ARTUS: Synthesis and Audiovisual Watermarking of the Movements of a Virtual Agent Interpreting Subtitling using Cued Speech for Deaf Televiewers

Gérard Bailly[1*], Virginie Attina[1], Cléo Baras[7], Patrick Bas[2], Séverine Baudry[3], Denis Beautemps[1], Rémi Brun[4], Jean-Marc Chassery[2], Frank Davoine[5], Frédéric Elisei[1], Guillaume Gibert[1], Laurent Girin[1], Denis Grison[6], Jean-Pierre Léoni[6], Joël Liénard[2], Nicolas Moreau[7], Philippe Nguyen[3]

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

(2) Laboratoire Image et Signaux, CNRS/INPG, BP 46, 38402 Saint Martin d'Hères – France

(3) Nextamp, 12, square du Chêne Germain, 35510 Cesson-Sévigné – France

(4) Attitude Studio, 50 avenue du Président Wilson , 93214 St Denis-la-Plaine – France

(5) Heudiasyc, CNRS/UTC, Centre de Recherches de Royallieu, 60205 Compiègne – France

(6) ARTE, 8, rue Marceau, 92785 Issy-les-Moulineaux Cedex 9 – France

(7) ENST, 46, rue Barrault, 75013 Paris – France

Abstract. The ARTUS project provides deaf televiewers with an alternative substitute for subtitles using Cued Speech: an animated agent can be superimposed - on demand and at the reception - to the original broadcast. The hand and face gestures of the agent are generated automatically by a text-to-cued speech synthesizer and watermarked in the broadcasted audiovisual signals. We describe here the technological blocks of our demonstrator. First evaluation of the complete system by end-users is presented.

I. Introduction

The ARTUS project offers a new service for deaf televiewers: a virtual talking head (cf. Figure 1)., that communicates using cued speech, can be superimposed on demand on the end-user TV screen. The movements of the face, head and hands of this virtual character are computed automatically from pre-existing subtitles. These movements are then compressed and inserted into the audiovisual content of the original TV program using watermarking techniques. This paper describes the different components of the ARTUS demonstrator presented to the TV company ARTE in

[*] Corresponding author : gerard.bailly@icp.inpg.fr

December 2005 and currently under evaluation by a panel of deaf televiewers with a good level of cued speech practice.

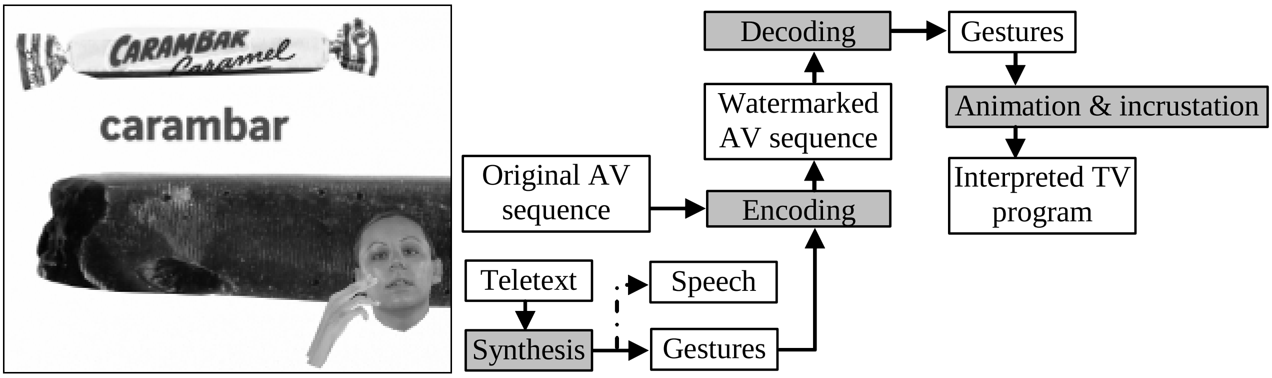

Figure 1. Incrusting the ARTUS virtual speech cuer in a TV program. Left: resulting display for the deaf televiewers. Right: synopsis of the system. Note that synthetic speech is also computed and could be broadcasted for enhanced multimodal help (e.g. blind people).

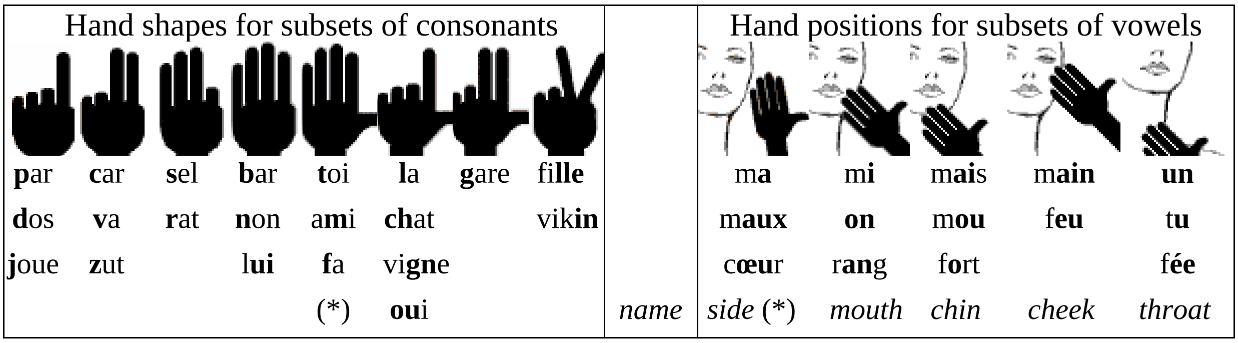
Hand shapes for subsets of consonants
Hand positions for subsets of vowels
par car sel bar toi la gare fille
dos va rat non ami chat vikin
joue zut lui fa vigne
(*) oui
name
ma mi mais main un
maux on mou feu tu
cœur rang fort fée
side (*) mouth chin cheek throat

Figure 2. Hand shapes and positions used to cue subsets of French consonants and vowels. The stars indicate default positions (see text for explanation).

II. French cued speech

Listeners with hearing loss and orally educated typically rely heavily on speechreading based on lips and face visual information. However speechreading alone is not sufficient due to the lack of information on the place of tongue articulation and the mode of articulation (nasality or voicing) as well as to the similarity of the lip shapes of some speech units (so called labial *sosies* as [u] vs. [y]). Indeed, even the best speechreaders do not identify more than 50 percent of phonemes in nonsense syllables as well as words or sentences (Bernstein, Demorest et al. 2000).

Cued Speech (CS) was designed to complement speechreading. Developed by Cornett (1967) and adapted to more than 50 languages (Cornett 1994), this system is based on the association of speech articulation with cues formed by the hand. While uttering, the speaker uses one of his hand to point out specific positions on the face (indicating a subset of vowels) with a hand shape (indicating a subset of consonants). Numerous studies have demonstrated the drastic increase of intelligibility provided by CS compared to speechreading alone (Nicholls and Ling 1982; Uchanski, Delhorne et al. 1994) and the effective facilitation of language learning using CS (Leybaert and Alegria 2003).

The French CS (FCS) system is described in Figure 2: it distinguishes between 8 subsets of consonants and 5 subsets of vowels that groups sounds of the language that may be further distinguished by the lip reading. A consonant-vowel sequence is coded in one gesture where the hand points to parts of the face with a certain hand shape. Isolated vowels or consonants in initial or final clusters are coded respectively with default configurations (see Figure 2). When recording our target talker, we imposed a rest position with the hand far away from the face and with a closed fist.

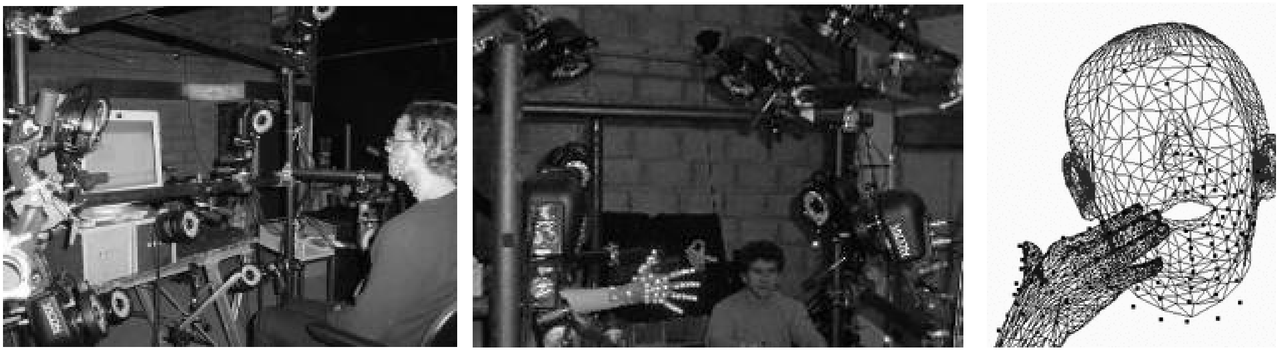

Figure 3. Capturing and cloning face and hand movements of a human cuer. Left and centre: the experimental settings. Right: the resulting shape model used to animate the virtual speech cuer.

III. Description of the proposed system

The system is described in Figure 1. Numerous technological and scientific bolts have been faced:

- Analysis and modeling of the FCS gestures produced by numerous cuers (Attina, Beautemps et al. 2004) and in particular by our female target speaker.
- Development of a multimodal subtitling-to-FCS synthesis system (Gibert, Bailly et al. 2005)
- Joint coding of head, face and hand gestures using matrix quantization (Girin 2004)
- Watermarking of audio and video streams
- Real-time videorealistic animation of a 3D virtual clone (Elisei, Bailly et al. 2005)

Key features of these different modules are detailed below.

A. Analyzing FCS gestures

Our system is based on the analysis and modelling of hearing FCS talker. She is a non-professional with a dozen years of intensive practice in her familial environment. The evaluation of her cueing performance (less than 1% of cueing errors and high intelligibility scores) demonstrates her excellent proficiency.

The head, face and hand gestures of our target FCS talker have been captured using a VICON® system with 12 cameras (cf. Figure 3). The positions of 113 semi-spherical reflective markers glued on her right hand and the left part of her face were recorded at 120 Hz synchronously with the speech sound. In addition to elementary gestures used for building statistical shape models, she uttered 239 phonetically balanced sentences that were designed so as to retrieve at least two instances of each French diphones. This resource is then used for multimodal synthesis.

The analysis of these movements confirms the hand/lips synchronization evidenced by Attina et al (2004; 2006): (a) deployment of hand shape and position are almost synchronous; (b) target hand shape and position for a CV unit is reached around the acoustic target of the consonant C… and thus FCS cues for the vowel are available well in advance from the lip information provided by the segment. Our cuer however departs from professional cuers in the contribution of head movements to hand/face contacts: if the arm contributes essentially to the hand displacement, the head contributes significantly to the narrowing (7.7% on average). This more important use of postural movements have been quoted for early learners of sign language (Boyes Braem 1999).

B. Modeling FCS gestures

Statistical shape models for the face and the hand of our target speaker have been developed: We reduced the 3*113 (xyz coordinates of our fleshpoints) degrees of freedom of our motion capture data to respectively 9 and 7 elementary gestural components for the face and hand. 12 additional translation and rotation parameters are added for decomposing the movements of the head and the arm. This reduction of dimension that preserves geometric accuracy (respectively less than 1 and 1.5 millimeters RMS error for predicted positions of face and hand fleshpoints) is done using series of principal component analysis and linear regressions on geometric or angular data (see Gibert, Bailly et al. 2005, for details). These statistical shape models have been used to regularize the noisy and incomplete motion capture data.

C. Synthesizing FCS gestures

Several components have been added to the multimodal text-to-speech synthesis system COMPOST (Bailly and Alissali 1992) to cope with subtitle input and cued speech output.

Subtitling consists of chunks of text (not always corresponding to sentences and often with little punctuation) with the time interval allocated to its display on screen. End-of-sentence detection exploits this chunking information and delivers time checkpoints to the rhythmic model mentioned below that copes with inter-sentence pause generation.

Prosody generation computes multimodal cues that users extract to ease their analysis of discourse structure. Prosody includes the organization of phoneme durations, audible cues such as pauses or melody and visible cues such as head movements. Even though experienced speech cuers can minimize the impact of hand gestures on speech rate, the inertia of the arm imposes a slower articulation especially for complex syllables. A trainable prosodic generator (Bailly and Holm 2005) has been fed with cued speech data. Its main task is here to generate phoneme durations and thus to allocate time intervals to speech chunks. This allocation may overestimate the time interval dedicated by the authors of subtitling. However in the three TV programs we have processed so far, only 4 sentences could not be pronounced in the time interval devoted to their initial subtitling. The average delay for these sentences was 120ms (corresponding roughly to one syllable).

Generation of speech and gestures is performed by selecting, smoothing, stretching and concatenating speech segments. Two types of segments are considered: “polysounds” that capture the signal and lip gestures from one stable part of a sound to the next (sounds such as semi-vowels and glides are not considered as having a sufficiently stable part and are included into larger segments) and “dikeys” that encompass hand gestures from one target gesture to the next. The warping function that maps original short-term speech signals to the synthetic time axis is also used for facial movements such as the proper synchronization between lip movements and their acoustic consequences is preserved. Another warping function is also computed to map original head and hand movements that respect the timing of the gestural targets computed according to the average phasing relations between speech and FCS gestures evidenced in section A, i.e. the alignment of gestural targets with acoustic targets.

Continuous gestures are generated using an anticipatory smoothing procedure (Bailly, Gibert et al. 2002) that preserves initial targets of adjacent segments

D. Coding FCS gestures

An additional dimensionality reduction of the gestures is performed in order to further reduce the necessary transmission bandwidth (typically a few 100 bits/s offered by up-to-date audiovisual watermarking techniques). Separate matrix quantization of head, face, hand and arm movements is thus performed that encodes blocks of several frames of articulatory parameters. Training consists in identifying typical blocks that are shared between the coder and the decoder. Indexes in the table of blocks are determined so that RMS error between the original parametric sequence and the series of chosen blocks is minimized. They are then transmitted via on-line watermarking. The variance explained by each parameter is used as a weighting factor for RMS calculation.

This technique has been applied to each speech segment stored in the polysounds and dikeys dictionaries. This choice may cause problems to live applications where gestures are captured and encoded on-line. In case of concatenative synthesis, coding is performed within a limited set of realizations. Finally, blocks of 80ms are used and the sizes of the dictionaries of blocks have been designed so as to reach a rate of 200 bits/s with an acceptable distortion.

E. Watermarking

Watermarking techniques aim at inserting imperceptible and indelible information in carrier signals (invisible in video and inaudible in audio). Classically this information is used to assert intellectual property on parts of the signals. In our case, watermarking is used to encode a continuous and synchronous flow of encoded FCS parameters. The application is then termed as delivering an “augmented content” or a “hidden channel”.

Watermarking is based on the basic principles of numeric transmission that consists of transmitting as many bits as possible through a noisy channel. The only difference here is that the “noise” is the audiovisual signal itself and that the signal-to-noise ratio is very small since watermarks should be imperceptible. Since watermarks should be robust to perturbations of the transmission channel (audiovisual compression, filtering, resampling, changes between analogic vs. numeric formats) and guaranty a very low bit error rate (BER), the bit rate is thus often limited to a few hundred bits per second. The most popular watermarking technique is the Direct Sequence Spread Spectrum (DSSS) that consists in encoding each bit by a short-term signal that is just added – and adapted - to the carrier signal.

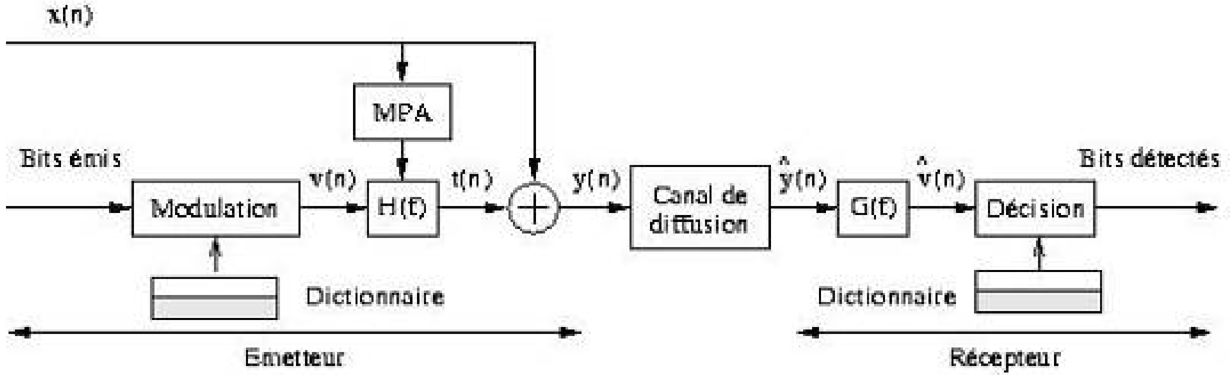

Figure 4 : Synopsis of the audio watermarking. The masking filters H(f) and G(f) are estimated from the carrier audio signal.

For the ARTUS project, a substitutive watermarking technique (Bas, Chassery et al. 2002) is used for the video channel and adaptive modulation (LoboGuerrero 2004; Baras 2005) for the audio channel. The substitutive watermarking technique modulates the average luminance of blocks (here 32x32 pixels) of the image taking into account the average luminance of neighboring blocks. Adaptive audio modulation uses well-known perceptual masking effects to shape spectrum of short-term watermarks (see Figure 4) so that inaudibility is guaranteed.

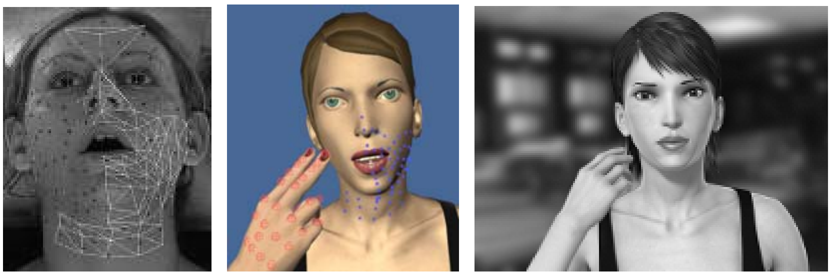

Figure 5. Video realistic shape and appearance face model. From left to right: the face model trained with motion capture data superposed to a high definition image of the speech cuer, then to a first virtual avatar; finally a the final video realistic rendering with a background video.

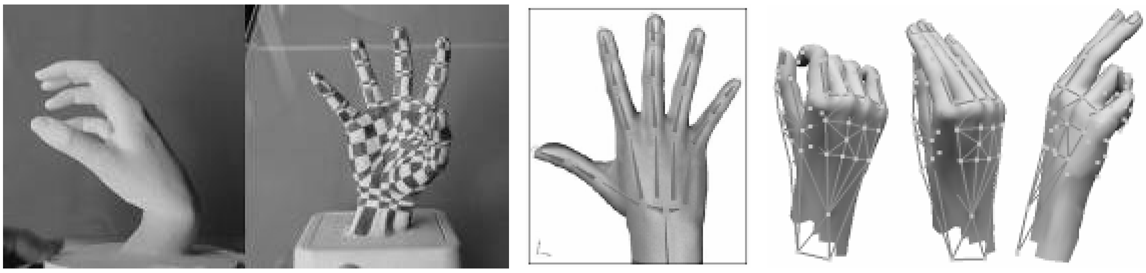

Figure 6. Video realistic shape and appearance hand model. From left to right: sample cast of the hand of our speech cuer, definition of the reference mesh, result after skinning and control by the model trained using motion capture data.

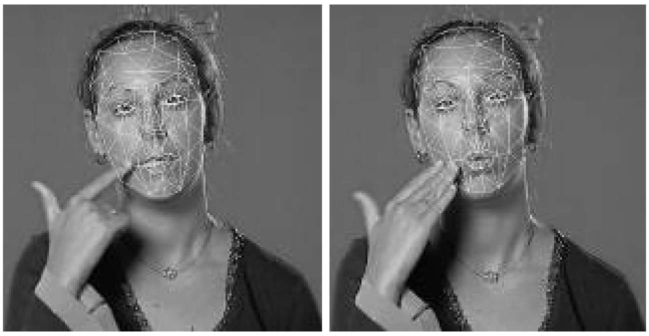

Figure 7. Tracking gestures using an adapted 3D face model with partial occlusions by the hand.

F. Videorealistic animation

Diverse models of the articulated shape of the upper body and of its video realistic appearance have been developed either from photogrammetric data of the original speaker or by adapting a generic model to her morphology (Figure 5). From casts of the hand of the cuer in various postures, a subject-specific high definition shape model of her hand has been built using classical “skinning” algorithms. Since reflexive markers used for motion capture are clearly visible, it is possible to drive the shape model from the sparse motion capture data (see Figure 6).

G. Tracking markerless cuers

Methods for tracking markerless face and hand gestures have also been developed for watermarking interactive shows in real-time: gestures of a human FCS transliterator might be estimated from video, scaled, encoded and rendered using the ARTUS virtual cuer. Robust tracking methods using robust Active Appearance Models (Cootes, Edwards et al. 2001) and stochastic filtering techniques (Dornaika and Davoine 2005) compensate for rotation and partial occultation of the face. This tracking is for now performed at a rate of 4.5 images per second using non optimized C/C++ code running on a modern PC with a Intel chip at 3,6 GHz. Jaw, lips and eyebrows are tracked. Hand gestures are under consideration.

IV. Evaluation of the cued speech synthesis system

A first series of experiments have been conducted to evaluate the intelligibility of this virtual cuer with skilled deaf users of the French cued speech (Gibert, Bailly et al. 2006). The first evaluation campaign was dedicated to segmental intelligibility and the second one to discourse comprehension.

A. Segmental intelligibility

The test mirrors the Modified Diagnostic Rime Test developed for French (Peckels and Rossi 1973): the minimal pairs do not here test acoustic phonetic features but gestural ones. A list of CVC word pairs – all current French words – has thus been developed that test systematically pairs of consonants in initial positions that differ almost only in hand shapes: we choose the consonants in all pairs of 8 subsets of consonants that are highly visually confusable (Summerfield 1991). Due to the fact that minimal pairs cannot be found for all vocalic substrates, we end up with a list of 196 word pairs. We compared lip-reading performance (displaying only the speaking face) with cued speech (displaying hand movements).

Mean intelligibility rate for lipreading condition is 52.36%. It is not different from hazard: minimal pairs are undistinguishable. Mean intelligibility rate for “CS” condition is 94.26%. The difference in terms of intelligibility rate between these two conditions shows our virtual cuer gives significant information in terms of hand movements. In terms of cognitive efforts, the “CS” task is easier: the response time is significantly lower ($F(1,3134)=7.5$, $p<0.01$) than for the lipreading condition.

B. Long-term comprehension

To evaluate the global comprehension of our system, we asked the same subjects to watch a TV program where subtitles were replaced by the incrustation of the virtual cuer. Ten questions were asked. The results show that all the information is not perceived. On average, the subjects replied correctly to 3 questions. Differences between performances of the virtual cuer and the human interpreter were not statistically significant. The difficulties of the task (proper names, high speaking rate) could certainly explain these results.

We conducted further experiments using a Tobii© eye tracker. We asked 4 deaf people to watch a TV program divided in 2 parts: one part subtitled and another part with the inlay of a cuer video. The results show the subjects spend 56.36% of the time on the teletext and 80.70% on the video of the cuer with a significant difference $F(1,6)=9.06$, $p<0.05$. A control group of 16 hearing people spend 40.14% of the time reading teletext. No significant difference was found.

C. Comments

The results of the preliminaries perceptive tests show significant linguistic information with minimal cognitive effort is transmitted by our system. This series of experiments must be continued on more subjects and other experiments must be added to quantify exactly the cognitive effort involved. Discourse segmentation and part of speech emphasis by multimodal prosodic cues (not yet identified nor implemented) is expected to reduce this effort.

V. Conclusions

ARTUS proposes a new audiovisual service to deaf televiewers. ARTUS encodes indelible and imperceptible information in the original audiovisual documents. This additional information channel is used to transmit gestural scores that are used to animate the upper body of a virtual avatar cueing speech. A first demonstrator that exploits the diverse modules described in this paper has been developed and is now under active assessment by content providers and end-users. Additional data on real-world transmission performance and further comparative studies with other alternatives are required before a full deployment of this technology. The impact of this technology on the long-term attention and cognitive load impacted to televiewers should also be quantified, especially in case of a long exposure to augmented audiovisual contents.

This system and parts of these techniques could benefit to other linguistic system (sign languages) or be used to carry iconic (insertion of icons) or paralinguistic information for the avatar (gaze, facial expressions, pointing gestures, etc). Note finally that this system could be extended to other situations (augmented face-to-face communication, etc), to other communication supports (interactive CDs) and to other applications of digital entertainment (computer-assisted CS learning).

VI. Acknowledgements

We thank our subject Yasmine Badsi for her high valuable contribution. We also thank our project officer Ilarion Pavel and scientific advisor Jean-Jacques Rigoni from ElanSoft for their support.

VII. References

Attina, V. (2006). La Langue française Parlée Complétée (LPC) : Production et Perception. Grenoble - France, Institut National Polytechnique.

Attina, V., D. Beautemps, et al. (2004). "A pilot study of temporal organization in cued speech production of French syllables: rules for a cued speech synthesizer." Speech Communication **44**: 197-214.

Bailly, G. and M. Alissali (1992). "COMPOST: a server for multilingual text-to-speech system." Traitement du Signal **9**(4): 359-366.

Bailly, G., G. Gibert, et al. (2002). Evaluation of movement generation systems using the point-light technique. IEEE Workshop on Speech Synthesis, Santa Monica, CA.

Bailly, G. and B. Holm (2005). "SFC: a trainable prosodic model." Speech Communication **46**(3-4): 348-364.

Baras, C. (2005). Tatouage informé de signaux audio numériques. Paris, Ecole Nationale Supérieure des Télécommunications.

Bas, P., J.-M. Chassery, et al. (2002). "Image Watermarking: an evolution to content based approaches." Pattern Recognition **35**(3): 545-561.

Bernstein, L. E., M. E. Demorest, et al. (2000). "Speech perception without hearing." Perception & Psychophysics **62**: 233-252.

Boyes Braem, P. (1999). "Rhythmic temporal patterns in the signing of early and late learners of German Swiss Sign Language." Language and Speech **42**: 177-208.

Cootes, T. F., G. J. Edwards, et al. (2001). "Active Appearance Models." IEEE Transactions on Pattern Analysis and Machine Intelligence **23**(6): 681-685.

Cornett, R. O. (1967). "Cued Speech." American Annals of the Deaf **112**: 3-13.

Cornett, R. O. (1994). "Adapting cued speech to additional languages." Cued Speech Journal **5**: 19-29.

Dornaika, F. and F. Davoine (2005). Simultaneous facial action tracking and expression recognition using a particle filter. IEEE International Conference on Computer Vision, Beijing, China.

Elisei, F., G. Bailly, et al. (2005). Capturing data and realistic 3D models for cued speech analysis and audiovisual synthesis. Auditory-Visual Speech Processing Workshop, Vancouver, Canada.

Gibert, G., G. Bailly, et al. (2005). "Analysis and synthesis of the 3D movements of the head, face and hand of a speaker using cued speech." Journal of Acoustical Society of America **118**(2): 1144-1153.

Gibert, G., G. Bailly, et al. (2006). Evaluating a virtual speech cuer. InterSpeech, Pittsburgh, PE.

Girin, L. (2004). "Joint matrix quantization of face parameters and LPC coefficients for low bit rate audiovisual speech coding." IEEE Transactions on Speech and Audio Processing **12**(3): 265-276.

Leybaert, J. and J. Alegria (2003). The Role of Cued Speech in Language Development of Deaf Children. Oxford Handbook of Deaf Studies, Language, and Education. M. Marschark and P. E. Spencer. Oxford, Oxford University Press**:** 261-274.

LoboGuerrero, A. (2004). Etude de techniques de tatouage audio pour la transmission de données. Grenoble - France, Institut National Polytechnique.

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

Peckels, J. P. and M. Rossi (1973). "Le test de diagnostic par paires minimales. Adaptation au francais du 'Diagnostic Rhyme Test' de W.D. Voiers." Revue d'Acoustique **27**: 245-262.

Summerfield, Q. (1991). Visual perception of phonetic gestures. Modularity and the motor theory of speech perception. I. G. Mattingly and M. Studdert-Kennedy. Hillsdale, NJ, Lawrence Erlbaum Associates**:** 117-138.

Uchanski, R., L. Delhorne, et al. (1994). "Automatic speech recognition to aid the hearing impaired: Prospects for the automatic generation of cued speech." Journal of Rehabilitation Research and Development **31**: 20-41.