modelling these typical human abilities represents a main objective in human communication research, including technological applications like the development of human-machine interaction systems (Cassell and Stone, 1999).
A specific question we would like to arise is the following: when listeners judge two communicative behaviours as different, how can we quantitatively estimate such differences in parametric terms? Since human communication is intrinsically multimodal, can "multimodal" or "audio-visual" parameters be determined which can contribute – or be more effective – to describe perceived differences in communicative behaviours?
In this paper we report on an exploratory study we carried out, aiming to provide some very preliminary answers to these questions. We analysed some prosodic event types and speech-accompanying hand gestures produced by one subject in two different moments of the same video, where the related communicative behaviours were perceived as different. We were particularly interested in measuring prosodic and gestural events which were temporally co-occurring, and verifying how they could contribute to the description of such perceived differences.
This pilot study represents the preliminary step of a larger project aiming to collecting and analysing a corpus of similar audio-video materials in Italian.

2. Audio-Video Materials

The audiovisual materials analysed consist of two excerpts – each of them having the same duration of approx. 1 min – taken from a very popular Italian talk show where the politician Silvio Berlusconi, at that time Prime Minister of the Italian goverment, was interviewed during the elections campaign in Italy in 2005 (elections which eventually resulted in Berlusconi's defeat).
In the first excerpt, Berlusconi is interviewed only by the journalist conducing the talk show, who is notoriously "sympathetic" with the Prime Minister, asking questions not related to the achievements of Berlusconi government, and showing himself overtly as being "by the Prime Minister's side". As a consequence, Berlusconi looks relaxed in answering the questions, and his communicative behaviour is perceived as non-aggressive, controlled and non-emphatic.
In the second excerpt, the situation changes, as this time the Prime Minister has to answer to politically "hot" questions – concerning his government's (lack of) achievements, criticism-biased questions posed by non-sympathetic guest journalists sitting in the audience. In this second excerpt, the subject appears not relaxed, and his communicative behaviour is clearly perceived as aggressive, less controlled and more emphatic (i.e., aiming at convincing also the larger TV audience at home that his government did achieve all the promised goals).
Given the exploratory nature of our study, we decided to check our hypothesis of perceived differences by means of an informal evaluation task which involved a number of students and collegues (most of them without any background in multimodal communication). Subjects confirmed our judgments, and most of them commented

that the main features for interpreting the perceived differences in the two communicative behaviours were the "more emphasis/assertiveness in uttered words" and the "more gesticulation when speaking" in the second video excerpt with respect to the first.

3. Speech and Gesture Analysis

We decided to focus our preliminary analysis on some audio-visual parameters which we hypothesised could correlate with the observed/perceived differences in terms of emphasis and incisivity of Berlusconi communicative behaviours. On the speech level, we looked at the occurrence of pitch prominent syllables, and on the visual level on the occurrence of gestures which have been referred to in the literature as speech-accompanying gestures, i.e. gestures which are rhythmically synchronised with speech (Kendon, 1980; McNeill, 1992; Valbonesi et al, 2002; Loehr, 2004; Yannisik et al, 2004; Jannedy & Mendoza-Denton, 2005).

Synchrony between speech and gesture represents a strong evidence of the co-expressive, intrinsically multimodal nature of human communication (see also McNeill, 2005). Therefore, we looked not only the occurrence of specific events on each level separately, but also at the occurrence of the overlappings between some types of speech and gesture events, namely between prominent/non prominent syllables and gestural peaks.

Our hypothesis was to verify whether such measured parameters could contribute to giving an account of the differences perceived in the Berlusconi's communicative behaviour in the two video sections.

Gesture Labelling

Since in the two video excerpts the great majority of Berlusconi's speech-accompanying gestures are realised with (one or both) arms/hands, gesture analysis in our study has ben focussed on this type of gestures. At this preliminary stage of labelling, we simply marked the begin and end of every hand gesture and, for each of them, we identified the phases, according to the following criteria:

1. the preparation phase, when the subject is moving his hand/arm with respect to his start position
2. the stroke, i.e. "the peak of effort in the gesture" (McNeill, 1992:83)
3. (post-stroke) hold, which is "the final position reached by the hand at the end of the stroke" (McNeill, 1992: ibidem) and whose duration is variable (pre-stroke holds are also possible, but we did not find any in our data)
4. the retraction phase, when the hand/arm returs to its initial position (rarely found in our data because of the particular type of gesture analysed).

where phases 1, 3 and 4 are optional.

Moreover, in our data we observed and labeled speech-accompanying hand gestures we classified as "single stroke" vs "multiple stroke" gestures (recalling the "single" vs "repeated" discrete gestures defined by Yasinnik et al, 2004), the former being realised by a single gestural peak, the latter being characterised by a sequence of repeated gestural peaks within a gestural phrase.

We also marked whether each gesture was realised with one or both arms/hands.

The two video excerpts have been labelled with the software ViMar (Video Marker), a tool for multimodal annotation. This tool is a component of ViSuite, a software package for creating, labelling, visualising, and organising a multimodal database, which has been developed at the Human-Machine Interaction Systems Lab of Polythechnics of Bari (for a detailed account of ViSuite see section 6). A snapshot of the ViMar software during the gesture labelling process is shown in Figure 1. During gesture labelling the audio channel was always kept disabled.

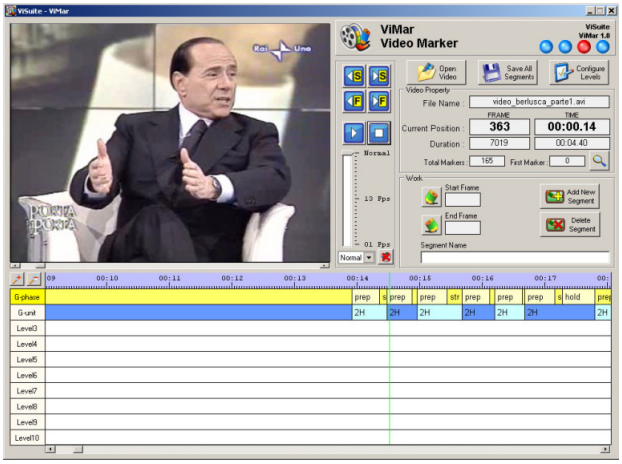

Figure 1: Snapshot of our multimodal annotation software ViMar during hand gesture labelling.

Speech Labelling

Speech productions have been separately labelled with the software package for speech analysis PRAAT (Boersma & Weenink, 1999) on two different levels: one marking words boundaries, and the other one marking syllables boundaries, specifying whether each syllable was characterised or not by a pitch prominence. A number of non-verbal phenomena like empty and filled pauses, inspirations, etc.. were also labelled.

4. Results

Analyses on different aspects at both audio and video levels has been performed. Given the exploratory nature of our investigation, and the resulting small sample of data examined, we are aware that results cannot be generalised, but nevertheless we believe they can give some indications with respect to our main goal, i.e. testing the possibility of determining audio-visual parameters able to measure perceived differences in communication.

Speech events occurrences are shown in Table 1. First of all, it can be noted that the number of words/syllables is

roughly the same in the two video selections, making them comparable not simply in terms of total duration (1 min each) but also commensurable from the speech productions viewpoint.
In speech productions, we would expect to find a higher percentage of prominent syllables over the total amount of syllables in the second part of the video: since more emphatic communication had been perceived there, we hypothesised a stronger presence of pitch prominence phenomena in the second excerpt. Yet results obtained do not show any significant difference between the two video selections, as it is shown in Table 1.

Prosodic event	**1st video excerpt**	**2nd video excerpt**
# words	187	173
# syllables	379	356
# prominent sylls	111	117
# non prom. sylls	268	239

Table 1: Occurrences of the prosodic events analysed in each of the two video excerpts.

More interesting results are obtained by analysing gestural events alone, which are shown in Table 2. While the same amount of speech-accompanying hand gestures is found in the two video excerpts (54 in the first, 53 in the second), they show a different distribution of the hand gestures types we classified as "single stroke" vs "multiple stroke". In the first video, communicative behaviour is characterised by a prevailing number of "single stroke" speech-accompanying hand gestures, whereas in the second part an increased number of "multiple stroke" ones is present. These results suggest that the perceived difference in terms of "more gesticulation" in the second video is not related to a higher number of hand gestures produced by the locutor, but by the type of stroke realisation in the gesture, namely multiple rather than single.

Gestural type/modality	**1st video excerpt**	**2nd video excerpt**
# speech-accomp. hand gestures	54	53
# "single stroke" hand gestures	18	2
# "multiple stroke" hand gestures	36	51
# holds	18	25
# gest. strokes 1 hand	26	19
# gest. strokes 2 hands	28	34

Table 2: Occurrences of gestural event types in each of the two video excerpts.

Hand gesture modalities like the presence of post-stroke holds, and the realisation of a gesture with both hands instead of one hand can be considered as means for adding emphasis and assertiveness to the accompanying speech events. With this respect, we found that whereas in the first video the subject makes use of one-handed and two-handed gestural strokes practically to the same extent, in the second part of the video his communicative behaviour is characterised by a larger number of two-handed than one-handed gestures. We also found a higher presence of (post-stroke)holds in the second video selection, which can also account for added emphasis.
Interesting results are also obtained by looking at the occurences of speech and gesture events which are temporally synchronised, as shown in Table 3. For example, we found that the number of prominent syllables overlapping with gestural strokes in the second video is almost twice with respect to the first one, and that again in the second video the percentage of gestural strokes overlapping with (one or more) prominent syllable(s) is higher than in the first part of the video (83% in the second, 50% in the first). In other words, in the second video the subject makes larger use of gestures for accompanying his speech, where gestures plays the role of "supporting" pitch prominences for adding emphasis.

Gesture/prosody synchronised event	**1st video excerpt**	**2nd video excerpt**
# strokes/prominent sylls	27	44
# strokes/non prominent sylls	22	8
# strokes/other	5	1

Table 3: Occurrences of gestural strokes overlapping with prosodic events in each of the two video excerpts. The third row refers to the number of strokes overlapping with nonverbal events like filled or empty pauses.

It can be also noted the almost equal distribution of strokes overlapping with prominent and non prominent syllables in the first video, as opposed to the largely prevailing number of overlappings with only prominent syllables in the second video. This observation might lead to the following speculation: since it can be assumed that Silvio Berlusconi, as other politicians, may be trained to use gestures "abundantly" in public speeches for persuasion purposes, subject may tend to produce gestures accompanying speech events without particular care of pitch prominence, as it is the case of the first video selection. In the second video, on the other hand, Berlusconi is emotionally urged by the non-sympathetic questions to convince the audience that he is right, this time by using gestures more appropriately for adding emphasis to his argumentations. This kind of speculations call for further investigations.
All the obtained results indicate that the parameters described above can contribute to accounting for the

perception of Berlusconi's communicative behaviour as "more emphatic" and "more gesticulating when speaking" in the second part of the video.

5. ViSuite, a suite of software tools for multimodal analysis

As mentioned before, the exploratory study described in this paper represents the preliminary step of a future larger project aiming to collecting and analysing a corpus of similar audio-video materials in Italian. A not negligible aspect in the creation of any database is the implementation of suitable software tools which could:

- assist the annotators in all steps of the process
- ensure enough flexibility in terms of annotation schemes to be used
- allow the complete control over the consistency and correctness of the database
- ensure easy, intuitive use by human trascribers also without particular technical skills, both in terms of user interface and output file formats

Currently a number of software tools are already available for multimodal annotation (for a detailed description and evaluation of the most popular tools see Rohlfing et al., 2006), notably ANVIL (Kipp, 2001) among the others. All these tools are XML-based: even though we are fully aware of all the potentialities of XML scheme, we agree with the view that working with XML annotation schemes and file format could not be straightforward for any kind of end-users (Rohlfing et al., 2006).

Moreover, since our aim is to build a multimodal database for further analysis, we thought that an additonal set of features for assisting the end-users in all the steps preceding the annotation would be useful if integrated in a software package.

For these main reasons, we decided to develop and implement a suite of tools named ViSuite (Video Suite) for collecting, organising, annotating and visualing a multimodal database. This system shares the basic functionalities with all the existing annotation tools, but implements few others which are specific with respects to the above mentioned aspects.

ViSuite is a Windows-based system, consisting of the following software tools, each one devoted to a specific task:

ViCom (**Vi**deo **Com**pression and join), which allows the compression of the video files created by a digital camera (any format) into video files compressed in the video format selected for ViMar (avi Indeo). It also codes the audio part of it as PCM, and reconstructs the audio/video file accordingly. This step may be necessary in all the cases when the video-camera used for the recording produces a video/audio file in different format and compression codes.

ViCut (**Vi**deo **Cut**ter), through wich it is possible to cut the original big video file in a number of smaller ones, along with the related audio files. This functionality stems from the consideration that long video sequences call for heavy computational resources, and in a medium size machine its elaboration may take some considerable amount of time. Shorter sequences let the user complete a labelling session in a reasonable time and possibly examine its results quickly before dealing with other materials.

ViMar (**Vi**deo **Mar**ker) is the core of the suite as it allows the labelling of the video files. It is characterised by the common basic features shared by other video labelling tools, as it can be seen from the system snapshot shown in Figure 1.

ViMar's output data format is the simple SAM (Fourcin, 1993)-like format, i.e.

<starting time> <ending time> <label>

This makes the system very easy to be controlled by human transcribers with different backgrounds, and therefore even by those without particular technical skills. ViMar is equipped with specific features for making all the produced data easily insertable into a coherent database. So, for example, the filenames related to annotation tiers are automatically assigned by the system, in order to prevent possible mistakes by human operators. In addition, label files produced by ViMar are generated with a protection algorithm which checks the coherence of the produced labels, and prevent possible manipulation of the label files content (typically, by means of text editors). Label files produced by ViMar are exportable for statistical processing in popular stastistical packages like Excel and SPSS.

ViView (**Vi**deo and audio labels **View**er) is the tool for visualising both video and audio annotation tiers, along with the video and speech information (waveform, fundamental frequency), the latter imported by existing speech analysis software tools (a snapshot is shown in Figure 2). It can import annotation files produced by systems like PRAAT (Boersma & Weenink, 1999), SegWin (Refice et al., 2000), Entropic XWAVES. A special conversion procedure has been implemented for PRAAT annotation files, whose file format is the less similar to SAM-like format.

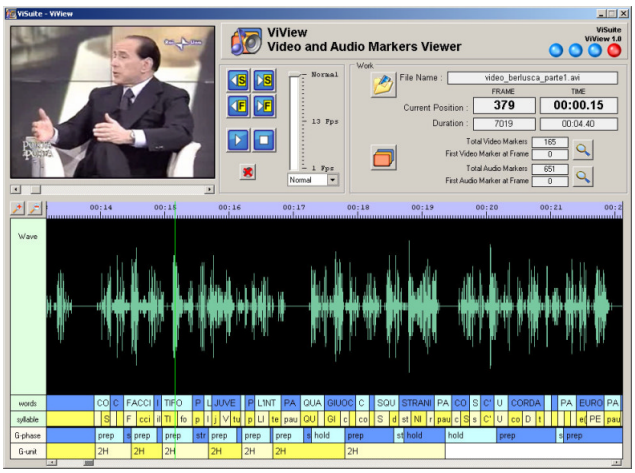

Figure 2: Snapshot of our multimodal visualiser ViView. Here gesture labels produced with ViMar, along with the speech labels imported from Praat are shown.

ViSuite software package is still under revision for further improvements. The pilot study presented in this paper represented also a testbed for this software.

6. Conclusions and Future Work

In this pilot study we explored the possibility of using audio-visual parameters for describing/measuring the differences perceived in interlocutor's communicative behaviours.

Preliminary results derived from the multimodal analysis of a single subject seem to indicate that measuring the distribution of some prosodic and (hand) gesture events which are temporally co-occurring contribute to the accounting of such perceived differences.

Moreover, as far as gesture events are concerned, it has been observed that relevant information are not simply to be found in the occurences of single gestures, but mainly in some gesture modalities (for example, "single stroke" vs "multiple stroke" gestures, one-hand vs both-hands gestures, etc…).

We believe that such preliminary indications can be useful for our future work of analysing a larger set of comparable data for Italian, where we plan to include also further aspects of audio-visual parameters. It would be interesting, for example, to look at the correlation between gestural peaks and pitch accent types (Loehr, 2004, Jannedy & Mendoza-Denton, 2005), and also between "gesticular phrasing" (Kendon, 1980) and prosodic phrasing.

In this paper we also introduced a software package for multimodal annotation, ViSuite, we developed and used for the work described in this paper.

7. Acknowledgements

We would like to acknowledge the contribution of our students Antonio Danese and Francesco A. Santangelo who implemented the ViSuite software package.

We are grateful to Adam Kendon and David McNeill for helpful discussion and advice on a very preliminary version of this work, and to three anonymous reviewers for useful comments and suggestions.

8. References

Boersma, P. & D. Weenink (1999). Praat. A system for doing phonetics by computer, http://www.fon.hum.uva.nl/praat/.

Cassell, J. and Stone, M. (1999). Living Hand to Mouth: Psychological Theories about Speech and Gesture in Interactive Dialogue Systems. In *Proceedings of the AAAI 1999 Fall Symposium on Psychological Models of Communication in Collaborative Systems*, North Falmouth, MA, Nov. 5-7, pp.34--42.

Fourcin, A. (1993). *The SAM project*. Chichester: Ellis Horwood.

Jannedy, S. & N. Mendoza-Denton (2005). Structuring Information through Gesture and Intonation. In S. Ishihara, M. Shmitz and A. Schwartz (eds), *Interdisciplinary Studies on Information Structure* 03, pp. 199--244.

Kendon, A. (1972). Some relationships between body motion and speech. In A. Siegmand and B. Pope (eds.), *Studies in dyadic communication*, New York: Pergamon Press, pp.17--210.

Kendon, A. (1980). Gesticulation and Speech: Two Aspects of the Process of Utterance. In M.R. Key (ed.) *The relation between verbal and nonverbal communication*, The Hague: Mouton, pp. 207--227.

Kipp, M. (2001). ANVIL – A generic annotation tool for multimodal dialogue. In *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech 2001)*, Alborg, pp.136--1370.

Loehr, D. (2004). Gesture and Intonation, Doctoral Dissertation, Georgetown University, Washington, DC.

McNeill, D. (1992), *Hand and Mind*, Chigago: Chicago University Press.

McNeill, D. (2005), *Gesture and Thought*, Chigago: Chicago University Press.

Refice M., Savino M, Altieri M. Altieri R. (2000). SegWin: A Tool for Segmenting, Annotating and Controlling the Creation of a Database of Spoken Italian Varieties. *In Proceedings of LREC 2000*, Athens, vol. 3, pp. 1531—1536.

Rohlfing K., Loehr, D., Duncan, S., Brown, A., Franklin, A., Kimbara, I., Milde J-T., Parrill F., Rose, T., Schmidt, T., Sloetjes, H., Thies, A., Wellinghoff, S. (2006). Comparison of multimodal annotation tools – workshop report. In *Gespraechsforschung* 7 (2006), pp.99--123 (www.gespraechsforschung-ozs.de).

Valbonesi, L., R. Ansari, D. McNeill, F. Quek, S. Duncan, K.E. McCullough, R. Bryll (2002). Multimodal signal analysis of prosody and hand motion: temporal correlation of speech and gesture. In *Proceedings of EUSIPCO 2002*, Toulouse, FR.

Yasinnik, Y., M. Renwick, S. Shattuck-Hufnagel (2004). The Timing of Speech-Accompanying Gestures with Respect to Prosody. In *Proceedings of the International Conference "From Sound to Sense"*, C97-C102, MIT, Cambridge, June 10-13.