CAPTURING DATA AND REALISTIC 3D MODELS FOR CUED SPEECH ANALYSIS AND AUDIOVISUAL SYNTHESIS

Frédéric Elisei[1], Gérard Bailly[1], Guillaume Gibert[1], and Rémi Brun[2]

(1) Institut de la Communication Parlée (ICP), UMR CNRS 5009, INPG/U3
46, av. Félix Viallet – 38031 Grenoble – France

(2) Attitude Studio SA – 50, avenue du Président Wilson – Bâtiment 126
93200 La Plaine Saint Denis – France

ABSTRACT

We have implemented a complete text-to-speech synthesis system by concatenation that addresses French Manual Cued Speech (FMCS). It uses two separate dictionaries, one for multimodal diphones with audio and facial articulation, and the other with the gestures between two consecutive FMCS keys ("dikeys"). Dictionaries were built from real data.

This paper presents our methodology and the final results, illustrated by the accompanying videos. We recorded and analyzed the 3D trajectories of 50 hand and 63 facial fleshpoints during the production of 238 utterances carefully designed to cover all possible diphones of French. Linear and non-linear statistical models of hand and face deformations and postures were developed using both separate and joint corpora. Additional data allowed us to capture the shape of the hand and face with a higher spatial density (2,600 points for the hand and forearm and 2,000 for the face), as well as their appearance. We succeeded in building new high-density articulated models that were compatible with the previous emerging set of control parameters. This allows the outputted synthesis parameters to drive the more realistic 3D models instead of the low-density ones.

1. INTRODUCTION

Speech articulation has some clear visible consequences. For example, the movements of the jaw, the lips and the cheeks are perfectly visible. However, the movements of the underlying organs that shape the vocal tract and thus help to shape the sound structure (e.g. larynx, velum and tongue) are not so visible: tongue movements are only weakly correlated with visible movements (R ~ 0.7) [1-3] and this correlation is insufficient for recovering essential phonetic cues such as place of articulation [4, 5].

Hearing-impaired people typically rely heavily on speech reading based on visual information from the lips and face. However, lip reading alone fails to identify many aspects of the place of articulation (for the tongue), manner of articulation (nasality), or voicing. The similarity in the lip shapes of some speech units gives rises to labial look-alikes such as [u] vs. [y]. Indeed, even the best speech readers do not identify more than 50 percent of phonemes in nonsense syllables [6] or in words or sentences [7].

Manual Cued Speech (MCS) was designed to complement speech reading. Developed by Cornett [8] and adapted for more than 50 languages [9], this system is based on the association of speech articulation with cues formed by the hand.

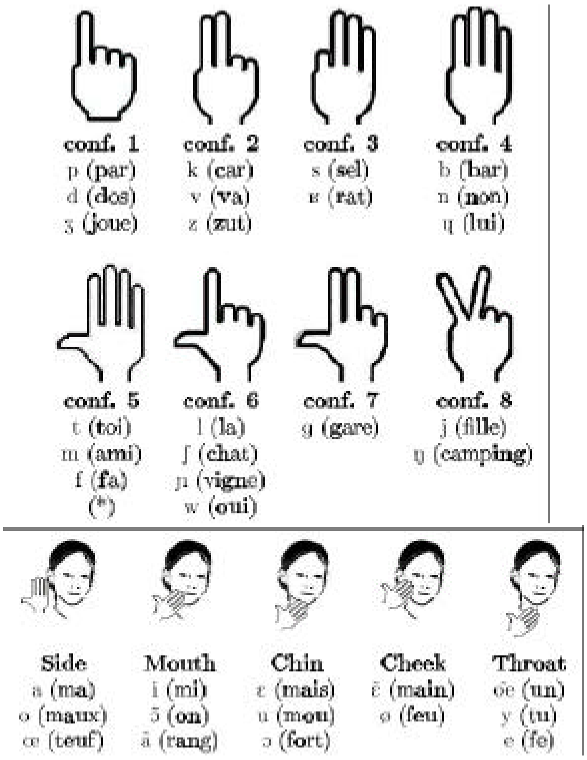

Figure 1. French Manual Cued Speech combines eight hand shapes and five hand placements to complement lip reading.

While speaking, a cued speech user points out specific positions on the face to indicate a subset of vowels, and shapes his or her hand to indicate a consonant group (Figure 1). A large amount of work has been devoted to MCS perception (e.g. [10, 11]), but very few studies have provided insights into MCS production [12-14], which might be useful to address MCS synthesis [15-17].

Section 2 describes the methodology used to gather data from our French MCS speaker. Section 3 addresses the modeling of her hand and face movements, for analysis or synthesis purposes. Section 4 explains how we provide a more realistic rendered appearance. The synthesizer module and its working are not presented in detail here (but see [17] for details).

2. CAPTURING DATA

ICP has developed and used a methodology to construct articulatory models of speech organs, such as the face and the tongue [18, 19]. It consists of an iterative linear analysis using the first principal components of various subsets of fleshpoints, with a subtraction of the successive contributions.

For example, the contributions of the jaw rotation, the lip rounding/spreading gesture, the proper vertical movements of the upper and lower lips, and of the lip corners, as well as the movement of the throat have been retained for facial articulatory cloning. This lead to successful results for several speakers and languages [20, 21].

The corpus consists in static images of the realization of isolated vowels and of consonants in symmetrical context VCV, where V is one the extreme vowels [a], [i] and [u]. Such images were recorded using beads glued on the face, with up to three studio cameras (50Hz, PAL) and two mirrors, as in the clone construction example of Figure 2.

2.1 MCS specificities and difficulties

The hand is a complex organ with many degrees of freedom, where joints, flesh and skin generate non-linear movements. Our three-cameras setup would not be able to capture all bent fingers, nor to track a face partially masked with cued speech keys robustly. Our usual linear-models construction paradigm was not adapted to the hand structure.

Still, we wanted to capture the hand and face movements of a FMCS user with high temporal and spatial resolution. And to compactly represent the variance of the associated trajectory units that will make up the dictionaries, we needed a non-linear articulatory model construction paradigm.

We could not solve the visibility and resolution problems with a unique capture device nor a single capture session. To emulate a "perfect dataset", we recorded several corpora in various conditions, and built a few extra models to merge their qualities.

2.2 Three dynamic corpus

We first used a Vicon© motion capture system with 12 cameras. The basic system delivers the 3D positions of candidate markers at 120Hz, with restrictions about the number of markers or their vicinity. We tracked 113 markers glued on the subject (Figure 3). We recorded three corpora, using two different camera settings:

Corpus 1: Close views of handshapes transitions produced in free space, with every possible transition between the eight consonantal hand shapes.

Corpus 2: A facial-only corpus of dynamic visemes, with all isolated French vowels and the previous VCV.

Corpus 3: A corpus of 238 sentences pronounced while cueing the FMCS.

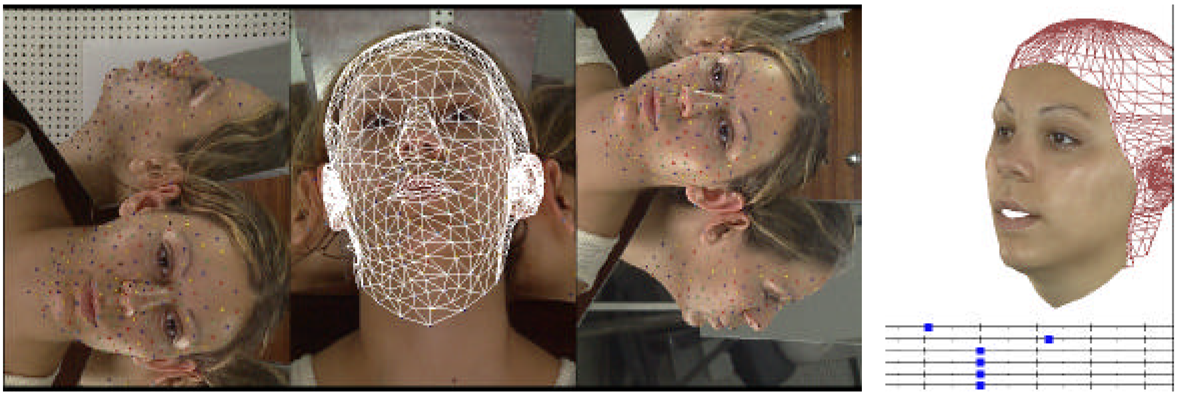

Figure 2. Construction of an articulatory face model. *Left:* The five synchronic view of a static viseme with 247 beads glued on face. *Right:* The clone created with the associated corpus, including the lip model and completed with a generic rigid mesh.

Corpora 1 and 2 are used to build statistical models of the hand and face movements separately. In the carefully designed camera setup, visibility of the markers was almost never a problem.

The models are then used to recover missing data in corpus 3: when cueing the FMCS, face and hand often hides phalanges or face regions.

2.3 High-density data for the face

We used ICP's three cameras setup to collect a spatially-denser set of articulations for our FMCS subject (Figure 2). The 3D coordinates of 247 face fleshpoints were reconstructed, along with the lip shapes. To capture natural-looking textures, a special set of images featuring fewer glued beads and more view angles was also recorded.

2.4 High-density data for the hand

The right hand of our subject was molded with alginate (Figure 4), capturing also the position of the reflective markers. The resulting mold was then digitized to capture the geometry and scales.

3. MODELS FOR FMCS

Constructing models and defining control parameters for the trajectories of the captured points have several practical benefits (data variance reduction, outlier filtering). Scientific motivations concern the study of FMCS: they offer a unique way for studying the cued speech production and the specific laws governing the coordination between acoustics and face/hand movements.

3.1 Low-density animated face model

Our basic articulatory face-cloning methodology was previously applied to heads alone, the quasi-static movements of which could be subtracted for analysis and synthesis. Since the head is now moving much more (to accompany the hand gesture) in Corpora 2 and 3, we need to solve the problem of the progressive deformation seen on the 18 markers placed on the throat. This problem is solved in three steps:

1. Estimation of the head movement using the hypothesis of a rigid motion of markers placed on the nose and forehead. A principal component analysis of the six parameters of the roto-translation extracted for Corpus 3 was then performed and the *nmF* first components were retained as control parameters for the head motion.
2. Facial motion cloning with the inverted rigid motion of the full data. Only *naF* components were retained as control parameters for the facial motion.
3. Throat movements were considered to be equal to head movements weighted by factors less than one. A joint optimization of these weights and the directions of *nmF* facial deformations was then performed, keeping the same values for the *nmF* and *naF* predictors.

These operations were performed using the facial data from Corpora 2 and 3 where all markers were visible. A simple vector quantization guaranteeing a minimum 3D distance between selected training frames (equal here to 2mm) was performed before modeling. This pruning step provided statistical models with conditioned data.

The final algorithm for computing the 3D positions *P3DF* of the 63 face markers of a given frame is:

```
mvt = mean_mF + pmF*eigv_mF;
P3D = reshape(mean_F + paF*eigv_F,3,63);
FOR i := 1 TO 63
  M = mvt.*wmF(:,i);
  P3DF(:,i) = Rigid_Motion(P3D(:,i),M);
END
```

where *mvt* is the head movement controlled by the *nmF* parameters *pmF*, *M* is the movement weighted for each marker (equal to 1 for all face markers, less than 1 for markers on the throat) and *P3D* are the 3D positions of the markers without head movements controlled by *naF* parameters *paF*.

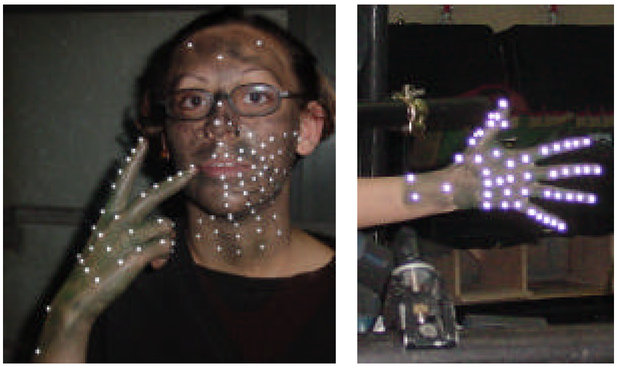

Figure 3. Our female FMCS user and the reflective markers used for the dynamic corpus. She wears 50 markers on her right hand and 63 markers on her face, mainly on her left side.

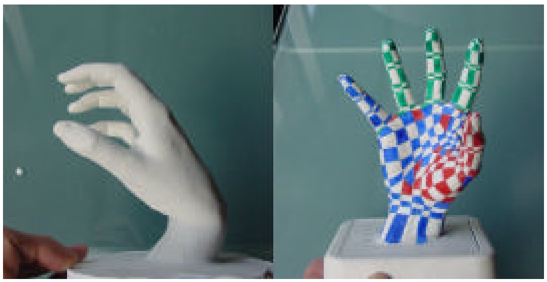

Figure 4. Alginate molds of the right hand of our subject.

3.2 Low-density, animated hand model

Building a statistical model of hand deformations was more complex. We considered the palm to be the carrier of the hand, giving the basic rigid movement for the 50 markers. This movement was computed using 11 markers placed on the back on the hand. The hand model was built in four steps:

1. Estimation of the hand movement using the hypothesis of a rigid motion from markers of the palm in corpus 1. A principal component analysis of the six parameters of this hand motion was then performed and the *nmH* first components were retained as control parameters for the hand motion.
2. All possible angles (rotation, twisting, spreading) between each hand segment and the palm as well as between successive phalanges (using the inverted rigid motion of the full hand data) were computed.
3. A principal component analysis of these angles was performed. The *naH* first components were retained as the hand shaping control parameters.
4. We then computed the sin() and cos() of these predicted values and performed a linear regression between these *2*naH+1* values (see vector *P* below) and the marker coordinates.

Step 4 allows an elliptic movement with scaling to better capture the complex displacement induced on the skin surface by the distant pure-joint rotation.

The final algorithm for computing the 3D positions *P3DH* of the 50 hand markers for a given frame is:

```
mvt = mean_mH + pmH*eigv_mH;
ang = mean_A + paH*eigv_A;
P = [1 cos(ang) sin(ang)];
P3DH = Rigid_Motion(reshape(P*Xang,3,50),mvt);
```

where *mvt* is the forearm movement controlled by the *nmH* parameters *pmH,* and *ang* is the set of angles controlled by the *naH* parameters *paH.*

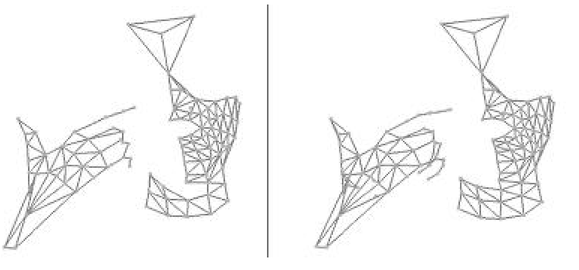

Figure 5. Missing data on a captured frame (left, with truncated bent fingers and incomplete left throat region) is filtered and reconstructed with the two articulatory models (right image).

3.3 Resulting low-density models

More than 8,000 frames in less than half the sentences were used to define the hand model. We retained *naH=12* parameters (emerging from the 23 possible angles) for the hand shape, and *nmH=6* parameters for the hand movements. Mean reconstruction error is around 1.2 mm, also with sentences not used for the model learning.

We used almost 5,000 frames to construct the facial model. We retained *naF=7* articulatory parameters, and *nmF=6* displacement parameters. Mean reconstruction error is around 1 mm.

Using both models, we can recover missing data, as illustrated in Figure 5. These models can be used to analyze the dynamic corpus and the FMCS phasing of the organs recruited. Indeed, we verified in [17] that cued speech information was delivered well in advance of the lip reading information, as in [12].

The articulatory models can also encode the trajectories to be concatenated by synthesis. But they are not sufficient for subjective evaluation purposes (even point-light based ones), as the lips could only be captured by their external contours: mouth always looks open even when it is not.

4. RENDERING MODELS

For communication applications or evaluation tasks, we need to provide better shapes for the hand and a completed face (with both sides and lips), without sacrificing the already captured dynamics quality.

4.1 Compatible high-density hand model

Algorithms classically used in computer graphics software (such as *Maya*™) were used to scale a generic articulated 3D hand model to our subject's personal geometry and dimensions (digitized mold), to derive a "pseudo-bone structure" and to link the displacements of the external markers to some of those for the pseudo-bones and the skin (Figure 6).

We chose 128 low-density postures that still span the observed articulatory space. High-density versions of the same postures were generated with a skinning algorithm. Though numbering 2,600 points each, these static shapes should be modeled by the exactly same *pmH* articulatory parameters as their lower density counterparts. A linear regression captures in a new *Xang* matrix how the 3D coordinates can be explained by the shared set of *P* values. Residual error is important in the forearm region, which is far from the one we tried to model.

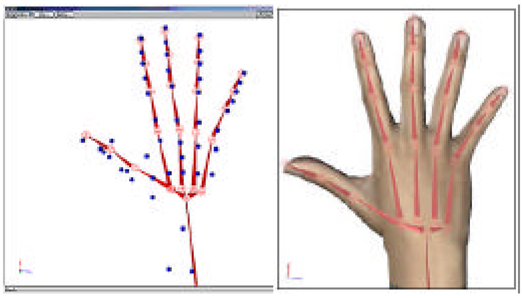

Figure 6. Markers can reshape and animate a generic hand model. *Left:* Markers (blue dots) can drive a joints/bones structure (red). *Right:* In the skinning algorithm, the pseudo-bones will influence the vertexes of the textured 3D object.

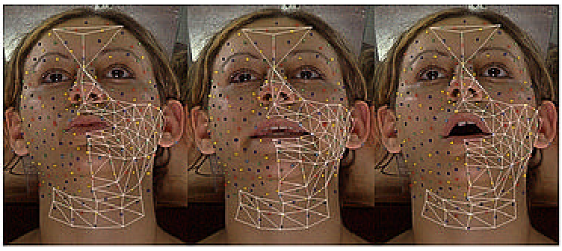

Figure 7. Fitting the low-density face model learned with the dynamic corpus on various static visemes (high-density face corpus). Although the side views are not shown of this picture, they were used in the fitting process.

4.2 Compatible high-density face model

For the face, we could not construct a synthetic corpus for a set of given control parameters, as we did for the hand. Instead, we had to use a pertinent real corpus: the static images with 247 glued beads. We had to find the set of parameters (position and articulation) for the low-density face model that would best fit the surface. This was achieved by manually defining a few matches between beads and markers or pseudo-markers (defined as linear combinations of neighbor markers). Once the parameters associated with the set of high-density 3D positions (beads and lip points) were inverted, a linear regression constructed the new head model. It inherits the features of the model: 247 articulated facial points, an articulated lip model and 1,700 extra points for rigidly linked parts. With the available corpus, the inverted high-density throat region is not influenced by head movements and will not seem connected to shoulders.

4.3 Rendering results

Figure 8 shows the inter-operability of the low-density and high-density models. For captured or synthesized frames, the same set of parameters can drive either version of the hand or face models.

Videos on the conference CD illustrate the perceived rigidity between the different models, for captured streams as well as synthetic ones. Please note the specific prosody, modeled and synthesized using a superposition of functional contours [22] that were learned on the dynamic corpus presented here. To smoothly connect the hand trajectory units, a specific filtering method was also designed for the synthesizer [17].

5. CONCLUSIONS, PERSPECTIVES

The observation of measured cued speech is a prerequisite for developing FMCS technologies, for communication but also educational purposes. The methodology and models presented here have already been used to study the phonetic structure of cued speech, notably the phasing relations between hand gestures and sound production. The hand and face gesture scores were also studied in reference to acoustic segmentation. This knowledge is embedded in ICP's multimodal text-to-FMCS system.

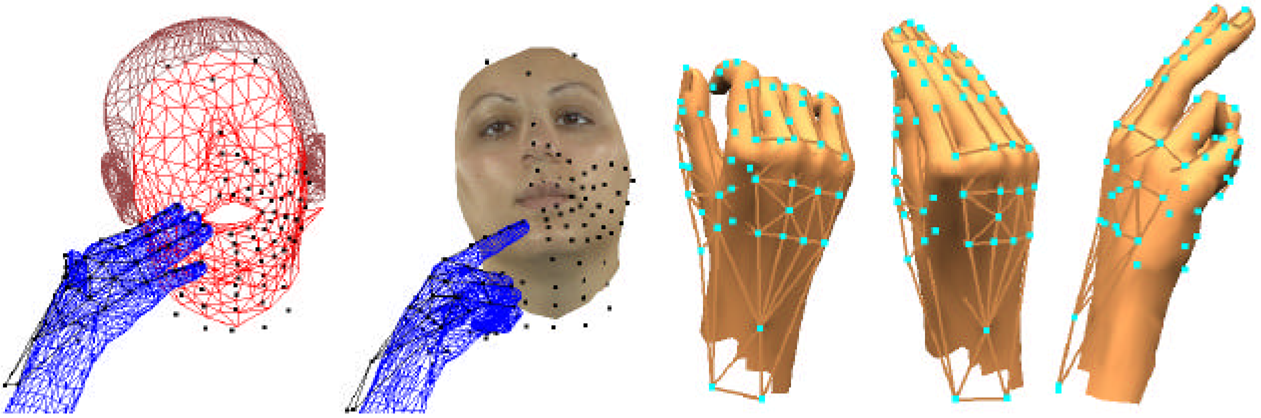

Figure 8. Rendering various sets of control parameters with the captured and reconstructed models. Models can be rendered as wire-frame, textured or shaded surfaces. Square dots (black or light blue) are from the captured low-density models, and correspond to the reflective markers. They are drawn connected to be more visible, though this does not match the real surface curvatures. Note how the model diverges for the forearm and the throat, where some extra degrees of freedom were lost (face) or not retained (hand). With the lip model, closed mouths can be perceived correctly and the protrusion becomes more visible.

Within the ARTUS project, ICP and Attitude Studio collaborate with academic and industrial partners to provide the French-German TV channel Arte with on-demand MCS dubbing. Face and hand parameters were first synthesized from existing textual subtitle data, then compactly watermarked within the broadcasted video and acoustic channels, and rendered locally by enhanced television sets. Further study and better understanding of the kinematics of the different segments involved in the production of MCS might lower the number of necessary parameters. We also need to remove the suprasegmental-related head and posture movements. Subjective evaluations of the reconstructed gestures and the synthesized ones is also planned, with point-lights or realistic rendering.

ACKNOWLEDGMENTS

Many thanks to Yasmine, our CS speaker, and to C. Corréani, X. Jacolot, J. Meunier, F. Vandenbergh and F. Vayssettes who processed the raw motion capture data. We acknowledge also V. Attina for her cued speech expertise when needed. Parts of this work were financed by the ARTUS RNRT project.

6. REFERENCES

[1] Yehia, H.C., Rubin, P.E., and Vatikiotis-Bateson, E. (1998) Quantitative association of vocal-tract and facial behavior. Speech Communication, 26: p.23-43.

[2] Kuratate, T., Munhall, K.G., Rubin, P.E., Vatikiotis-Bateson, E., and Yehia, H. (1999) Audio-visual synthesis of talking faces from speech production correlates. in EuroSpeech. p.1279-1282.

[3] Jiang, J., Alwan, A., Bernstein, L., Keating, P., and Auer, E. (2000) On the Correlation between facial movements, tongue movements and speech acoustics. in International Conference on Speech and Language Processing. Beijing, China. p.42-45.

[4] Bailly, G. and Badin, P. (2002) Seeing tongue movements from outside. in International Conference on Speech and Language Processing. Boulder - Colorado. p.1913-1916.

[5] Engwall, O. and Beskow, J. (2003) Resynthesis of 3D tongue movements from facial data. in EuroSpeech. Geneva

[6] Owens, E. and Blazek, B. (1985) Visemes observed by hearing-impaired and normal-hearing adult viewers. Journal of Speech and Hearing Research, 28: p.381-393.

[7] Bernstein, L.E., Demorest, M.E., and Tucker, P.E. (2000) Speech perception without hearing. Perception & Psychophysics, 62: p.233-252.

[8] Cornett, R.O. (1967) Cued Speech. American Annals of the Deaf, 112: p.3-13.

[9] Cornett, R.O. (1988) Cued Speech, manual complement to lipreading, for visual reception of spoken language. Principles, practice and prospects for automation. Acta Oto-Rhino-Laryngologica Belgica, 42(3): p.375-384.

[10] Nicholls, G. and Ling, D. (1982) Cued Speech and the reception of spoken language. Journal of Speech and Hearing Research, 25: p.262-269.

[11] Uchanski, R., Delhorne, L., Dix, A., Braida, L., Reed, C., and Durlach, N. (1994) Automatic speech recognition to aid the hearing impaired: Prospects for the automatic generation of cued speech. Journal of Rehabilitation Research and Development, 31: p.20-41.

[12] Attina, V., Beautemps, D., and Cathiard, M.-A. (2002) Coordination of hand and orofacial movements for CV sequences in French Cued Speech. in International Conference on Speech and Language Processing. Boulder - USA. p.1945-1948.

[13] Attina, V., Beautemps, D., and Cathiard, M.-A. (2003) Temporal organization of French Cued Speech production. in International Conference of Phonetic Sciences. Barcelona, Spain

[14] Attina, V., Beautemps, D., Cathiard, M.-A., and Odisio, M. (2004) A pilot study of temporal organization in Cued Speech production of French syllables: rules for a Cued Speech synthesizer. Speech Communication, 44: p.197-214.

[15] Duchnowski, P., Lum, D.S., Krause, J.C., Sexton, M.G., Bratakos, M.S., and Braida, L.D. (2000) Development of speechreading supplements based on automatic speech recognition. IEEE Transactions on Biomedical Engineering, 47(4): p.487-496.

[16] Attina, V., Beautemps, D., Cathiard, M.-A., and Odisio, M. (2003) Towards an audiovisual synthesizer for Cued Speech: rules for CV French syllables. in Auditory-Visual Speech Processing. St Jorioz - France. p.227-232.

[17] Gibert, G., Bailly, G., and Elisei, F. (2004) Audiovisual text-to-cued speech synthesis. in 5th Speech Synthesis Workshop. Pittsburgh, PA. p.85-90.

[18] Revéret, L. and Benoît, C. (1998) A new 3D lip model for analysis and synthesis of lip motion in speech production. in Auditory-visual Speech Processing Workshop. Terrigal, Australia. p.207-212.

[19] Badin, P., Bailly, G., Revéret, L., Baciu, M., Segebarth, C., and Savariaux, C. (2002) Three-dimensional linear articulatory modeling of tongue, lips and face based on MRI and video images. Journal of Phonetics, 30(3): p.533-553.

[20] Elisei, F., Odisio, M., Bailly, G., and Badin, P. (2001) Creating and controlling video-realistic talking heads. in Auditory-Visual Speech Processing Workshop. Scheelsminde, Denmark. p.90-97.

[21] Odisio, M. and Bailly, G. (2004) Shape and appearance models of talking faces for model-based tracking. Speech Communication, 44(1-4): p.63-82.

[22] Holm, B. and Bailly, G. (2002) Learning the hidden structure of intonation: implementing various functions of prosody. in Speech Prosody. Aix-en-Provence, France. p.399-402.