New Technologies for Simultaneous Acquisition of Speech Articulatory Data: 3D Articulograph, Ultrasound and Electroglottograph

Mirko Grimaldi[1]*, Barbara Gili Fivela*[1]*, Francesco Sigona*[1]*, Michele Tavella*[2]*, Paul Fitzpatrick*[3]*,*
Laila Craighero[3]*, Luciano Fadiga*[3]*, Giulio Sandini*[2]*, Giorgio Metta*[2]

[1] Centro di Ricerca Interdisciplinare sul Linguaggio, University of Salento, Lecce, Italy
[2] Laboratory for Integrated Advanced Robotics, University of Genoa, Italy
[3] Dep. S.B.T.A., Section of Human Physiology, University of Ferrara, Italy

{mirko.grimaldi, barbara.gili, francesco.sigona}@ateneo.unile.it;
{michele, paulfitz, pasa}@liralab.it; {fdl, crh}@unife.it; sandini@dist.unige.it

Abstract

The study of articulatory features in speech requires highly sophisticated instruments. Most of them are available at the CRIL research centre (*Centro di Ricerca Interdisciplinare sul Linguaggio* – University of Salento – Lecce) where the equipment usually found for articulatory studies in the most advanced international research centres is accessible: 3D articulograph, ultrasound system, electroglottograph, electropalatograph, and aerophone. Some of these instruments have been synchronized within the European CONTACT project, and have already been used for simultaneous recording.

In this paper, we will point out the peculiarity of each instrument, the added information related to the simultaneous recording, and the main steps towards their complete fruition in the specific phonetic-phonology field. In order to do that, we will briefly describe the recording set up and the first data collection, we will focus on a subset of the recordings in order to provide examples of the achieved articulatory-phonetic information, and describe the work-in-progress software analysis environment we use for linguistic purposes.

Index Terms: 3D articulography, ultrasound, electroglottography.

1. Introduction

The study of articulatory features of speech requires the use of an appropriate technology, often specifically developed for this purpose. In CRIL (*Centro di Ricerca Interdisciplinare sul Linguaggio* – University of Salento – Lecce) the main equipment used for articulatory studies in the most advanced international research centres is available: 3D articulograph, ultrasound system, electroglottograph, and aerophone [1], [2]and [3]. Within the European CONTACT project, some of these instruments have been synchronized and are usable to register various material simultaneously [4]. The aim of the CONTACT project is to verify if the development of language can be in parallel linked to the motoric control acquisition, especially the control that is necessary for precision movements [5], [6]. In this perspective, 3D articulograph, ultrasound system and electroglottograph have been synchronized (cf. [6]) in order to acquire articulatory material, which – supplied as input to an automatic speech recognition system – could eventually improve the speech recognition abilities. The material which was registered within the project – that is, a corpus of words and one of pseudowords, read by 9 speakers of the area of Lecce – was chosen according to the main objective of the project, and the methodology was adapted to its purposes. Actually, the registered material is not entirely suitable to verify specific linguistic hypotheses. However, in this paper we will take into account the pseudoword corpus, in order to point out the peculiarity of each instrument and the main aspects of their simultaneous use, and, furthermore, the main steps towards their complete fruition in the specific the phonetic-phonologic field.

Also, we describe the analysis environment and provide examples of the articulatory information which can be obtained by this type of material. The analysis environment – developed in MatLab® [7] – allows to display the wave shape corresponding to the electroglottographic signal, any interval which has been identified thanks to previous segmentation and labelling phases, the pictures obtained simultaneously by the ultrasound equipment and the graphs of the articulograph sensors.

2. Instrumentation

A constellation of hardware and software that allows the simultaneous recording of acoustic and articulatory data has been implemented. It is made up of several interconnected devices; some of those are synchronized in hardware, while the remaining rely on post-processing synchronization (for more details see [6]:49-96).

2.1. 3-D Electromagnetic articulograph

An electromagnetic articulograph (EMA), model AG500 by Carstens Medizinelektronik GmbH, Germany [8], was used to track the movements of a set of sensors glued on the tongue, on the lips, on the front teeth and on a pair of glasses (the latter, used as reference, to compensate head movement), during speech production (Figure 1 shows how the sensors are

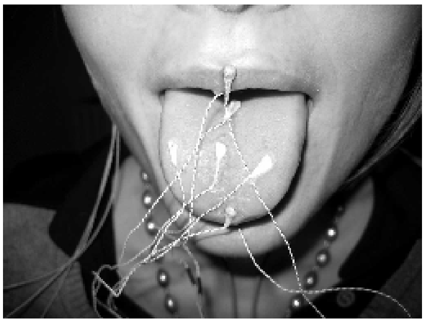

Figure 1: *EMA sensors glued on the subjcet's tongue*

positioned onto the tongue). The EMA device is able to determine the Cartesian coordinates x, y and z, as well as azimuth and elevation of up to twelve directionally sensitive magnetic field sensors at a sampling frequency of 200Hz. The measuring sensors are single axis coils. Six reference coils, arranged to form a three-dimensional frame of reference, emit six magnetic fields at different well known frequencies between 7500 and 13750 Hz. During a recording session, the alternating currents induced in the sensors by the magnetic fields of the reference coils are separated by their frequencies, digitized and sent in real-time to the AG500 control PC (using an Ethernet connection). Software provided by Carstens Medizinelektronik GmbH stores the current values on the hard drive of the control PC, making them available for the spatial arrangement determination process.

It is important to stress that the speech signal coming from a microphone is also synchronously recorded with positional data.

2.2. Ultrasound system

The investigation of the tongue motion is also accomplished by means of an ultrasound system, which is both non-invasive and non obtrusive, thus non affecting speech production, and is able to provide the full profile of the tongue dorsum, although the apex and the radix are often occluded during speech production by the jaw and by the hyoid bone respectively.

Medical sonography uses high frequency (2-14 MHz) sound waves emitted from an array of piezoelectric transducers (crystals), multiplexed in time so that only one crystal emits sound waves in a given time interval, while all the remaining crystals are used to convert the received echoes to voltage values. Indeed, when an ultrasound wave reaches an interface between materials with different impedance properties, it is reflected. Voltage values of the received echoes are processed and the image is reconstructed.

As it can be easy verified in the following echography pictures, the brightest area in the achieved images is the subject’s tongue. The brighter the pixel is, the higher the density gradient is, thus the acquired echoes. The tongue lamina is clearly visible since a great density gradient exists.

Actually, an Aplio XV machine, by Toshiba Medical System corp. [9] has been used, which also allows exporting ultrasound pictures as a continuous video stream (at 25 Hz) by means of a dedicated S-Video output. The video stream is synchronously acquired together with the audio signal, by means of an external a/v analog-to-digital acquisition card, and then recorder in real-time on a dedicated PC.

2.3. Electroglottograph

Electroglottography is a technique used to register laryngeal movements measuring the change in electrical impedance across the throat during speech production. The EGG device used for the Linguometer is a “Laryngograph Microprocessor“ by Laryngograph Ltd, London, United Kingdom [10]. An alternated current generator supplies a high frequency (usually from 300 kHz to 5 MHz) sinusoidal current to a pair of copper-made electrodes, kept by an elastic band on the surface of the throat at the level of the thyroid cartilage. The current amplitude is in the order of few milliampere while the applied voltage is around 0.5 V. When the vocal folds move, a rapid variation of the conductance is observed in the electroglottographic (EGG) signal applied across the larynx.

Driven by a dedicated software, the elettroglottograph acquires both the EGG signal and the speech signal synchronously, at 16 kHz, and sends them to an attached control PC using a standard USB interface.

The system enables the study of the regularity of the vocal folds vibrations, the single opening and closing phases, their correlation and shape.

2.4. Facial expressions recording

During each experiment, facial expressions of the subject have been recorded on a tape, in DV format, by means of a camcorder. The tape content is exported off-line on a PC as a single a/v stream, to be segmented later. At the moment, the actual purpose of facial expression recording is not to be matter of study, but to test the segmentation and the synchronization procedures on a type of signal which does not embed the segmentation pulses in itself, while heavily deploying cross-correlation between the speech component and the speech reference signal ([6] : 58-59).

2.5. Stimuli presentation, signal recording and control

In the data acquisition setup for the CONTACT project, a GNU/Linux-based workstation handles the stimuli (i.e. the corpora) presentation. Furthermore, since all of the signals are acquired as a set of continuous streams (and they need to be segmented off-line in order to make the analysis of single words and pseudowords easier), the stimuli presentation software is also in charge to generate a segmentation signal (a set of well-known acoustic pulses).

The segmentation signal is entered into an audio mixer together with the speech signal coming from a couple of microphones. The resulting audio signal is then entered in a a/v analog-to-digital acquisition card together with the continuous video stream, as depicted before.

Another Microsoft Windows-based PC run the Cartsens Medizinelektronik GmbH software to calibrate and record the data of the articulograph. Furthermore, it run the software required by the laryngograph.

A server shared on the network is used to store the data recorded during the investigations.

Finally, a notebook computer was used to control the recording apparatus and to monitor the execution of the stimuli presentation software.

2.6. Post-processing

As already said, signals are acquired as continuous streams: an off-line software procedure is in charge to segment those streams, i.e., to separate each word the submitted corpus is made of.

Also, data coming from different sources, such as EGG and US data, need to be time-aligned, i.e. for each word/pseudoword, the starting instant for a signal must be related to the starting instant of the others, in order to make it possible to see what happens to all of the signals at a given time interval. This task is accomplished by another software tool, which basically deploys the cross-correlation between the audio signals synchronously recorded together with EGG and US respectively.

Finally, the resulting data are packaged in a Matlab-compliant format, so that they can be analyzed and shared among the scientific community in a *de facto* standard format.

Post processing software includes C/C++, Perl, Matlab programming technologies and run almost full-automatically (it may requires several hours of computation, depending on a lot of factors such as the CPU speed, the amount of data to be processed, etc.): just a little human control is required at the very beginning of operations.

2.7. Software for data analysis

In order to show the result of the data acquisition procedure, a simple software analysis environment has been set up. This environment is under development, and relies on the main idea to integrate as much as possible already existing software tools. Actually, the environment relies on PRAAT [11], Edgetrak [12], Mplayer [13][13] and a Matlab-based GUI developed at CRIL (let's call it "mygui") which is able to interact with the above-mentioned companions within certain limits, also thanks to the scripting capabilities of the operating system, and the command-line interface of Praat and Mplayer.

Driven by the Matlab GUI commands, Mplayer is used to extract all of the pictures from the ultrasound and facial expression movies, in order to be displayed with ease when requested. Mplayer is also used to build demonstrative a/v movies of the simultaneous playback of all the synchronized signals.

PRAAT is deployed directly by the user at the very beginning of the data analysis stage, to build multiple levels of labels attached to one or more selected time intervals of interest, related to a word/pseudoword: the text file describing labels can be automatically imported in the Matlab environment by means of an ad-hoc script. Even if Matlab is already equipped with a toolbox to generate spectrograms, "mygui" has been designed to drive Praat to generate spectrogram (with superimposed formants) plots and to export them as text files, which are then imported and displayed within "mygui".

Edgetrack is a software that allows to generate a dotted contour of the tongue on the mid-sagittal plane, for a sequence of ultrasound images. The user is asked to select the sequence and help the program tuning some parameters to generate the contour for the first picture: Edgetrak will generate dotted contours for the remaining ones, following the tongue movements picture by picture. The set of contours can be exported in text format: "mygui" can import and use the file to superimpose the contours over the appropriate pictures again.

"Mygui" is the core of the environment, and is used to automatically import and show all of the available and time-aligned information, including EGG signal, facial expression pictures, speech signal, spectrogram and formants, x-y-z coordinates of the EMA sensors, belonging to a selected word/pseudoword. In addition, the user can display single ultrasound pictures, with or without superimposed Edgetrak tongue contours, reference grids and the EMA sensors glued on the tongue dorsum (within the current implementation of the system, the latter feature is a challenge, and actually is just an estimation of the most likely positions: at the moment it should not be considered for quantitative relative measurements between EMA sensors and the tongue dorsum resulting in the ultrasound picture). Of course the user can filter the data in order to show only a subset, or choose a different time interval with respect to the ones associated to the Praat labels at the very beginning of the operations.

3. First data collection and observations

As already mentioned, a corpus of data was recorded within the CONTACT project. The aim was getting articulatory data to train an automatic learning system. Part of these data will be considered here to point out the impact of the simultaneous acquisition of articulatory data on phonetic studies, focusing on specific linguistic hypotheses.

Nine speakers of Lecce Italian were asked to read three times a corpus of 74 words and 68 pseudo-words (both words and pseudo-words were read as declarative sentences; words were also read as questions). Both words and pseudo-words were chosen in order to include all consonantal and vocalic phonemes attested in Italian. In particular, words were mainly stress-initial (e.g., /'matto, 'nome, 'strada/ <mad, name, street>) and were chosen in order to show the various consonants in word initial position, followed by different vowels (e.g., /'matto, 'muffa, 'moro/ <mad, mould, dark >); moreover, instances of words with different stress positions were also inserted (e.g., /mat'tone, pa'pa/ <brick, dad>). Pseudo-words were monosyllables, chosen in order to get data on all the Italian consonants followed by /a, u, i/ vowels (e.g., /'na, 'nu, 'ni, 'λa, 'λu, 'λi, /).

We will focus here on the recordings of two speakers regarding pseudo-words composed by unvoiced alveo-dental/post-alveolar articulation (/t, s, ts, tS/) and a vowel (/a, i, u/). The acoustic signal was first segmented and labelled by means of the software PRAAT [9]. For each monosyllable, both the syllable segment boundaries and the transition(s) between segments were identified (e.g., for [tSu], the transitions between [t] and [S], and [S] and [u]). The inspection of articulatory data was performed by means of the Matlab script described above (see previous section). The data analysed allow us to point out the possible specific contribution of the instruments described above, used simultaneously for linguistic analysis.

3.1. Integrating information on articulatory gestures

The electroglottograph offers an unambiguous indication of the boundaries between voiced and unvoiced segments: the boundary is clearly detectable, independently from the segmentation procedure and the settings for acoustic analysis (e.g., spectrogram settings); in voiced stretches, it offers information on vocal fold vibration during voicing.

The ultrasound system offers information on the morphology of the tongue and integrates the information on specific points obtained by means of the AG500 sensors. In particular, the surface of the whole tongue during the affricate production allowed us to observe the specific transition gesture in the plosive-fricative and fricative-vowel segments. The ultrasound images are definitely useful in relation to the back of the tongue, an area which is particularly difficult to track – see Figure 2. Notice that no AG500 sensor may be glued as back as the point indicated in the figure without causing problems to subjects.

The AG500 offers highly detailed information on specific points located on the articulators, as for both the spatial and the temporal domain. First of all, the system tracks the lip kinematics that would not be available with the other instruments alone – see Figure 3; moreover it integrates the information on tongue morphology obtained by means of the ultrasound system, both in terms of spatial-temporal resolution and in terms of specific points than may be monitored, e.g., the tip of tongue is often missing or difficult to detect in ultrasound images.

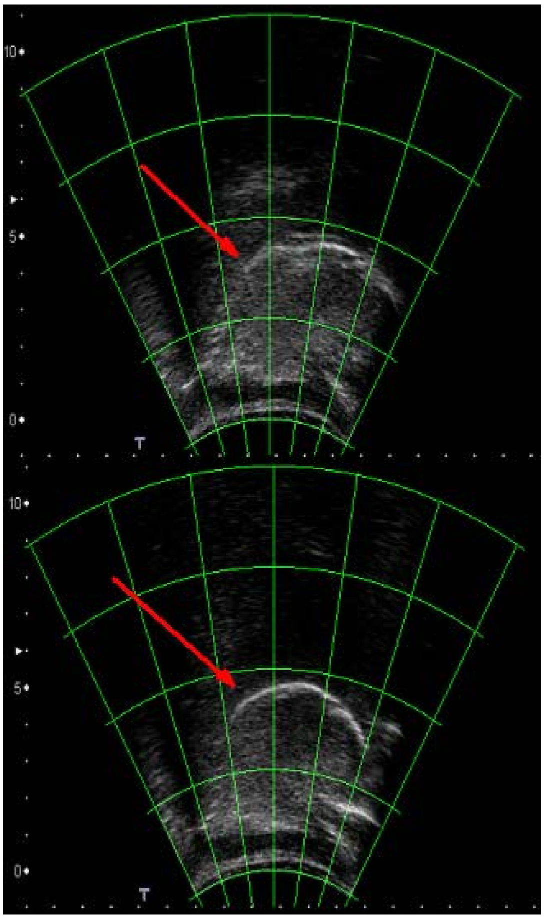

Figure 2*: Tongue position during the production of pseudo-word 'ciu' /'tSu/: back of the tongue position (red arrow) at the beginning and at the end (after 120 ms) of the consonant-vowel transition – upper and lower panel, respectively.*

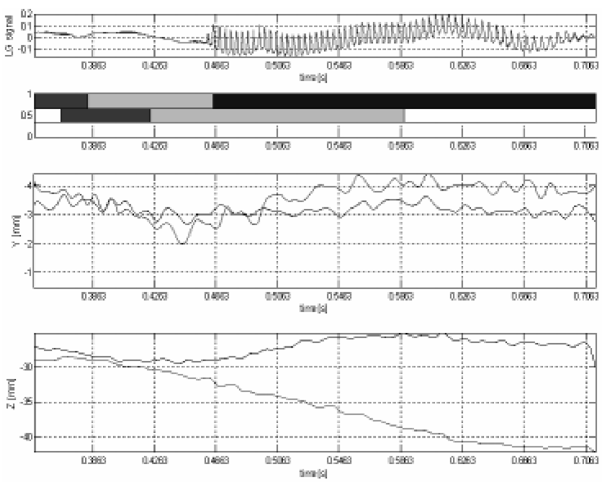

Figure 3 *EGG signal (1st panel from top), segments, transition intervals and slots for ultrasound images (2nd panel), and tracks of AG500 sensors during the production of 'ciu' /tSu/: sensors placed on the lower and upper lips – (3rd panel) – and on the tip and dorsum of the tongue – lower panel.*

3.2. Ultrasound tongue contour and EMA data

One of the possible perspectives of analysis achievable using US and EMA techniques is based on the observation that during speech production the tongue deforms in a complex manner, probably because the gesture is influenced by more than one phone. A problem is deeply related to this: how is this highly deformable tissue of the tongue controlled to obtain fine gestures necessary for speech? Classical hypotheses postulated a tip and body subdivision of the tongue executing quasi-independent motion back and forth and bottom and up directions. However, thanks to recent imaging techniques, it is possible to suppose that tongue deformations are far more complex than the motion of a tongue body and a tongue tip [14], [15]. A more recent hypothesis is that the tongue would be controlled by the synergistic coordination of muscular 'functional segments'. These segments are laid out orthogonally to the longitudinal axis of the vocal tract and composed of multiple muscle systems. From this point of view, the tongue can be divided into quasi-independently controlled functional segments based on regions of the tongue and vocal tract, rather than gross muscle architecture. Instead of entire muscles aligning to execute a gesture, segments would be controlled independently or aggregated into larger units to form coordinative structures determined by language dependent phonetic considerations [16].

Among the muscles that act on the tongue structure, there are three extrinsic muscles that originate on bone structures and insert into the tongue (responsible for the main displacement and shaping of the overall tongue structure): the genioglossus, GG (different fibres produces a forward and upward movement), the styloglossus (raises and retracts the tongue, causing a bunching of the dorsum in the velar region), and the hyoglossus (retracts and lowers the tongue body). The floor of the mouth contains also the geniohyoid (GH) and the mylohyoid (MH) muscles, that elevate the hyoid bone and the base of the tongue. Three additional intrinsic muscles contribute to a minor extent to the sagittal tongue shape. The superior longitudinalis muscle shortens the tongue, and bends its blade upwards. The inferior longitudinalis muscle depresses the tip. The verticalis fibres depress the tongue and flatten its surface ([17]: 171-177; see Figure 4).

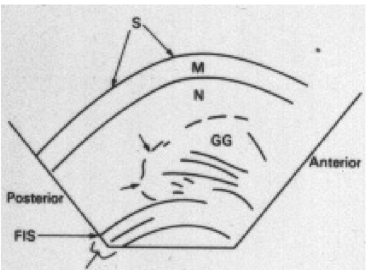

Figure 4: *schematic of US mid-sagittal scan (adapted from [1]: 12). S = Tongue Surface; M = mucosa; N = Superior, Inferiorior logitudinalis, and Vertical, Horizontal network of fibres; FIS = Floor Intermuscular septum.*

Using our data, a preliminary attempt in this direction was made observing the role of tongue muscles and the position of the tongue surface, both in US images and in EMA sensors tacks.

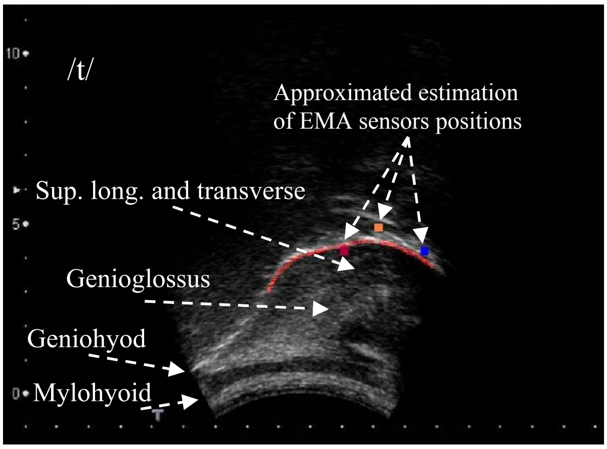

Figure 5: *closure phase and burst in phoneme /t/.*

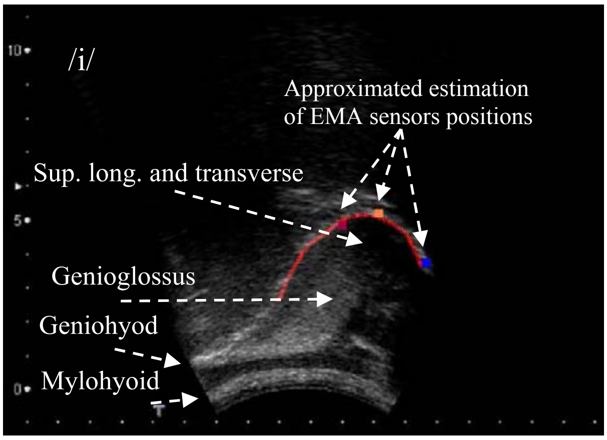

Figure 6 raising and fronting in phoneme /i/.

We can try to analyze a complex segment as /ti/ consisting of an occlusive and an anterior high vowel. We can observe the realization of /t/ in Figure 5 and of /i/ in Figure 6.

It's possible to note that phonemic differentiation in mid-sagittal tongue motions is realized through a functional independence of the tongue muscles. In particular, in the closure phase the expansion and compression of the GG together with superior longitudinal and transverse muscles seem more involved (Figure 5). On the other hand the fronting and raising of the tongue seems to be controlled by the expansion and compression of the posterior fibres of the GG muscle (Figure 6). In the same way GH and MH muscle could be involved to the realization of the Advanced Tongue Root (ATR) feature in vowel /i/. Finally, EMA sensors, whose position on the images is only probabilistic estimated, emphasize the gestural goal reached by the synergistic coupling of the tongue muscles.

4. Conclusions

The most relevant instruments for articulatory studies are available at the CRIL research center, in Lecce, and have been synchronized within the CONTACT project. In the paper we report some notes on the information added by the simultaneous recording, and the main steps towards the complete fruition of the equipment for specific phonetic-phonology purposes. Some examples of the achieved articulatory-phonetic information are provided, and the (currently being developed) analysis environment used for addressing specific linguistic issues is described.

The synchronization of the articulograph, ultrasound and electroglottograph allows us to integrate information on the vocal fold vibration and both internal and external articulators. In particular, the synchronization of the instruments may offer richer information as it enables both to track different points and segments and to track them with different resolution qualities. Moreover, the simultaneous recording by means of the instruments offers different information on the same point/segment. This 'redundancy' may, in any case, be useful to have a clearer idea of articulatory gestures during speech, and to help scholars in the elaboration of more accurate theoretical assumptions about the nature of speech and language.

5. References

[1] Stone, M., (2005), "A Guide to Analyzing Tongue Motion from Ultrasound Images", *Clinical Linguistics and Phonetics*, 19, (6-7), 455-502

[2] Wrench, A.A. (2007), "Advances in EPG palate design", Advances in Speech-Language Pathology, 99, Issue 1 March 2007 , pages 3 – 12

[3] Zierdt, A., Hoole, P., Tillmann, H.G., (1999), "Development of a System for Three-Dimensional Fleshpoint Measurement of Speech Movements", Proc. ICPhS'99, San Francisco. August

[4] Grimaldi, M., Gili Fivela, B., Tavella, M., Sigona, F., Fitzpatrick, P., Metta, G., Craighero, L., Fadiga, L., Sandini, G., (2007). "Synchronized acquisition of Italian speech articulatory data using ultrasound, 3D-articulograph and laryngograph: First results." Presented at *Ultrafest VII,* New York, September 2007 (http://jerome.linguistics.fas.nyu.edu/ultrafest.html), and to be published in Proceedings of AISV "La fonetica sperimentale: metodo ed applicazioni", December 2007.

[5] http://eris.liralab.it/contact/

[6] Tavella, M., (2007), *Simultaneous recording of phono-articulatory parameters during speech production*, unpublished Master Thesis, University of Genova, LiraLab (Laboratory for Integrated Advanced Robotics), 2006-2007.

[7] http://www.mathworks.com

[8] http://www.articulograph.de

[9] http://www.medical.toshiba.com

[10] http://www.laryngograph.com

[11] Boersma and Weenink, University of Amsterdaam. http://www.fon.hum.uva.nl/praat

[12] Li, M., Kambhamettu, C., and Stone, M. (2005) Automatic contour tracking in ultrasound images. *Clinical Linguistics and Phonetics*, 19 (6-7), 545-554.

[13] http://www.mplayerhq.hu

[14] Wilhelms-Tricarico, R. (1995), Physiological modeling of speech production: Methods for modeling soft-tissue articulators. *Journal of the Acoustical Society of America*, 97, 3085–3098.

[15] Stone, M. & Lundberg A., (1996), Three-dimensional tongue surface shapes of English consonants and vowels. *Journal of the Acoustical Society of America*, 99, 3728–3737.

[16] Stone, M. Epstein, M. A., Iskarous, K., (2004), Functional Segments in Tongue Movement, *Clinical Linguistics and Phonetics*, 18, 6-8, 507-521.

[17] Kent R. D., (1997), *The Speech Sciences*, Singular Publishing, Group, San Diego-London.