A low-cost stereovision based system for acquisition of visible articulatory data

Brigitte Wrobel-Dautcourt, Marie-Odile Berger, Blaise Potard, Yves Laprie, Slim Ouni

A LOW-COST STEREOVISION BASED SYSTEM FOR ACQUISITION OF VISIBLE ARTICULATORY DATA

B. Wrobel-Dautcourt, M. O. Berger, B. Potard, Y. Laprie, S. Ouni

LORIA - Campus Scientifique - BP 239 - 54506 Vandœuvre-lès-Nancy Cedex - France

ABSTRACT

In this paper, we present the 3D acquisition infrastructure we developed for building a talking face and studying some aspects of visual speech. A short-term aim is to study coarticulation for the French language and to develop a model which respects a real talker articulation. One key factor is to be able to acquire a large amount of 3D data with a low-cost system more flexible than existing motion capture systems (using infrared cameras and glued markers).

Our system only uses two standard cameras, a PC and painted markers that do not change speech articulation and provides a sufficiently fast acquisition rate to enable an efficient temporal tracking of 3D points. We present here our stereovision data capture system and how these data can be used in acoustic-to-articulatory inversion.

1. INTRODUCTION

The study of labial coarticulation, face expressions and more generally face deformations involved in speech production is often hindered by the limited amount of data available.

To a great extent it is due to the cost and utilization constraints of acquisition systems. Indeed, most of the motion capture systems require markers to be glued onto the face and do not allow an arbitrary number of markers to be used simultaneously.

The first part of this paper describes a low cost acquisition system together with deformation modes for two French speakers. The second part shows how this system was used to investigate the incorporation of articulatory constraints derived from tracking results in acoustic-to-articulatory inversion.

2. VISUAL DATA ACQUISITION

2.1 Set-up

The investigation of labial coarticulation requires the same physical points to be tracked over time. As the natural skin is not textured enough, we chose to paint markers on the speaker's face. This method allows the size, the density and the position of the interesting points to be controlled. For example, 140 white markers were painted on lips, cheeks and jaw; 6 markers on the upper part of the face were used to compensate for the global motion of the head (Figure 1).

For another data capture sequence, we have painted 168 white markers on the speaker’s face (Figure 2).

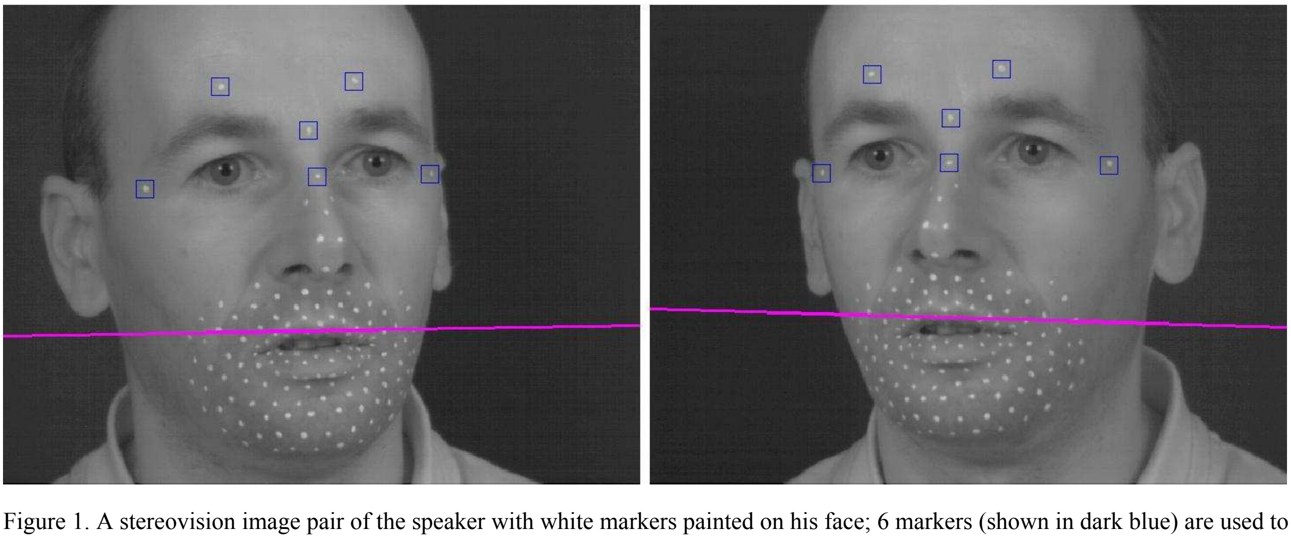

Figure 1. A stereovision image pair of the speaker with white markers painted on his face; 6 markers (shown in dark blue) are used to compute global motion of the head. The epipolar geometry, computed with a calibration object, describes the relations that exist between two images: corresponding points must belong on epipolar lines like the ones shown on the images.

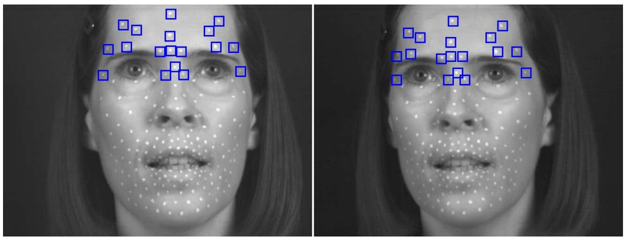

Figure 2. A stereovision image pair of a female speaker with markers painted on her face. In this example, 18 markers (dark blue) are used to compute global motion of the head.

Two monochrome cameras (JAI A33) are used for acquisition because their acquisition rate (120 fps) is faster than that of color cameras and thus enable the tracking of very fast movements of the articulators, i.e. release bursts for instance.

During the recording, the speaker sat in front of the stereo camera pair with a microphone placed 50-60 cm far from his mouth. The face was illuminated by two light sources so that no shadows appear on it (Figure 3). The logatoms and sentences to pronounce are projected on a screen in front of the speaker; this allows him to keep his head straight forward.

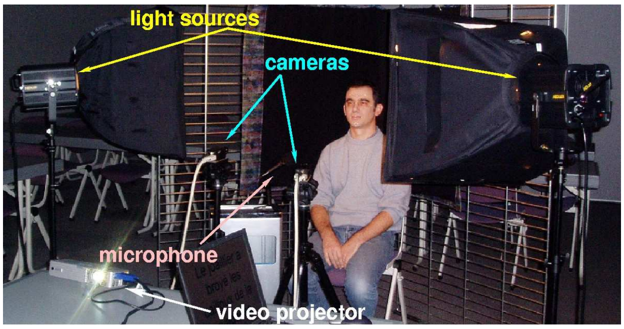

Figure 3. Video and sound recording. The speaker sits at ease in front of two stereovision cameras.

2.2 Reconstruction

A preprocessing stage is first used on the images to detect the markers based on their average grayscale, shape and size (white circular points with a radius less than 3 pixels). This low-level process is able to detect most of the markers except some points that are not visible in one image of some stereo image pairs. Examples of problems are points on the temples which may disappear when the speaker moves his head or some points on the lips that are occluded during protrusion or closing of the mouth. Conversely, this process may find erroneous points that are not markers but have the same photometric features, as light reflects on eyes or teeth. A classical stereovision algorithm is then applied to recover the set of 3D corresponding points for every image pair of the sequence. This algorithm respects the traditional constraints such as unicity and epipolar geometry (corresponding points must belong to epipolar lines, cf Figure 1). After triangulation of homologous points, an average of 86% of physical points are well reconstructed, 10% of 3D points are erroneous and 4% of markers are missing. Most of the erroneous points correspond to hidden markers (mainly on the inner part of the upper lip, temples and cheeks).

Building temporal trajectories of points is often based on closeness criteria. A point at time *t* is connected to the nearest point from the set of points computed at *t+1*. Unfortunately, due to high deformation of face during articulation, ambiguities may appear in the choice of the corresponding point: the trajectories of upper and lower lip markers can be confused when the mouth closes and then opens.

Reliable temporal trajectories, possibly incomplete, are built using position and velocity similarity criteria when non ambiguous matching is possible. Unlike classical tracking algorithms, we do not use smoothness criteria because the movements of the lips are very fast and have very low inertia: the changes of direction and velocity are very quick. A global strategy which makes use of a topological mesh of the face is finally used to fill in the gaps and to interpolate missing data. At time 0, an initial mesh is automatically built from the set of 3D markers [1] and corrected by hand. This step requires 5 to 20 minutes, it directly depends of the number of missing points. We can then define a mesh that evolves over time while keeping the same topology. Missing points are then recovered from the knowledge of their neighbours using a classical interpolation scheme. On average, about 7% of markers are estimated by this process, the others being directly computed by stereovision.

2.3 Evaluation

We first recorded the speaker shown on Figure 1 during 125 seconds. The corpus used was composed of logatoms (10 vowel sequences, 4 consonant sequences, 7 "vccv" sequences, 2 "cvccv" sequences) and 2 sentences; each sequence stands for 5 seconds of speech. To evaluate our acquisition system, we applied a principal component analysis on these data composed of 14,870 frames × 150 points × 3 coordinates. The first three principal modes explain more than 89% of variance, indicating a small number of functional dimensions of face movements. In detail, the first component accounts for protrusion which explains 60.4% of variance (Figure 4). This mode shows the importance of protrusion in the French language. The two remaining components (Figure 5) encapsulate the up/down lower jaw motion (22% of variance) and the aperture/closeness of the mouth concerning the upper lip (7.1% of variance). These modes are in very good agreement with those computed by S. Maeda [3].

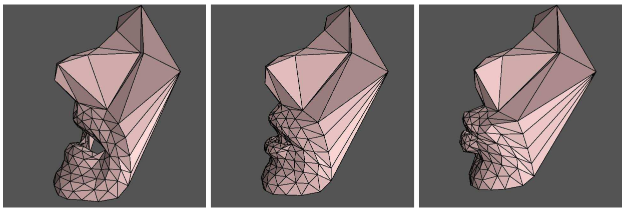

Figure 4. Effect of varying the first mode between ±3 standard deviation, this first mode accounts for protrusion (60.4% of variance).

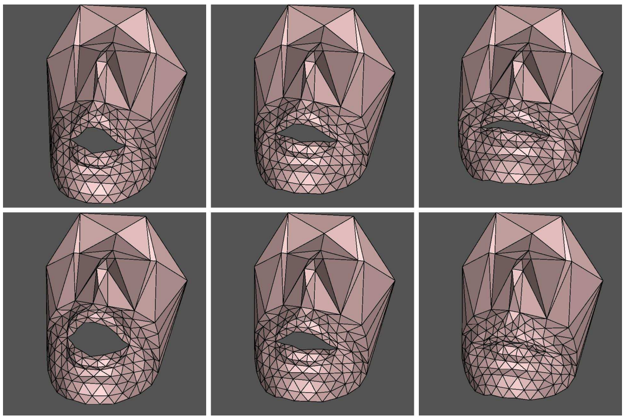

Figure 5. Effect of varying each of the second and third principal modes in turn between ±3 standard deviation; these modes account for the up/down lower jaw motion (22% of variance) and for the aperture/closeness of the mouth (7.1% of variance).

We also recorded another speaker (the woman shown on Figure 2) on a bigger corpus of 900 seconds. This corpus mixed logatoms and sentences. Many more points were painted on the face in order to capture, in addition of lip and jaw motion, cheeks and nose motion. We then applied a principal component analysis of these data composed of 107,796 frames × 190 points × 3 coordinates. The first three components explain 88% of variance (Figure 6 and Figure 7).

These two experiments, with two native French speakers are in good agreement, even though the corpus, the markers, and the speaker are different.

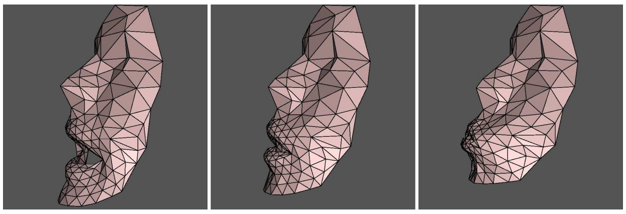

Figure 6. Effect of varying the first mode between ±3 standard deviation, this first mode accounts for protrusion (65.2% of variance).

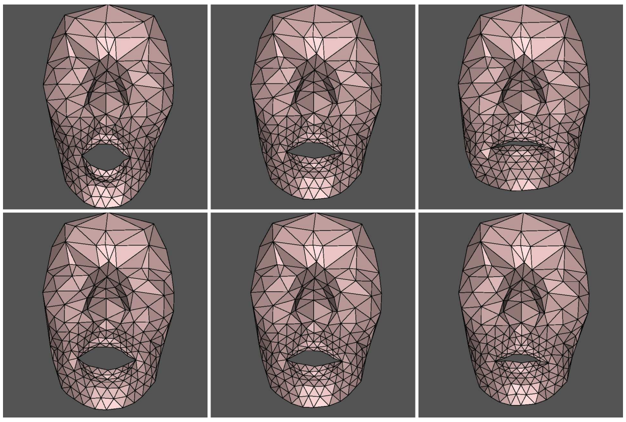

Figure 7. Effect of varying each of the second and third principal modes between ±3 standard deviation; these modes account for the up/down lower jaw motion (19.3% of variance) and for the aperture/closeness of the mouth (3.4% of variance).

2.4 Exploitation of the stereo acquisition system

This system is now used to investigate the variability of labial coarticulation (paper of Robert et al. in these proceedings) and to build a labial coarticulation model for one speaker from a vast corpus of 3D data. The massless property of markers is especially important because the number of markers tracked on lips is substantially higher (46) than in comparable studies. This enables a more precise lip shape information recovery. We also investigated the incorporation of 3D information about the lower jaw and lips into the process of acoustic-to-articulatory inversion. Face parameters provide 3 articulatory parameters (lower jaw position, lip aperture and protrusion) out of the 7 articulatory parameters of Maeda's model as we explain it in the next section. In the future, we intend to investigate 3D tracking without markers as in [2], however, we first need to check if this automatic tracking method is in good agreement with the articulatory motion especially for the lip motion.

3. OUR INVERSION METHOD

Our acoustic-to-articulatory inversion method exploits an articulatory table that associates vectors of articulatory parameters, i.e. 7-tuples in the case of our model, with their corresponding 3-tuples of the first three formant frequencies. This table represents the synthesis facet of the inversion. It is used to recover all the possible 7-tuples of articulatory parameters corresponding to the formant frequencies extracted from a vowel signal at each time frame. The second stage consists of reconstructing articulatory trajectories that are sufficiently regular along time. This is achieved by a dynamic programming algorithm that minimizes a cost function. In the current study, we only minimize the overall velocity of articulators. The final stage improves the articulatory regularity and the acoustical proximity of formants derived from inversion results to the original formants measured on the speech segment.

3.1 Construction of the articulatory table

The strength of our inverse method resides at the quasi-uniform acoustic resolution of the articulatory table. This property originates in the construction method that evaluates the linearity of the articulatory-to-acoustic mapping at each step. The codebook inscribes a root hypercube. Sampling the articulatory space amounts to find reference points that limit linear regions (see [7] for further details).

3.2 Exploration of the null space of the articulatory-to-acoustic mapping

For every acoustic entry given by the first three formants, the inversion process consists in finding all the possible hypercubes, i.e. those for which the articulatory-to-acoustic mapping can produce the 3-tuple of formants observed. Then for each of these hypercubes precise solutions have to be found out. As the inversion consists of recovering seven parameters from the first three formants, the solution space has four degrees of freedom. This means that the null space of the local articulatory-to-acoustic mapping (i.e. the inverse mapping to zero of the linear application corresponding to the Jacobian matrix of the local articulatory-to-acoustic mapping) has to be sampled to get a good description of the solution space. This is not a trivial problem since it corresponds to finding the intersection of a 4-dimensional space and a 7-dimensional hypercube. A first approximation of the intersection is obtained by linear programming. Then the belonging of each articulatory sample to the hypercube is tested [4].

4. VISUAL CONSTRAINTS

In this part, we describe how the visual data we recorded were integrated into the inversion process.

4.1 Integrating visual parameters into Maeda's articulatory model

Maeda's articulatory model has been derived from X-ray sagittal images by applying a factor analysis [5] which enables the explicit choice of linear components that are known to be relevant. It is the case of the jaw whose movements can be readily determined by measuring the position of incisors which appear very clearly on X-ray images. On the other hand the lip stretching cannot be evaluated because no front images were available. 3D data of the speaker's face (see Figure 8) enable the direct measurement of lip opening and stretching by measuring the position of markers painted on the lips. The protrusion is also estimated from these points. However, protrusion corresponds to a complex movement that implies some "unfolding" of the lips. The movements of markers painted on the lips in the mid-sagittal plane thus only partially render this complex movement. Consequently, protrusion is probably slightly underestimated.

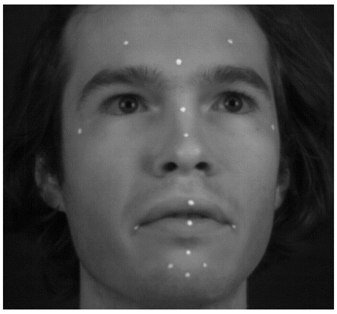

Figure 8. Positions of markers used to evaluate visible articulatory parameters

Contrary to X-ray images, 3D data of the speaker's face do not enable the precise measurement of lower jaw movements. Indeed, the movements of markers painted on the chin depend on the movement of the jaw, but also of that of the lower lip which pulls these markers when it moves. Consequently, the jaw movement cannot be known precisely by tracking points on the speaker's skin.

4.2 Adjusting visual data

From the acquired visual data, we compute four parameters : the mouth opening, lip stretching and jaw movement, which are straightforward to compute, and the lip protrusion, which is more complex. The mouth opening is simply the distance between the points located in the middle of the upper and lower lip. The jaw movement is the distance of the points on the chin to a fixed point (e.g., a point on the temple). In the current study, we assume that the influence of the lip movements is insignificant compared to that of the jaw (which, as said above, is not true). The lip protrusion is the hardest parameter to compute. We compute an estimation using a projection of the points of the upper and lower lip on a plane defined by the average positions of the lip points. Since Maeda's model does not take lip stretching into account, we will not mention it any longer.

4.2.1 Normalization

As our objective is to use the 3D data acquired by the stereovision tracking system as constraints onto the parameters corresponding to the visible articulators of the Maeda's model, we have to derive a direct correspondence between articulatory parameters of the Maeda's model and these data. Geometrical data measured by Maeda were centered and normalized before being processed by the factor analysis procedure. Each of the seven parameters of Maeda's articulatory model is thus allowed to vary over a range of ±3σ (σ is the standard deviation of that parameter). We thus apply the same transformation to 3D data recovered by stereovision.

4.2.2 Decorrelation

The normalization stage enables getting the same dimension for visible parameters and articulatory parameters of Maeda's model. However, the factor analysis used by Maeda allowed the effect of the jaw movement to be removed from other measures. We thus have to remove the influence of the lower jaw from the visible articulatory data. Similarly to Maeda we compute the correlations between jaw and each of the other two visible articulatory measures (lip opening and protrusion) and subtract this correlation from the measure considered. As previously mentioned, contrary to the case of X-ray data where the movement of the jaw is perfectly known through the position of the lower central incisor, here the jaw position is approximated by markers painted on the chin which is not a rigid structure and which is partly influenced by the lower lip.

4.3 Integrating visual data to the inversion process

We thus obtain three parameters compatible with Maeda's model. Unfortunately, due to the fact that the jaw is not known precisely, all visual parameters are imprecise (since their decorrelated value depends on the value of the jaw parameter). Therefore, we have to compensate this imprecision; this is done by accepting inverse solutions having visual parameters in a wider range of values than that that could be used with precisely measured data.

Currently, the use of the visual parameters is very rough; they are only used at the hypercube selection stage: we only sample hypercubes with centers that have visual parameters close enough to the observed parameters. Even this simple choice makes the inversion process significantly faster, and also improves dramatically the realism of solutions. The visual parameters are currently only used at this very early stage in the whole inversion process.

5. EXPERIMENTS

5.1 Corpus

The sentence we are using in this study is extracted from a vast corpus of 3D data recorded for 10 French native speakers (5 female et 5 male speakers), each speaker talked during about 120 seconds. Beside logatoms, our corpus comprises one sentence: "Le joaillier a broyé les cailloux de la voyageuse.", especially designed to evaluate inversion easily since most of the sounds are vowels, semivowels or other voiced sounds.

As the construction of the articulatory codebook used for inversion takes approximately one week on a standard PC we evaluated the audiovisual inversion only for one speaker. The articulatory model was adapted to our speaker using A. Galván-Rodriguez's method [5]. Although we chose a very low acoustic precision in the codebook construction (1 Bark), the resulting codebook had still a very good average acoustic precision: the average RMS error was of about 15Hz on F3.

5.2 Sequence inversion

Here we present some detailed results for two parts of the previous sentence: "joaillier" /Ζοαφε/ and "cailloux" (/καφυ/), since both are voiced. The inversion method performed quite well on both segments, slightly better on the second one. We display here the results obtained without smoothing, because they are more meaningful as regards to the strengths and weaknesses of the current system.

Figure 9 displays the results of inversion on "joaillier": we present the time evolution found for the four main parameters (jaw, mouth opening, lip protrusion and tongue dorsum position). All times are in ms. We also display in solid line the respective visual parameters. As it can be seen on the graph of the jaw parameter, the inversion had trouble in the middle of the sequence to render the /ɑj/ transition at time 21400 ms. The two other visual parameters are fairly close to their constraints. We can also observe that the tongue position trajectory is fairly consistent with what one would expect: it starts in the back of the mouth to pronounce / α/, then goes in the front for /φ/, and goes slightly back for /ε/ (when this parameter increases, it means the tongue gets further back). Due to space limitation we will not display the trajectories of the other articulators.

The "cailloux" (/αφυ/ actually, since the /κ/ cannot be inverted by our system) sequence performed quite well. Figure 10 presents the results for this sequence: all three visual parameters have trajectories very close to their constraints. The fourth parameter displayed is the tongue position. Once again, its trajectory is phonetically realist (the tongue is in the back for /α/, comes closer for /φ/, and goes back for /υ/)

6. CONCLUSION

The acquisition system described here has been used to study inter-speaker labial variability and to investigate how constraints put on visible articulatory parameters can be exploited in audiovisual-to-articulatory inversion. With these constraints the dimension of the null space reduces to one and we are thus studying the relation between inversion results, acoustic precision imposed on formants, adjustments of the articulatory model and phonetic relevancy of inverted articulatory trajectories.

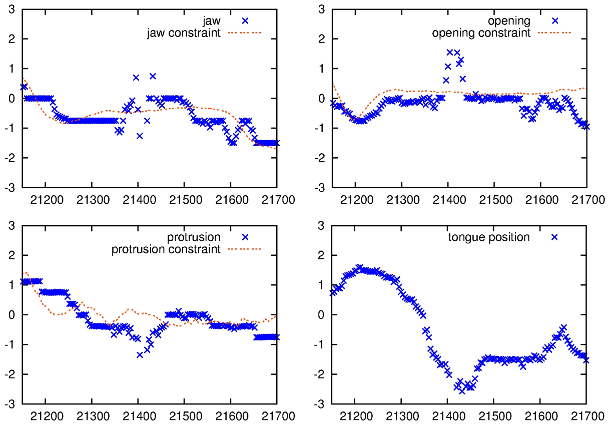

Figure 9. Inversion results for "joaillier". The measured data "constraints" is a red dotted line, inversion data are blue crosses.

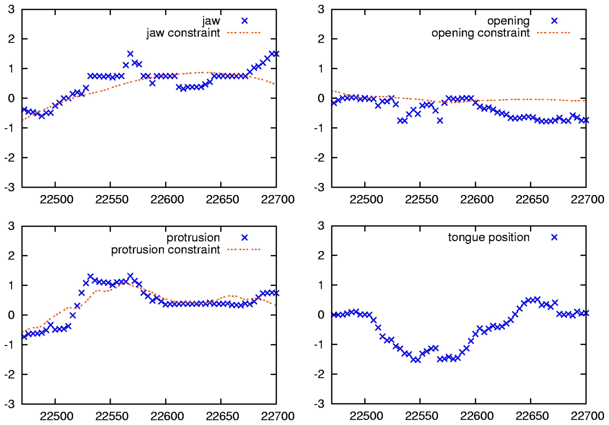

Figure 10. Inversion results for "cailloux".

7. REFERENCES

[1]. J.R. Shewchuk. Triangle: Engineering a 2D Quality Mesh Generator and Delaunay Triangulator. In *Proceedings of the First Workshop on Applied Computational Geometry, Philadelphia, Pennsylvania*, pages 124-133, 1996.

[2]. L. Zhang, N. Snavely, B. Curless, and S. Seitz. Spacetime Faces: High-resolution capture for modeling and animation. In *ACM SIGGRAPH Proceedings*, Los Angeles, CA, pages 548-558, 2004.

[3]. S. Maeda, M. Toda, A. Carlen, and L. Meftahi. Functional Modeling of Face Movements during Speech. In *Proceedings of the International Congress on Speech and Language Processing, Denver,* pages 1529-1532, 2002.

[4]. S.Ouni, and Y.Laprie. Exploring the Null Space of the Acoustic-to-Articulatory Inversion Using a Hypercube Codebook. In *Eurospeech, Danemark*, vol.1, pages 277-280, 2001.

[5]. S.Maeda. Improved articulatory model. In *Journal of Acoustical Society of America*, vol. 81, S146, 1988.

[6]. A.Galván-Rodriguez and R. Laboissière. Speaker normalization using an articulatory model. In *Proceedings of the 15th International Congress of Phonetic Sciences*, pages 1325-1328, 2003.

[7]. Y. Laprie, S. Ouni, B. Potard and S. Maeda. Inversion experiments based on a descriptive articulatory model. In *Proceedings of the 6th International Seminar on Speech Production*, Sydney, 2003.