TELMA : Telephony for the Hearing-Impaired People. From Models to User Tests

Denis Beautemps[1], Laurent Girin[1], Noureddine Aboutabit[1], Gérard Bailly[1], Laurent Besacier[6], Gaspard Breton[5], Thomas Burger[5], Alice Caplier[3], Marie-Agnès Cathiard[2], Denis Chêne[5], Jeanne Clarke[1], Frédéric Elisei[1], Oxana Govokhina[5], Christian Jutten[3], Viet-Bac Le[6], Martine Marthouret[4], Stéphane Mancini[3], Yves Mathieu[7], Pascal Perret[5], Bertrand Rivet[1], Pablo Sacher[1], Christophe Savariaux[1], Sébastien Schmerber[4], Jean-François Sérignat[6], Mélody Tribout[8], Sylvie Vidal[5]

(1) GIPSA-lab, Département Parole et Cognition, CNRS UMR 5216, (2) Université Stendhal, Grenoble, (3) GIPSA-lab, Département Images et Signaux, (4) CHU de Grenoble, Service ORL, (5) France Télécom, division R&D, (6) Laboratoire d'Informatique de Grenoble, CNRS UMR 5217 (7) GET / Télécoms Paris, CNRS LTCI (8) Polylogies

Contact: Denis.Beautemps@gipsa-lab.inpg.fr

Abstract

Opening the New Technologies of Information and Communication to the disabled people is a question of increasing interest nowadays. The TELMA project aims at developing software and hardware bricks for a telecommunication terminal (cellular phone) for hearing impaired users. This terminal will be augmented with original audiovisual functionalities. More specifically, the TELMA terminal will exploit the visual modality of speech in two main tasks. On the one hand, visual speech information is used to improve speech enhancement techniques in adverse environment (environmental noise reduction enables the hearing-impaired to better exploit his/her residual acoustic abilities). On the other hand, the terminal will provide analysis/synthesis of lip movements and Cued Speech gestures. The Cued Speech is a face-to-face communication method used by a part of the oralist hearing-impaired community. It is based on the association of lip shapes with cues formed by the hand at specific locations. The TELMA terminal will translate lipreading + Cued Spech towards acoustic speech, and vice-versa, so that hearing-impaired people can communicate between them and with normal hearing people through telephone networks. To associate scientific developments, economic perspectives and efficient integration of disabled people concerns, the project is build on a partnership between universities (INPG and ENST), industrial/service company (France Télécom, R&D division) and potential users from the hearing-impaired community, under the supervision of health professionals (Grenoble Hospital Center / ORL).

Categories and subject descriptors

Telecommunications, Cued Speech, speech enhancement.

Key-words

Cued Speech; hearing-impaired; automatic analysis, recognition, modeling and synthesis of speech and gestures; speech enhancement; user tests; Wizard of Oz.

1. Introduction

The adaptation of our environment to the disabled people is a social question of increasing interest, given their difficulties to access many services proposed to the consumer citizens. This evolution has led to introduce some constraints to adapt the services and their associated environments to the needs of the disabled. The field of information technologies follows this rule: The needs of the hearing-impaired people for telecommunication tools are increasing and are more and more taken into account, as the new technologies allow for new possibilities. The UMTS technology could be useful for the hearing-impaired community. New projects concern Internet access interfaces adapted for the hearing-impaired people. Videophones, virtual reality, video analysis and synthesis, software and hardware performances are solutions (among others) for supplying the deficiency of the hearing-impaired people.

Concerning the face-to-face communication, the channels used by the hearing-impaired people can be classified in three categories:

- The hearing-impaired that use lip-reading, as a complement to the voice. People using hearing aid (only one million people in France) or cochlear implant (550 implantations per year) are exploiting both visual and auditory channels. Even if the auditory channel is deficient, these people can perceive some auditory residual information.

- The profoundly deaf people of the oralist category, whose auditory channel is severely damaged, can have their lip-reading ability enhanced by using the Cued Speech method, which is at the heart of this project.

- The profoundly deaf people of the gestual category use the Sign Language, which is very well known, but not considered in this project.

We can also add other modes of communication or helping technologies, possibly implemented in communication networks, like noise reduction, auditory correction, or subtitles. Given this diversity, the ideal service should allow for multi-modality and interoperability. The hearing-impaired user should be free to choose his preferred communication mode, irrespective of the network and the terminal.

The TELMA project aims at elaborating and developing original technological (software and hardware) bricks for a telecommunication (phone) terminal for hearing impaired users. This (prototype) terminal will be augmented with original audiovisual functionalities. This paper presents the stakes of the TELMA project, the current technological mature developments, and the original simulation platform which has been developed for assessing user tests. This platform is also used as a framework for the progressive integration of the TELMA (software and hardware) functionalities.

2. The stakes of the TELMA project

To respond to the multi-modality needs, and to demark ourselves from already existing tools and techniques, the aim of this project is to develop technological modules exploiting the major contribution of the visual modality of speech in the face-to-face communication. The two main functionalities of the TELMA terminal that are addressed are the followings:

- On the one hand, we propose a system for enhancing the speech acoustic signal, degraded by environmental noise, exploiting the visual information related to the speaker's face movements (more specifically speech visible articulators like the lips and the jaw are concerned). We know that those visual speech signals carry significant information which is useful for hearing-impaired people and also for normal-hearing people. Indeed, in adverse environment, the visual speech information can compensate for the decrease of sound intelligibility. Thus, the major theoretical and technical point of this part of the project is: given that the project includes a speaking face video analysis task, how to exploit the particular coherence and complementarity of acoustic and visual speech signals, which are strong properties to fight adverse conditions of communication, so that automatic noise reduction techniques (which are currently purely audio) can be improved.

- On the other hand, we propose a face and Cued Speech analysis-synthesis system used for communication with oralist hearing-impaired speech-readers. The Cued Speech (CS) method was developed by Cornett in 1967 for American English language [1], and then extended to more than 56 languages. It was designed to complement lip-reading, which is intrinsically incomplete (several different sounds can correspond to one single lip-shape). It is based on the association of lip shapes with cues formed by the hand at specific locations around the face (Fig. 1). The cues are formed along two parameters: hand location and hand-shape. The location of the hand (among five possible locations in French) codes a set of vowels whereas the hand-shape (among eight possible configurations) codes a set of consonants. The grouping of the phonemes into sets is made according to their corresponding lip shape contrast (see Fig. 1). For example, the phonemes [p], [b] and [m] have identical visual shapes and are associated to different hand-shapes. In contrast, phonemes that are easily discriminated from lip shapes are grouped in the same hand location/configuration. In this way, the identification of a group of look-alike phonemes at the lips, and the simultaneous identification of a group of phonemes by the hand, results in the identification of a single phoneme (the intersection of two hand/lip sets is always a singleton). Further, the combination of hand-shape, hand location and consecutive lip-shapes, identifies a single consonant-vowel syllable.

The interest in this method was motivated by its effectiveness for the access to complete phonological representations of speech for deaf people, since the first months of life. This involves a positive effect on access to language, and performances for reading and writing get similar to those of hearing people. Finally, considering the current increasing development of hearing implants, this method contributes to facilitate the access toward the auditory modality.

In the TELMA project, we want to implement the above-mentioned audiovisual approach to noise reduction and the Cued Speech analysis-synthesis in a system where these different functionalities are flexible according to environmental conditions and the communication type (hearing impaired / normal hearing people). The different scenarios for a communication using the TELMA terminal are presented in the next section.

It must be noted that the project mainly focuses on the development and implementation of algorithms that will demonstrate the feasibility of the proposed functionalities. The project will also consider the problem of developing software and hardware modules that can allow the implementation of these functionalities in real-time and mobile environment (e.g. cellular terminals). These functionalities will be first implemented on software environment, and then, as far as possible, on hardware targets. The complete specification of the ergonomics and the software and hardware architecture of the terminal providing all the functionalities of TELMA is currently out of the scope of this project, but this project will allow to provide many technological basis for such aim.

It is also important to note that the project will largely consider the user aspects, at all conception levels, by including the participation of the interested hearing-impaired community, under the supervision of deaf rehabilitation specialists (Oto-Rhino-Laryngology service of the Grenoble Hospital). Hearing-impaired subjects will actively participate to the identification of the usability contexts and to the evaluation of the implemented functionalities. The results of such evaluation may drive the future specifications of the terminal ergonomics, since the project aims at providing a terminal usable in real communication conditions.

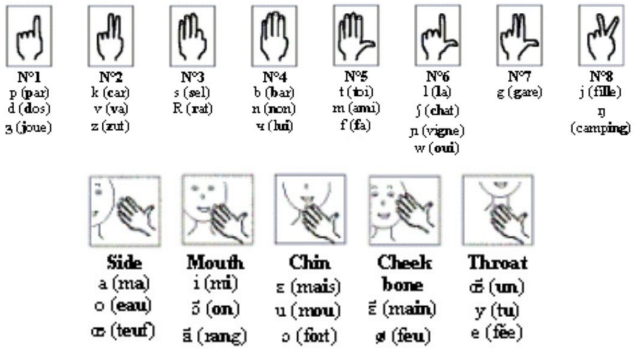

Figure 1: Hand-shapes for coding consonant sets (up) and hand locations for coding vowel sets (down) in French Cued Speech (after [2]).

3. The communication scenarios of TELMA

The different configurations for a communication using the TELMA terminal are the followings. At the source, the acoustic signal produced by a hearing speaker can be transmitted with no signal processing in case of clear environmental conditions. The acoustic signal can also be enhanced by using visual information captured on the speaker's face in case of noisy conditions and/or if a hearing-impaired user with auditory remains requires such signal processing (Fig. 2).

If the hearing impaired speaker uses Cued Speech with no phonation, a video analysis system is used to extract parameters of the face and the hand. A recognition system using these parameters delivers the related phonemic chain, which is then used as input of an acoustic speech synthesiser (Fig. 3).

At the receiver, the acoustic signal can be directly provided to normal hearing users. Alternately, for hearing impaired users, it can be converted into Cued Speech using an acoustic analyser module coupled with a visual synthesizer (Fig. 4). Due to the current limitations of speech recognition systems in unconstrained conditions (environmental noise, multi-speakers, spontaneous speech, etc.), we also keep the possibility (not shown on the figure) of transmitting some labial parameters (resulting from the speaker video analysis). If these parameters are available at the receiver, they can significantly simplify the face and hand synthesis task (lip information is part of Cued Speech and the hand configuration can be estimated in this case from both sound and lip parameters). Lip parameters occupy very low bandwidth and can be transmitted jointly to the sound by using watermarking techniques. Let us remark that, in any case, the visual synthesis should be possibly synchronised with the corresponding (transmitted) acoustic speech signal.

The system is flexible in the sense that it enables any communication scenario between normal-hearing people and Cued Speech users, including a communication between two Cued speech users, through an acoustic speech channel (in classical telephony). Of course, if the system is integrated within a modern videophone network, direct audiovisual communication between two Cued Speech users can be used. But the essence of the TELMA project is not affected, since it is mainly a bi-directional translation system between acoustic speech and Cued Speech. Positively enough, a videophone system may allow to choose to process the translation task at the source level or at the receiver level, or even on a third-party server, depending on the local computational resources. A deepened presentation of the potential material architecture for such a TELMA service is out of the scope of the present paper, which focuses on the acoustic speech / Cued Speech translation (and noise reduction) problem.

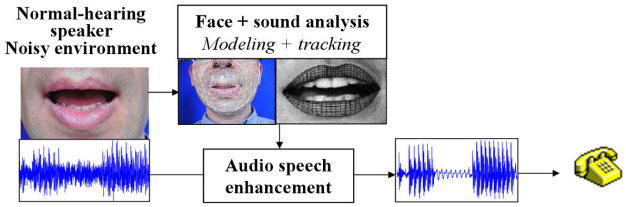

Figure 2: Video-based speech enhancement (before transmission)

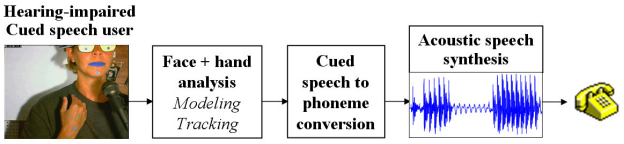

Figure 3: Speech sound generation from Cued Speech analysis (before transmission)

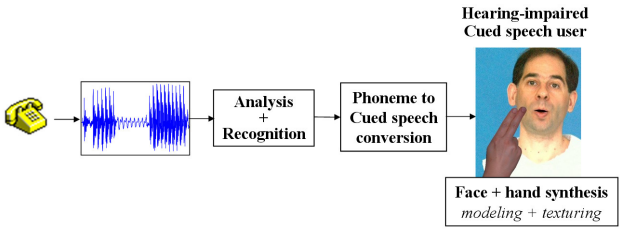

Figure 4: Cued Speech synthesis from acoustic speech (after transmission)

4. Overview of the models and technological modules in TELMA

4.1 Audiovisual speech enhancement

The main objective of this sub-project is to develop a multi-modal approach to the problem of speech enhancement in background noise, by using visual speech information.

4.1.1 Theoretical foundations

This work is based on the signal processing area called *source separation*. The difficult realistic case of a multi-dimensional audio convolutive mixture is addressed: several microphones are used to record the noisy speech audio signal. The different sources (*i.e.* the speech signal of interest and competing interfering sources) are assumed to be mixed at the microphones level by a linear time-invariant process (*i.e.* "natural" linear filters). We aim at using the visual information from one specific speaker (more specifically, we use here the lip movements information) to extract the corresponding speech sound from the mixture (see Fig. 5). Following the basics of source separation techniques, this problem is addressed in the frequency domain where the (multi-dimensional) convolutive mixture is transformed into a (multi-dimensional) linear instantaneous mixture for every frequency bin of the signals (Short-Term) Fourier Transform. Several approaches for the proposed audiovisual processing have been developed.

4.1.2 Exploiting the audiovisual coherence of speech

The first approach has consisted of using the audiovisual coherence of speech to solve the problem of the classical "gain and permutation" indeterminations that are present at the output of purely-audio source separation systems. These indeterminations result from the fact that the separation methods are based on the statistical independence between the signals at the outputs of the system. Therefore, the output signals can only be estimated up to a gain, and up to a permutation between the sources (for instance, an audio system cannot identify on which output channel is the relevant speech signal, and what is its appropriate power level, unless an expertise or an additional information is provided, which is the role of the visual information in our studies). This kind of limitation is even more crucial in "block-by-block" processing on consecutive frames of signal, as is generally done to respect pseudo-real-time constraint.

In a first series of study, demixing filter matrices were estimated for each frequency (up to a gain and a permutation) using a source separation algorithm that has been developed in collaboration between our lab and the Jean Kuntzmann lab in Grenoble [3]. The audiovisual part of the process was to model the audiovisual

coherence of speech signals using a probabilistic audiovisual model, which parameters were trained on a large audiovisual corpus. Using this probabilistic model, the possible permutations and gain errors at the output of the separation system are regularized so that the reconstructed speech signal is as much coherent as possible with the speaker's lip movements (see Fig. 5) [4]. The probabilistic model that has been developed to measure the audiovisual coherence is adapted to both the nature of the data and the separation task [5]. This method has provided good results, but it has several drawbacks: it is computationally heavy and speaker-dependant (the probabilistic model is tuned with the data from one speaker, and it must be adapted or generalized to other speakers for an extended use).

4.1.3 Using a visual voice activity detector

A simpler second approach has been proposed to address those limitations: the use of a visual voice activity detector (V-VAD). In that case, the video information is used to detect the presence or the absence of the relevant speaker in the audio mixture. We have shown in [6] that this V-VAD can be used to solve efficiently the gain/permutation problems at the output of the audio separation system. This is done at very low additional computational cost (given that the basic video processing of TELMA is activated). The processing of the video information for the VAD task is very simple and quite robust: a single dynamic parameter calculated from lip descriptors provides a very good detection. Furthermore, the V-VAD can be adapted very easily to any speaker.

Eventually, a third series of work has been proposed and has been shown to provide very encouraging results. This time, the V-VAD is combined with a completely novel separation method of the geometric kind. This method is based on the projection (for each frequency bin) of the noisy audio data on the complementary space of the sub-space spanned by the mixed signals, the latter being considered when the relevant speech signal to be extracted is detected as absent by the V-VAD [7]. This complementary space, by nature, is assumed to contain the components of the useful speech signal. This method appears very efficient, both in terms of computational cost and separation power: the estimated speech signal is of good quality and clearly emerges from the mixture. Moreover, this method intrinsically circumvents the gain/permutation problems. In the perspective of a mobile TELMA terminal, the future works will concern the development, the evaluation, and the implementation of an online real-time version of this method.

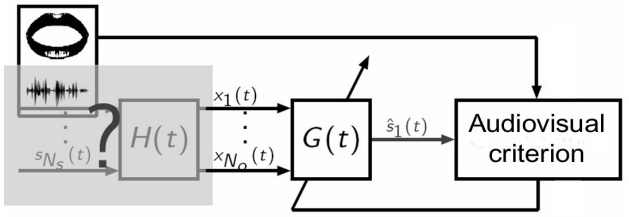

Figure 5: Schema of a speech source separation system assisted by the video. An unknown speech signal $s_1(t)$ is mixed with competing sources through an unknown filter matrix $\boldsymbol{H}(t)$. An audiovisual probabilistic criterion uses the video information to estimate/regularize a demixing filter matrix $\boldsymbol{G}(t)$ to provide the estimated speech signal.

4.2 Transcription of the Cued Speech gestures into phonemes and word sequences

The automatic recognition of the Cued Speech gestures, *i.e.* going from hand and lip movements towards phonemes and words, is the objective of this part. Within the TELMA project, the solution requires (i) the extraction of Cued Speech parameters related to the hand and the lips from the processing of the video sequence (ii) the fusion of the Cued Speech parameters to recognize the corresponding phonetic chain, and (iii) a higher level (language) model for the complete recognition of words and word sequences. These three steps are described in the following sub-sections. Note that, at this point of the project, the complete transcoding of Cued Speech into acoustic speech requires to add a text-to-speech (TTS) module, such as the one developed at GIPSA-lab, or any other system. Since this last step is a classical issue in speech processing, it is not developed in this paper.

4.2.1 Hand segmentation from video image

The global architecture for automatic recognition of LPC gestures is made of the following modules:

- The hand segmentation module which involves a learning step about the colour glove of the coder (see Fig.6).

- The extraction of target images (images related to a given LPC configuration): the proposed algorithm uses a dedicated retina filter which yields to the detection of the images of the video sequence for which the hand and the fingers movements significantly slow down [8];

- The localization of the face, eyes and mouth based on the algorithm of [9];

- The detection of the pointing area: this task is related to the position of the pointing finger with respect to the permanent facial features such as mouth and eyes (see Fig. 6);

- The recognition of the hand configuration: the classification algorithm combines the advantages of the Belief theory, the Support Vectors Machines, and the generalization of the Pignistic transform [10, 11].

Figure 6: left, hand segmentation; right, different pointing areas.

4.2.2 Lip contour extraction

Lip contour segmentation is of crucial importance for lip reading applications. Lot of works has been done for the extraction of the outer lip contour but few studies deal with the problem of inner lip segmentation, especially when dealing with natural lips (*i.e.* no make-up is used to facilitate the extraction). The main reason is that inner contour extraction is much more difficult than outer contour extraction. Indeed, we can find different mouth shapes

and non-linear appearance variations during a conversation. Especially, inside the mouth, there are different areas which have similar color, texture or luminance than lips (gums and tongue). We can also see very bright zones (teeth) as well as very dark zones (oral cavity). Every area can continuously appear and disappear when people are talking.

For the detection of the outer lip contour, the algorithm described in [12] is used. The outer contour being available, some key points are used to initialize an active contour called "jumping snake" for the detection of the inner lip contour. According to some optimal information of luminance and chrominance gradient, this active contour fits the position of two parametric models; a first one composed of two cubic curves and a broken line in case of a closed mouth, and a second one composed of four cubic curves in case of an open mouth. These parametric models give a flexible and accurate final inner lip contour. Fig.7 presents some examples of lip segmentation results.

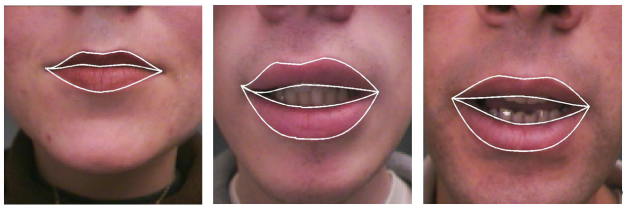

Figure 7: Some examples of lip contour segmentation.

4.2.3 Phonemes labial modeling

The previous processing provides the inner and outer contours of the lips. Geometric lip parameters that are relevant for the phonetic description of the lips are derived from these contours: The lip width *A*, the lip aperture *B*, and the lip area *S* of the inner contour, the lip width *A'*, the lip aperture *B'*, and the lip area *S'* of the outer contour. For the French vowel case, Aboutabit and colleagues have demonstrated that a simple Gaussian classifier based on the lip inner parameters was efficient [14][18]. Indeed, when the Cued Speech hand position is given, high vowel identification performances (89% on the average) are obtained with only one measure instant, the L2 instant of vowel lip target which corresponds to a local minimum of the *S* parameter variation [13]. This global score, based on a set of 1105 vowels, is a recognition score for which the hand position is given, *i.e.* known without error.

4.2.4 Phonemes automatic recognition

The phoneme recognition needs to fusion the lip classification process with the automatic Cued Speech hand decision. Since the detection processes are automatic (see for example Section 4.2.1), they can be corrupted by some errors. The two flows (*i.e.* the labial and the manual information) are by nature complementary in the Cued Speech method. But they are not in direct synchrony as shown in different studies conducted at the GIPSA-lab on Cued Speech production (see for example Attina *et al.* [2], Aboutabit *et al.* [17]). More precisely, the results on the French vowels show that the hand reaches quasi systematically its target (at the M2 instant) largely before the corresponding target at the lips (at the L2 instant), as illustrated in Fig. 8.

In their first attempt to fusion both flows of (lips and hand) information, Aboutabit and colleagues [14][18] have adapted the SI (Separated Identification) model, which is one of the models proposed in the framework of audiovisual speech identification [16], to the Cued Speech case. In this model, the decision on each flow is obtained independently, and the set of resulting phonetic features are combined in a second step to obtain the corresponding phonetic code. In the Cued Speech context, Aboutabit and colleagues considered a Gaussian classifier of the lips (as described in Section 4.2.3) to obtain a candidate vowel among the set of vowels for each of the five hand positions. Let us denote by E1 the resulting set of five candidate vowels. The automatic hand coding gives a hand position associated to a set E2 of two or three vowels, following the specifications of the Cued Speech system (see Fig. 1). The vowel identification is thus the result of the intersection between the E1 and E2 sets. To improve the efficiency of this selection process, Aboutabit *et al.* implemented a cascaded variant of the SI model named "hand first, then the lips", in which the hand decision at the M2 instant is used to select the Gaussian classifier of the lips at the L2 instant from the five possible ones (Fig. 9). The selected classifier directly provides the vowel phoneme. In this latter implementation, the considered L2 instant is the next L2 instant that immediately follows the M2 instant that is considered for the hand decision.

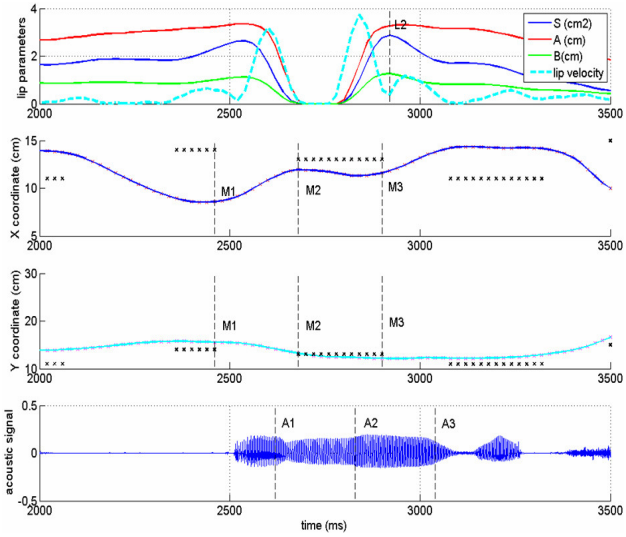

Figure 8: Example of measured signals for a Cued Speech sequence. From top to bottom, (i) the inner lip parameters and derived lip velocity; the lip target instant is denoted L2, (ii) the abscissa X of the gravity centre of the reference landmark on the back of the hand; in dotted lines, the maintained hand positions with the M2 instant of hand target, (iii) the ordinate Y of the gravity centre of the reference landmark on the back of the hand; in dotted lines, the maintained hand positions, (iv) the corresponding acoustic signal with the vowel segmentation instants A2 and A3.

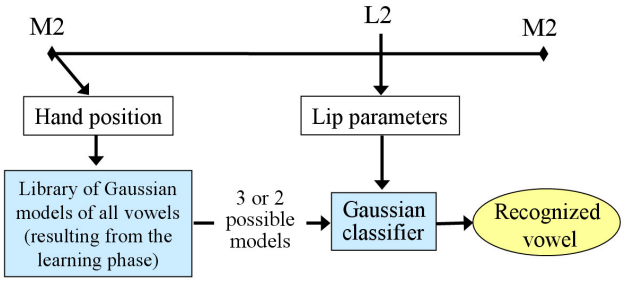

Figure 9: Fusion model for vowel recognition (after [14]).

In their experiment, Aboutabit and colleagues used an automatic Cued Speech hand coding process with an error accuracy of 3.5%. Using the proposed fusion model on 1105 vowels, the performance of vowel recognition was reduced but still reaches an honorable score of 75%. The difference with the above-mentioned 89% score (for which we shall recall that the Cued Speech hand position is given exactly, see Section 4.2.3), is due to the errors in the automatic decision of the hand position and in the matching between M2 and L2 instants.

The extension of the fusion model to the consonants is possible, but it is likely to be more complex. Indeed, the "degree of articulation" of the consonants is quite variable across speakers, utterance, etc. Moreover, the consonant realization at the lips is strongly influenced by the speech context. In their first attempts to include the vowel context, Aboutabit *et al.* obtained encouraging scores (more than 80% of CV identification on a set of 774 Consonant Vowel syllables), taking into account the whole lip movement between the consonant and the vowel with the use of 3-states HMM models [15].

4.2.5 Towards words recognition and beyond

To obtain a complete lexical chain, from hypotheses of phonemes (vowels and consonants) such as provided at the previous stage, resources of higher level are necessary. An analogy can be found with automatic speech recognition in which the acoustical models ("low level"), generating phonetic hypotheses, work jointly with a model of language and a dictionary of pronunciation to decode the most probable lexical chain. In our case, the dictionary of pronunciation represents each word as a sequence of units that can be recognized by "low level" models. These units can be phonemes or syllables according to the choices of "low level" models performed. The advantage of such approach is that a change in the vocabulary of the considered application does not require to modify or to learn again the "low level" models. In close future, we also plan to insert a model of language into the system, allowing to specify the statistics of continuity between words by likelihood of co-occurrences. Such models were efficient for automatic recognition of continuous word, and allow to constraint the research in the space of possible hypotheses.

4.3 Video synthesis of Cued Speech from acoustic speech

In this section, we describe the dual task of transcoding acoustic speech into Cued Speech. Similarly to the previous section, the automatic speech recognition system that is used at the beginning of the chain to deliver the phonetic chain is not presented here: At this point of the project, it can be any of the many existing systems. Our current developments are based on the system developed at the LIG. We focus next on the Cued Speech synthesizer, which is based on corpus-based modeling and concatenative synthesis. The first results of an evaluation campaign are also reported.

4.3.1 Corpus-based parametric models for a virtual clone

The system that is currently used for the synthesis of Cued Speech from acoustic speech is a video-realistic clone, designed from measures made on a reference human Cued Speech coder. This clone includes 3D shape and appearance models of the face and of the hand. These models have been developed within the framework of the RNRT project ARTUS [19]. The clone is controlled by 9 parameters for the hand (each parameter corresponds to an elementary gesture), 7 parameters for the face, and also 12 parameters to control the arm and the head. These parameters are derived from the analysis of the (hand/arm/face/head) movements of the human Cued Speech coder for 239 different sentences. This set of sentences contains all the possible French diphones, so that concatenative synthesis can be achieved (see below). The recording was made using a VICON® motion capture system with 12 cameras operating at 120Hz, the sound being recorded in synchrony [19].

4.3.2 Synthesis based on pre-stored units

Speech and hand/face gestures synthesis is achieved by selection, concatenation, and smoothing of the pre-stored units resulting from the analysis. Two kinds of segments are considered: On the one hand, the "polyphones" represent the signal and facial gestures from one acoustic target of a given allophone to the next (note that some sounds –e.g. the glides– do not have a clear target and are embedded in a larger segment); On the other hand, the "dikeys" represent the hand gestures, the movements of the arm, and the movements of the head, from one key target to the next. For synthesis, the di-keys boundaries are first synchronized with the polyphones boundaries, following a set of speech/gestures synchronization rules [20].

The dictionary of pre-stored units is redundant, *i.e.* several units of the dictionary represent the same phonological segment in different utterances. The optimal unit from one multi-represented set is selected by using Dynamic Programming techniques with specific selection and concatenation costs. The generation of the speech sound is then achieved using the classical TD-PSOLA technique. The generation of the hand/face gestures is achieved by time-stretching/compression of the pre-stored units, the transient parts being preserved as much as possible. A predictive filtering is applied to assume a smooth continuity of the movements.

4.3.3 Intelligibility tests

A first evaluation of the Cued Speech synthesizer intelligibility has been achieved using a segmental intelligibility test derived from the Diagnostic Rhyme Test (the DRT is based on the identification of pairs of sounds, generally monosyllable words, only differentiated by one phoneme, e.g. veal/feel; This test is commonly used for testing the acoustic intelligibility). In the present study, the phonetic contrasts are enhanced by the Cued Speech gestures. Minimal pairs of words of the form C1VC2, with a meaning in French (e.g. *bal* vs. *mal*, *sire* vs. *tir*) have been chosen so that a choice based on lip-reading only is difficult. Eight hearing-impaired persons participated to the test. For all of

them, the gain of intelligibility provided by the synthetic Cued Speech information was drastic: The choice between pairs using facial animation only is equivalent to random, whereas the discrimination score is close to 94% when the virtual clone produces Cued Speech (with high statistical significance for all subjects; $F(1,3134) = 7.5$, $p<0.01$) [21].

4.3.4 TDA system for improving face movements synthesis

The TDA (Task Dynamics for Animation) system has been proposed to improve the quality of the face movements synthesis, emphasizing on the fluidity. The model involves two steps: (a) A gesture planning step, which specifies the trajectories of the basic geometric parameters (opening; width, protrusion for the lips, distance from hand to face for the location, etc.) using statistical modeling of trajectories by Hidden Markov Models (HMM) [22], and (b) A gesture execution step, which generates the articulatory trajectories using pre-stored units (see Section 4.3.2). During this step, the distance between the pre-stored and the planned articulatory trajectories is integrated in the selection cost. This way, the TDA synthesis takes benefit from the advantages of both systems [23]: the HMM-based planning of face movements ensures a global coherence of the gestures, while the concatenative synthesis ensures to preserve the subtle articulatory details and the cross-parameters synchrony, that are both rooted in corpus-based information. This was confirmed by the objective evaluations that have been conducted on this system.

Perspectives to this work include an HMM-based model for trajectory forming that would integrate a multi-modal phasing model: i.e., a model that would be able to jointly learn the co-articulation between the gestures (hand, face, sound) and their relative (de)phasing [24].

5. Study of the user aspects

A multimodal communication platform that enables to simulate the functionalities of the TELMA terminal has been conceived. It can be used to test the interaction between a hearing impaired and a normal-hearing person, in real-life user situation. The "Wizard of Oz paradigm" has been used to ensure an optimal implication of the subjects in the course of the communication scenarios. The audiovisual data that have been recorded can be used as references for the adaptation and the optimization of the different models that compose the software bricks of TELMA.

5.1 The Wizard of Oz paradigm adapted to TELMA

The "Wizard of Oz" paradigm is an experimentation conducted in the field of human-machine interaction, in which the subjects are interacting with a system that they believe automatic, but which is in fact totally or partially controlled by a human. In the TELMA framework, a hearing impaired subject thinks that he/she is interacting with a normal hearing person by means of a technological interface that automatically converts speech sounds to Cued Speech and Cued Speech to speech sounds. In fact, the speech signal produced by the interlocutor is performed in Cued Speech by a Cued Speech human coder (hence the Wizard) and displayed on the screen of the hearing-impaired tester (the Wizard is hidden in a different room) (Fig. 10); Alternately, the Wizard can type on a computer keyboard the text corresponding to the acoustic speech produced by the normal-hearing interlocutor, and the corresponding Cued Speech sequence is delivered by the synthesizer (hence the Wizard is replacing the automatic recognition brick) (Fig. 11).

Figure 10: Wizard of Oz Experiment 1: The acoustic speech produced by the normal-hearing user is coded in Cued Speech by the (hidden) Wizard and presented to the hearing-impaired subject using a video screen.

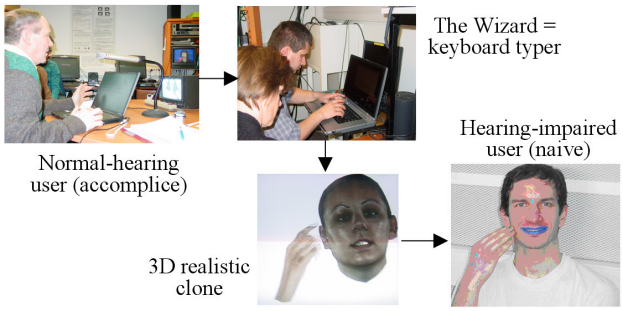

Figure 11: Wizard of Oz Experiment 2: The acoustic speech produced by the normal-hearing user is translated into text by the Wizard and then synthesized from the text by the Cued Speech synthesizer for presentation to the hearing-impaired user.

5.2 Results in terms of usability

The platform and the above experiments have allowed to validate several elements: (i) the communication loop between hearing-impaired and normal-hearing users, (ii) the different Cued speech functionalities of TELMA, (iii) the acceptation by the users of the half-duplex (i.e. alternate turns) communication mode, (iv) the related acceptation by the users of the latency (due to the half-duplex mode and to the translation/calculation time), (v) the target application of the interaction: in the above experiments, it consisted of the simulation of medical services (taking an appointment with a medical cabinet, and contacting an emergency care unit). The current platform is used as a reference for the progressive integration of the software/hardware bricks that will assume the automation of the TELMA functionalities in the future. For instance, a first version of the virtual clone has been evaluated on this basis (Fig. 11). Five tested subjects out of eight could fulfill the scenarios tasks completely. This shows that the virtual clone is effective in providing the useful information in Cued Speech, even if the users have identified several points that can be improved in future versions [25].

6. Acknowledgements

The TELMA project is supported by the ANR (French National Research Agency), as a RNTS (National Network of Health Technologies) project. We thank the eight hearing-impaired subjects that participated to the user tests.

7. References

[1] Cornett, R.O. Cued Speech, *American Annals of the Deaf*, Vol. 112, pp. 3-13, 1967.

[2] Attina, V., Beautemps, D., Cathiard, M. A. & Odisio, M. A pilot study of temporal organization in Cued Speech production of French syllables: rules for Cued Speech synthesizer, *Speech Communication*, Vol. 44, pp. 197-214, 2004.

[3] Pham, D.-T., Servière, C. & Boumaraf, H. "Blind separation of convolutive audio mixtures using nonstationary," *Proc. Int. Conf. Independent Components Analysis & Blind Source Separation (ICA)*, Nara, Japan, pp. 981–986, 2003.

[4] Rivet, B., Girin, L. & Jutten, C., Mixing audiovisual speech processing and blind source separation for the extraction of speech signals from convolutive mixtures, *IEEE Transactions on Audio, Speech, & Language Processing*, Vol. 15, No. 1, pp. 96-108, January 2007.

[5] Rivet, B., Girin, L. & Jutten, C., Log-Rayleigh distribution and its application to log-spectral coefficients statistical modeling, *IEEE Transactions on Audio, Speech, & Language Processing*, Vol. 15, No. 3, pp. 796-802, March 2007.

[6] Rivet, B., Girin, L., Servière, C., Pham, D.-T. & Jutten, C. Using a visual voice activity detector to regularize the permutations in blind separation of convolutive speech mixtures, *Proc. IEEE Int. Conf. on Digital Signal Processing (DSP 2007)*, Cardiff, Wales, 2007.

[7] Rivet, B., Girin, L. & Jutten, C. Visual voice activity detection as a help for speech source separation from convolutive mixtures, *Speech Communication*, Vol. 49, No. 7/8, pp. 667-677, July 2007.

[8] Burger, T., Benoit, A., & Caplier, A. Intercepting Static Hand Gestures in Dynamic Context, *Proc. Int. Conf. Image Processing*, Atlanta, USA, 2006.

[9] Garcia, C. & Delakis, M. Convolutional Face Finder: A Neural Architecture for Fast and Robust Face Detection, *IEEE Transactions on Pattern Analysis & Machine Intelligence*, Vol. 26, No. 11, pp. 1408-1423, 2004.

[10] Burger T., Aran, O. & Caplier, A., Modeling hesitation and conflict: a belief-based approach for multi-class problems, *Proc. Int. Conf. on Machine Learning & Applications (ICMLA 2006)*, Orlando, USA, 2006.

[11] Aran, O., Burger, T., Caplier, A. & Akarun, L. Sequential Belief-Based Fusion of Manual and Non-Manual Signs, *7th Int. Workshop on Gesture in Human-Computer Interaction & Simulation*, Lisbon, Portugal, 2007.

[12] Eveno, N., Caplier, A. & Coulon, P-Y. Jumping Snakes and Parametric Model for Lip Segmentation, *Proc. Int. Conf. on Image Processing*, Barcelona, Spain, 2003.

[13] Aboutabit, N., Beautemps, D. & Besacier, L. Characterization of Cued Speech Vowels from lip inner contour, *Proc. Int. Conf. on Spoken Language Processing*, Pittsburgh, Pennsylvania, USA, 2006.

[14] Aboutabit, N., Beautemps, D. & Besacier, L. Vowel classification from lips: the Cued Speech production case, *Proc. Int. Seminar on Speech Production*, Ubatuba, Brazil, 2006.

[15] Aboutabit, N., Beautemps, D. & Besacier, L. A HMM modeling for CV syllable recognition, *Proc. Int. Conf. on Spoken Language Processing*, Antwerp, Belgium, 2007.

[16] Schwartz, J. L., Robert-Ribes, J. & Escudier, P. Ten years after Summerfield: a taxonomy of models for audio-visual fusion in speech perception. *In Hearing by Eye II, Advances in the psychology of speechreading and auditory visual speech,* Psychology Press, pp. 85-108, Hove, UK, 1998.

[17] Aboutabit, N., Beautemps, D. & Besacier, L. Hand and lips desynchronization analysis in French Cued Speech: automatic temporal segmentation of visual hand flow, *Proc. Int. Conf. on Acoustics, Speech, & Signal Processing*, Toulouse, France, 2006.

[18] Aboutabit, N., Beautemps, D. & Besacier, L. Automatic identification of vowels in the Cued Speech context, *Proc. Int. Conf. on Audiovisual Speech Processing* (AVSP), Hilvarenbeck, Netherlands, 2007.

[19] Bailly, G., Attina, V., Baras, C., Bas, P., Baudry, S., Beautemps, D., Brun, R., Chassery, J.-M., Davoine, F., Elisei, F., Gibert, G., Girin, L., Grison, D., Léoni, J.-P., Liénard, J., Moreau, N., & Nguyen, P. ARTUS: calcul et tatouage audiovisuel des mouvements d'un personnage animé virtuel pour l'accessibilité d'émissions télévisuelles aux téléspectateurs sourds comprenant la Langue Française Parlée Complétée, *Actes de Handicap 2006*.

[20] Gibert, G., Bailly, G., Beautemps, D., Elisei, F. & Brun, R. Analysis and synthesis of the 3D movements of the head, face and hands of a speech cuer, *Journal of the Acoustical Society of America*, Vol. 118, No. 2, pp. 1144-1153, 2005.

[21] Gibert, G., Bailly, G. & Elisei, F. Evaluating a virtual speech cuer, *Proc. Int. Conf. on Spoken Language Processing*, Pittsburgh, Pennsylvania, USA, 2006.

[22] Zen, H., Tokuda, K. & Kitamura, T. An introduction of trajectory model into HMM-based speech synthesis, *Proc. ISCA Speech Synthesis Workshop*, Pittsburgh, Pennsylvania, USA, 2006.

[23] Govokhina, O., Bailly, G., Breton, G. & Bagshaw, P. TDA: A new trainable trajectory formation system for facial animation. *Proc. Int. Conf. on Spoken Language Processing*, Pittsburgh, Pennsylvania, USA, 2006.

[24] Govokhina, O., Bailly, G. & Breton, G. Learning optimal audiovisual phasing for a HMM-based control model for facial animation. *Submitted at the ISCA Speech Synthesis Workshop*, Bonn, Germany.

[25] Chêne, D., Vidal, S., Tribout, M. & Beautemps, D. Etude de l'interaction distante en mode LPC, *Soumis à la Conférence ASSISTH'2007*.