# BIOMETRIC DATA COLLECTION FOR BIMODAL APPLICATIONS

*Piero Cosi, Emanuela Magno Caldognetto, Graziano Tisato and Claudio Zmarich*

ISTC-SPFD CNR
Istituto di Scienze e Tecnologie della Cognizione
Sezione di Padova “Fonetica e Dialettologia”,
Consiglio Nazionale delle Ricerche
www: http://nts.csrf.pd.cnr.it/
e-mail: {cosi, magno, tisato, zmarich}@csrf.pd.cnr.it

## ABSTRACT

This work focuses on the description of the environment and the procedures utilized at ISTC-SPFD for the biometric data collection of visual face-related articulatory (spatio-temporal) movements useful for lip-reading, bimodal communication theory, and the development of bimodal talking head and bimodal speech recognition applications.

## 1. INTRODUCTION

The knowledge that both acoustic and visual signal simultaneously convey extra linguistic, paralinguistic and linguistic information it is rather spread in the speech communication community.

Considering and simplifying the knowledge made available in the literature by phoneticians, psychologists and computer science researchers, it is useful to remember that (see Figure 1 [1]):

- speech signal transmits:

- information related to segmental phonological units (i.e. vocalic and consonantal phonemes driving lexical access) focused on spectral configurations, presence or absence of voice bar, etc.;
- information on the sintagmatic structure of the utterance, on possible focalizations, on sentence performatives, due to prosodic and intonation characteristics, carried by F0, energy and duration variations;
- paralinguistic information related to emotions and attitudes carried by F0, energy and duration variations, but also by amodal characteristics of vibrations of the vocal folds, by specific modifications of the spectral shape, etc [2];

- visual signal transmits:

- segmental linguistic information, visemes, groups pf phonologically equivalent visibile articulatory movements [3-5], that is of labial, mandibular and tongue (obviously considering only the front part of the tongue associated with the visibility of teeth) gestures and movements etc.
- suprasegmental linguistic information (focalizations, performatives, old/new information) carried partially by variations of labial movements (amplitude and duration variations) and mainly by forehead, eyebrows, eyes, and glance movements which represent the so-called “visual prosody” [6],
- paralinguistic information related to emotions and attitudes carried by specific forehead, eyebrows, eyes, and mouth configurations, as indicated by FACS system by Ekman and Friesen [7].

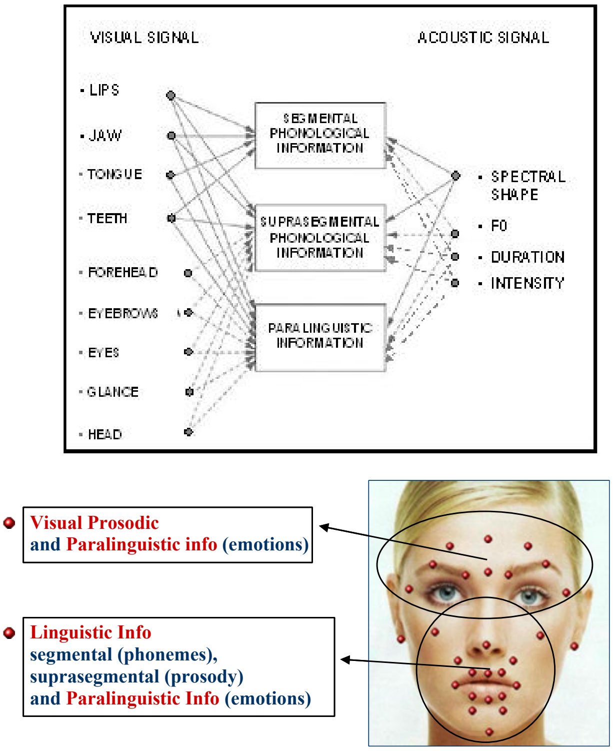
VISUAL SIGNAL
ACOUSTIC SIGNAL
• LIPS
• JAW
• TONGUE
• TEETH
• FOREHEAD
• EYEBROWS
• EYES
• GLANCE
• HEAD
SEGMENTAL PHONOLOGICAL INFORMATION
SUPRASEGMENTAL PHONOLOGICAL INFORMATION
PARALINGUISTIC INFORMATION
• SPECTRAL SHAPE
• F0
• DURATION
• INTENSITY
Visual Prosodic and Paralinguistic info (emotions)
Linguistic Info segmental (phonemes), suprasegmental (prosody) and Paralinguistic Info (emotions)

**Figure 1**. Visual and audio sources of linguistic and paralinguistic information.

The definition of the spatio-temporal characteristics of visible articulatory movements is actually important because it provides the basic experimental material with which various relevant theoretical problems could be tackled, such as, for example:

- the quantification of the available visible information relative to each phonological unit;
- the definition of the perceptive role of various articulatory parameters and of their relation and possible co occurrence with linguistic (distinctive) features;
- the identification of rules able to capture the variability induced by phonetic context, prosodic variations, speech rate;
- the determination of the iso- or aniso-morphism between articulatory movements and their correspondent acoustic product in order to formulate adequate rules for the integration of visual and auditory information useful for the synthesis of visible speech (talking heads);

but also various relevant applications could be conceived, such as, for example:

- the implementation of new audio-visual technological applications such as new talking head and new bimodal speech recognition systems.

The relevance of scientific studies resulting to the individuation of acoustic and visual correlates of all the above cited information it is quite evident at least for some languages, in particular for American English.

As for the “acoustic” speech signal, various multi-level labeling systems describing the co-occurrence of the different levels of linguistic information have been implemented [8].

As for the “visual” speech signal, a similar standard for this kind of systems is far to be reached mainly because the inventory of all the possible visual units in all the possible levels of information, in particular those related to the “visual prosody”, is far to be completed.

Segmentation and labeling of both acoustic and visual signals in representative units are particularly important for linguistic and psycholinguistic studies because looking into the coherence of bimodal information it is possible then to elaborate specific co-production theoretical models aiming to reach specific communicative tasks [9].

In fact, in order to develop theories on audio/visual production and perception of speech [10-12] and also to support various technological applications in telecommunications [13], in man machine interaction or in language teaching and rehabilitation, such as bimodal audio/visual speech synthesis [14-15] and recognition systems [16-19], it is essential to:

- identify the minimal units conveying visual linguistic information (visemes) [3-5], i.e. on the basis of articulatory parameters, specifying which are the consonants belonging to each of them;
- determine the relationships between the visible articulatory movements and the corresponding co-produced acoustic signal, i.e. determine the iso- or aniso-morphism between articulatory movements and their correspondent acoustic product.

This kind of information is partially language specific because, even though significant and expected cross-linguistic parallelisms are present due to the high versus low visibility of anterior versus posterior articulation loci, language specific characteristics arise due to the different size and structure of the phonological inventories ([20-21] for English visemes, and [14], [22] for French visemes).

# 1. BIOMETRIC DATA COLLECTION AT ISTC-SPFD: ENVIRONMENT AND PROCEDURES

For all the reasons mentioned above in the Introduction the importance of the collection and management of biometric data of visual face-related articulatory (spatio-temporal) movements is quite obvious.

During then last five years at ISTC-SPFD biometric visual face-related articulatory (spatio-temporal) movements were recorded and analyzed with ELITE [23].

ELITE is a fully automatic movement analyzer for 3D kinematics data acquisition. This system ensures a high accuracy and minimum discomfort to the subject. In fact, only small, non obtrusive, passive markers of 2mm of diameter, realized by reflective paper, are attached onto the speaking subject's face. The subjects are placed in the field of view of two CCD TV cameras at 1.5 meters from them. These cameras light up the markers by an infrared stroboscope, not visible in order to avoid any disturbance to the subject. ELITE is characterized by a two level architecture. The first level includes an interface to the environment and a fast processor for shape recognition (FPSR). The outputs of the TV cameras are sent at a frame rate of 100 Hz to the FPSR which provides for markers recognition based on a cross-correlation algorithm implemented in real-time by a pipe-lined parallel hardware. This algorithm allows the use of the system also in adverse lighting conditions, being able to discriminate between markers and reflexes of different shapes although brighter. Furthermore, since for each marker several pixels are recognized, the cross-correlation algorithm allows the computation of the weighted center of mass increasing the accuracy of the system up to 0.1mm on 28cm of field of view. The coordinates of the recognized markers are sent to the second level which is constituted by a general purpose personal computer. This level provides for 3D coordinate reconstruction, starting from the 2D perspective projections, by means of a stereophotogrammetric procedure which allows a free positioning of the TV cameras. The collinearity equations [24] are iteratively linearized and solved at least squares after the acquisition of a known control object [25]. The 3D data coordinates are then used to evaluate the parameters described hereinafter.

The input data consist of various speech stimuli such as simple disyllabic symmetric /'VCV/ and asymmetric /'$V_1CV_2$/ nonsense words, usually embedded in a carrier phrase, where C=consonants and V,$V_1$,$V_2$=vowels, or content words, sentences and “emotional” sentences uttered by a different set of male and female speakers, depending on the data collected. All the subjects producing the stimuli were northern Italian university students, aged between 19 and 22, and were paid volunteers. They usually repeated five times, in random order, each of the stimuli. The speaker comfortably sits on a chair, with a microphone in front of him, and utters the experimental paradigm words, under request of the operator.

The current complete acquisition pattern used also for “emotion” studies is graphically illustrated in Figure 2(a), but in the past a more simple pattern considering only lips/mouth movements, such as that illustrated in Figure 2(b) was considered.

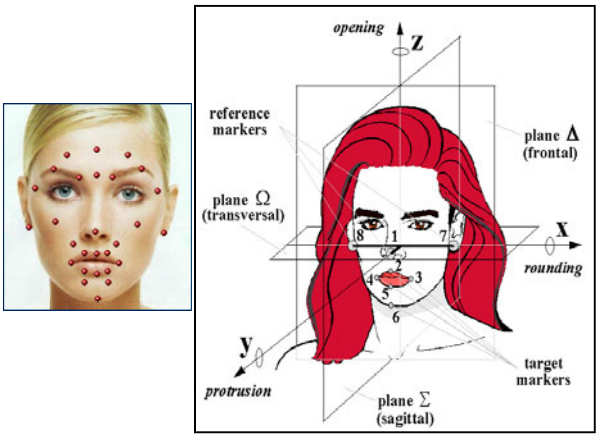

**Figure 2**. (a) Complete acquisition pattern for the data collection of articulatory movements; (b) position of the reflecting markers and of the reference planes during data collections for lips/mouth movements. Identification numbers are indicated next to their corresponding markers.

As illustrated in Figure 2(b), with this simpler acquisition pattern three reference points and five target points on the face of the subjects were considered.

In particular, the movements of the markers placed on the central points of the vermilion border of the upper lip (marker 2), and lower lip (marker 5), together with the movements of the marker placed on the corners of the mouth (markers 3, 4) were analyzed, while the markers placed on the tip of the nose (marker 1) and on the lobe of the ears (markers 7, 8) served only as reference points. In fact, in order to eliminate the effects of the head movement, the opening and closing gestures of the upper and lower lip movements were calculated as the distance of the markers 2 and 5 placed on the lips, from the transversal plane Ω depicted in Figure 3 and defined by the line crossing markers 7 and 8, placed on the ear lobes, and marker 1, placed on the tip of the nose. Similar distances with the frontal plane Δ perpendicular to the above one serve as a measure of upper and lower lip protrusion. A total of 14 values, defined as the difference between various markers or between markers and reference planes, plus the correspondent instantaneous velocity obtained by numerical differentiation, constitute the articulatory vector which has been used together with the acoustic vector in order to represent the target stimuli. The articulatory parameters, also listed in Table 1, were besides the upper and lower lip opening and closing movements (UL, LL), and the upper and lower lip protrusion (ULP, LLP), the lip opening height (LOH) calculated as the distance between markers 2 and 5, the lip opening width (LOW), calculated as the distance between markers 3 and 4, the jaw opening (JO), measured as the distance between the markers placed on the chin and on the tip of the nose, and the corresponding velocities.

| code | Meaning | definition |
|---|---|---|
| UL | upper lip vertical movement | d(m2,Ω) |
| LL | lower lip vertical movement | d(m5,Ω) |
| ULP | upper lip protrusion | d(m2,Δ) |
| LLP | lower lip protrusion | d(m5,Δ) |
| LOH | lip opening height | d(m2,m5) |
| LOW | lip opening width | d(m3,m4) |
| JO | jaw opening | d(m6,Ω) |
| ULv | ∂UL/∂t | ∂d(m2,Ω)/∂t |
| LLv | ∂LL/∂t | ∂d(m5,Ω)/∂t |
| ULPv | ∂ULP/∂t | ∂d(m2,Δ)/∂t |
| LLPv | ∂LLP/∂t | ∂d(m5,Δ)/∂t |
| LOHv | ∂LOH/∂t | ∂d(m2,m5)/∂t |
| LOWv | ∂LOW/∂t | ∂d(m3,m4)/∂t |
| JOv | ∂JO/∂t | ∂d(m6,Ω)/∂t |

**Table 1**. A subset of the articulatory parameter definitions used in data collections for lips/mouth movements.

An example of some the extracted articulatory parameters for the sequence /'apa/.is given in Figure 3.

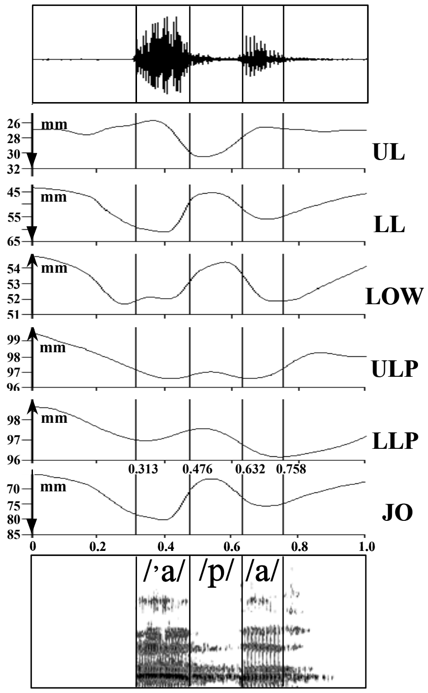

**Figure 3**. Time evolution of displacement and velocity of some of the markers and articulatory parameters illustrated in figure 2 (b) associated with the sequence /'apa/.

## 2. CONCLUDING REMARKS

We believe that the collected articulatory data could be useful both for theoretical modeling the way in which communicative informaton is conveyed by acoustic and visual signal simultaneously and for technological bimodal applications.

As for the implementation of "emotional speech" recognition systems or especially for the design of realistic "emotive" talking heads appropriate for various tasks and functional in various situations, we should be able to choose the information not only necessary but also more suitable to that aim.

As for bimodal synthesis, for example, besides extra linguistic information related to sex, race and age of the talker and those idiosyncratic (e.g.: morphology  of the face, eyes, nose etc.., frequency of vocal folds vibration, speech intensity, etc.) that are socially and socio-linguistic relevant for technological applications (e.g. virtual actor, handicap-aids for augmented speech communication etc.), a talking head shall correctly transmit selectable linguistic and  paralinguistic information depending on the specific tasks and applications for which it is designed.

All this information can be extracted by applying statistical or rule based approach working on real articulatory data such as those collected within this environment [26].

## 3. REFERENCES

1. Magno-Caldognetto E., Zmairch C. "Facce parlanti: problemi e potenzialità dal punto di vista della fonetica sperimentale", in "Il parlante e la sua lingua", *Atti delle X Giornate di studio del G.F.S.* (A.I.A.), Napoli, 13-15 Dicembre, 1999, pp. 117-137.
2. Magno-Caldognetto E., "I correlati fonetici delle emozioni", 2002, (in fase di stampa).
3. Magno-Caldognetto E., Vagges K., Ferrigno G., Zmarich C., "Articulatory Dynamics of Lips in italian /'VpV/ and /'VbV/ Sequences" *Proceedings of Eurospeech'93*, Berlin, 1993, vol. 1, pp. 409-413.
4. Magno-Caldognetto E., Zmarich C., Cosi P., Ferrero F.E., "Italian consonatal Visemes: Relationships between spatial/ temporal articulatory characteristics and co-produced acoustic signal", in Ch.Benoit and R. Campbell, (Eds.), *Proceedings of AVSP'97*, Rhodes (Greece), 1997, pp. 5-8.
5. Magno-Caldognetto E., Zmarich C., Cosi P., "Statistical Definition of Visual Information for Italian Vowels and Consonants", in D. Burnham, J. Robert-Ribes, E. Vatikiotis-Bateson, (Eds.), *Proceedings of AVSP'98*, Terrigal-Sydney (Australia), 1998, pp. 135-140.
6. Beskow J.,"Rule-based Visual Speech Synthesis", *Proceedings of Eurospeech'95*, Madrid, 1995, pp.299-302.
7. Eckman P., Friesen W., *Manual for the Facial Action Coding Systems*, Consulting Psych. Press, Palo Alto (CA), 1977.
8. *LinguisticAnnotation*, http://www.ldc.upenn.edu/annotation/.
9. Magno Caldognetto E., Poggi I., "Dall'analisi della multimodalità quotidiana alla costruzione di Agenti Animati con facce Parlanti ed Espressive", in Magno Caldognetto E. e Cosi P. (Eds.) *Multimodalità e Multimedialità ella Comunicazione*, Atti delle XI Giornate di Studio del GFS, Unipress, Padova 2001, pp. 47-55.
10. Summerfield Q., "Some Preliminaries to a Comprehensive Account of Audio-Visual Speech Perception", in Dodd B. and Campbell R. (Eds.), *Hearing by Eye: The Psychology of Lip-Reading*, Lawrence Erlbaum Associates, Hillsdale, New Jersey, 3-51.
11. Massaro D.W., "Speech Perception by Ear and Eye: a Paradigm for Psychological Inquiry", in Dodd B. and  Campbell R. (Eds.), *Hearing by Eye: The Psychology of Lip-Reading*, Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1987, pp. 53-83.
12. Massaro D.W., "Bimodal Speech Perception: A Progress Report", in Stork D.G. and Hennecke M.E. (Eds.) [13], 1996, pp. 79-102.
13. Storke D.G. and Henneke M.E. (Eds.),  *Speechreading by Humans and Machine: Models, Systems and Applications*, NATO ASI Sseries F: Computer and Systems Sciences, Vol. 150, 1996, Springer-Verlag.
14. Benoit C., Lallouache T., Mohamadi T., and Abry C., "A Set of French Visemes for Visual Speech Synthesis", in Bailly G., Benoit C., and Sawallis T.R. (Eds.), *Talking machines: Theories, Models, and Designs*, North-Holland, Amsterdam, 1992, pp. 485-504.
15. Cohen M.M, and Massaro D.W., "Behaviour Research Methods, Instruments and Computers", Vol. 22 (2), 1990, pp. 260-263.
16. Petajan E.D., "Automatic Lipreading to Enhance Speech Recognition", PhD Thesis, Univ. of Illinois at Urbana-Champaign. 1984.
17. Stork D.G., Wolff G. and Levine E., "Neural Network Lipreading System for Improved Speech Recognition", *Proc. of IEEE International Joint Conference on Neural Networks, IJCNN-92*, 1992, pp. 285-295
18. P.L. Silsbee and A.C. Allen. "Medium-Vocabulary Audio-Visual Speech Recognition", Proc. NATO ASI, New Advances and Trends in Speech Recognition and Coding, 1993, pp. 13-16.
19. A. Adjoudani and C. Benoit. "Audio-Visual Speech Recognition Compared Across Two Architectures", Proc. Eurospeech-95, Madrid, Spain, 18-21 September 1995, Vol. 2., pp. 1563-1566.
20. Walden B.E., Prosek R.A., Montgomery A.A. Scherr C.K. and Jones C.J., "Effects of Training on the Visual Recognition of Consonants", *Journal of Speech and Hearing Research*, Vol. 20, 1977, pp. 130-145.
21. Cohen M.M., Walker R.L. and Massaro D., "Perception of Synthetic Visual Speech", in Stork D.G. and Hennecke M.E. (Eds.) [13], 1996, pp. 153-168.
22. Benoit C., Guiard-Marigny T., Le Goff B. and Adjoudani A., "Which Components of the Face do Humans and Machines Best Speechread?", in Stork D.G. and Hennecke M.E. (Eds.) [13], 1996, pp. 315-328.
23. G. Ferrigno and A. Pedotti. "ELITE: A Digital Dedicated Hardware System for Movement Analysis via Real-Time TV Signal Processing", IEEE Trans. on Biomedical Engineering. BME-32, 1995, pp. 943-950.
24. R.P. Wolf. "*Elements of Photogrammetry*", Mc Graw-Hill Publisher, 1983.
25. Borghese N.A., Ferrigno G., Pedotti A.. "3D Movement Detection: a Hierarchical Approach", *Proc. of the 1988 International Conference on Systems, Man and Cybernetics, International Academic Publisher, 1988, pp. 333-336.*
26. Cosi P., Magno Caldognetto E., Perin G., Zmarich C., "Labial Coarticulation Modeling for Realistic Facial Animation", in *Proceedings of ICMI 2002*, Pittsburgh, PA, USA, October 14-16, 2002, pp. 505-510.