Evaluating a virtual speech cuer

G. Gibert, G. Bailly & F. Elisei

Institut de la Communication Parlée, UMR CNRS n°5009, INPG/Univ. Stendhal
46, av. Félix Viallet, 38031 Grenoble Cedex, France
{gibert,bailly,elisei}@icp.inpg.fr

Abstract

This paper presents the virtual speech cuer built in the context of the ARTUS project aiming at watermarking hand and face gestures of a virtual animated agent in a broadcasted audiovisual sequence. For deaf televiewers that master cued speech, the animated agent can be then superimposed - on demand and at the reception - on the original broadcast as an alternative to subtitling. The paper presents the multimodal text-to-speech synthesis system and the first evaluation performed by deaf users.

Index Terms: Cued speech, evaluation, audiovisual speech synthesis

1. Introduction

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 [16] or in words or sentences [5]. Cued Speech (CS) was designed to complement speechreading. Developed by Cornett [7, 9] and adapted to more than 50 languages [8], 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). The French CS (FCS) system is described in Figure 2 and Figure 3. Note that the basic CS coding unit is the CV sequence. Isolated vowels and consonants (resp. not preceded by a consonant or not followed by a vowel) are respectively coded with a default hand position or hand shape (indicated by stars in Figure 2 and Figure 3). Numerous studies have demonstrated the drastic increase of intelligibility provided by CS compared to speechreading alone [15, 19] and the effective facilitation of language learning using FCS [13, 14].

A large amount of work has been devoted to CS perception but few works have been devoted to CS synthesis [see rule-based systems described in 1, 10]. We describe here a multimodal text-to-speech system driving a virtual FCS speaker and its first evaluation by deaf users.

2. The multimodal text-to-speech system

The multimodal text-to-speech system developed in the framework of the ARTUS project [3] converts a series of subtitles into an acoustic signal and a stream of animation parameters for the head, face, arm and hand of a virtual cuer. The control, shape and appearance models of the virtual cuer have been determined using multiple multimodal recordings of one human speaker using cued speech.

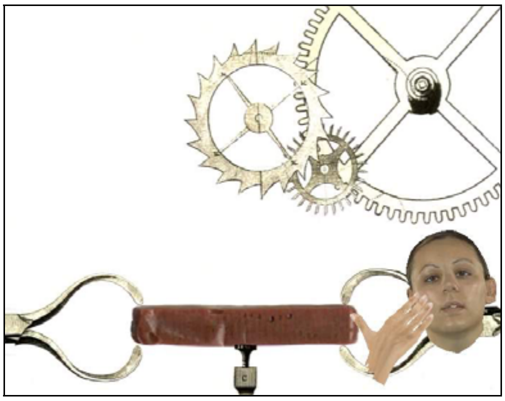

Figure 1: Superimposition of the ARTUS virtual speaker in a broadcast produced by ARTE.

2.1. Data collection and modelling

The different experimental settings used to record our target cuer and capture its gestures are described in a previous paper [12]. These settings include (a) intensive motion capture with good time resolution (120Hz) and high accuracy (0.1mm) of 113 infrared (IR) reflecting markers glued on the face and hand of the subject (see Figure 4), when cueing 238 sentences; (b) video capture of 247 coloured beads glued on the face of the subject (see Figure 5), when cueing simple syllables; and (c) scans of her head, moulds of her hand and her teeth. Using this data, accurate shape and appearance models of the head and face of the subjects have been developed [11]: errors are close to the millimetre for both models. Both shape and appearance models of the face and hand are driven by quasi-articulatory parameters emerging from statistical analysis of the geometric degrees-of-freedom of the observed shapes.

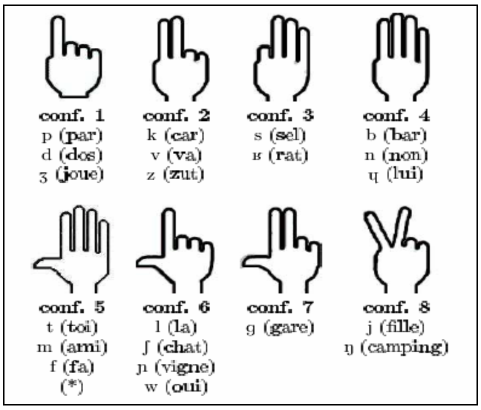

Figure 2: Coding subsets of consonants with hand shapes

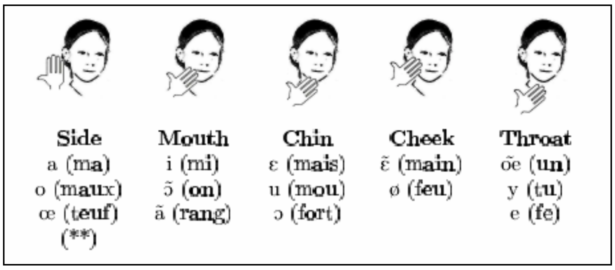

Figure 3: Coding subsets of vowels with hand positions.

2.2. Multimodal text-to-speech synthesis

The COMPOST text-to-speech synthesis system [2] has been parameterized and specific modules have been added to deal with FCS generation. These add-ons include the following items:

Linguistic processing and mark-up. The subtitles have little punctuation. A specific module considers the beginning of each text fragment as a potential sentence beginning and discards unlikely hypotheses. Synchronization marks equal to the time codes of retained start fragments are then inserted. The rhythmic model will then adapt inter-sentences pause durations to fulfil when possible these meeting points.

Prosody. Even though cuers are able to minimize the impact of the adding of the gesture modality on speech rate, the coding of isolated consonants (in the onset of a consonantal cluster or in the coda) using side hand position and the inertia of the arm impose a slower speech rate and an hyper-articulated pronunciation of complex syllables. Intonation is also affected. The trainable prosodic model SFC [4] was thus trained using data from experiment (a). For the three broadcasts tested so far, only four sentences were not pronounced in the time interval devoted to their display, with an average delay of 120ms. Note that rules governing the positioning of time stamps of subtitles usually consider the number of letters to display and not the time to read them.

Multimodal concatenative synthesis. Synthetic gestures and sound are produced by selecting, smoothing and concatenating multi-represented multimodal segments. Because of the specific coordination between face and hand gestures during Cued Speech production, the system proceeds in two steps: two types of segments are considered and synchronized by phasing gestural and acoustic landmarks according to specific rules [12]: “polysounds” that capture the signal and facial gestures between two stable acoustic targets (sounds such as glides are thus enclosed in larger units) and “dikeys” that encompass the arm, hand and head gestures between two successive hand position targets. Note that head movements of our cuer contribute significantly to the hand/face constrictions: if the hand carries out most of the path towards the hand position on the face, the head itself accomplishes on average 16.43% of that distance. Such an enhanced contribution of posture to discourse structure has also been reported for native signers [6]. Multi-represented segments are selected by a classical dynamic programming using specific selection and concatenation costs. The concatenation costs take into account the relative contribution of each quasi-articulatory parameter to the variance of the facial and hand shapes.

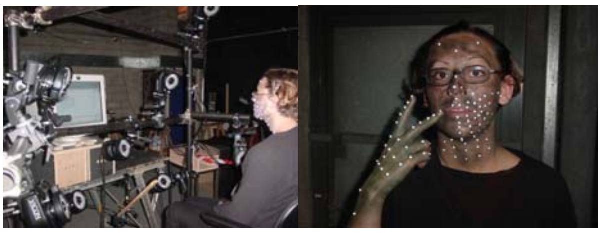

Figure 4: Motion capture using a VICON® system with 12 cameras, 50 beads glued on the hand and 63 on the face.

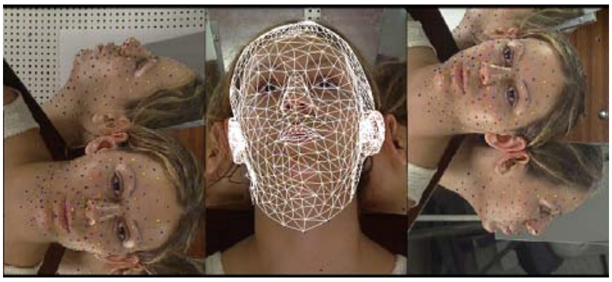

Figure 5: Video recordings of the face with 247 beads.

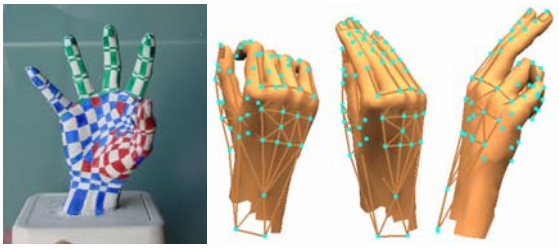

Figure 6: Left: meshing the cast of the hand. Right: controlling the skinned model with the shape model built from motion capture.

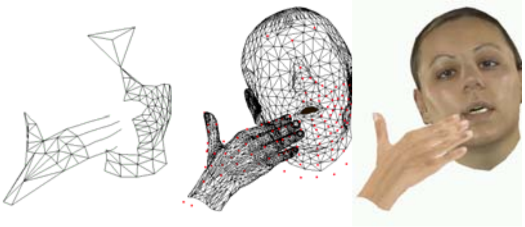

Figure 7: The virtual cuer: from motion capture to video-realistic animation.

2.3. Video-realistic animation

A video-realistic model of the hand of our cuer has been developed using casts and skinning procedures (see Figure 6). Generic models of the lips, skull, teeth and eyes have been adapted to the subject’s morphology. The resulting high-definition shape model is properly textured using blending of multiple cylindrical textures controlled by the quasi-articulatory parameters. The resulting videorealistic speech cuer is presented in Figure 7. It is controlled by the hand and face models developed using motion capture data (cf. §2.1) and thus combines a high and accurate surface resolution (more 10000 vertices for articulated hand and face meshes) with nice time resolution (motion capture data at 120Hz).

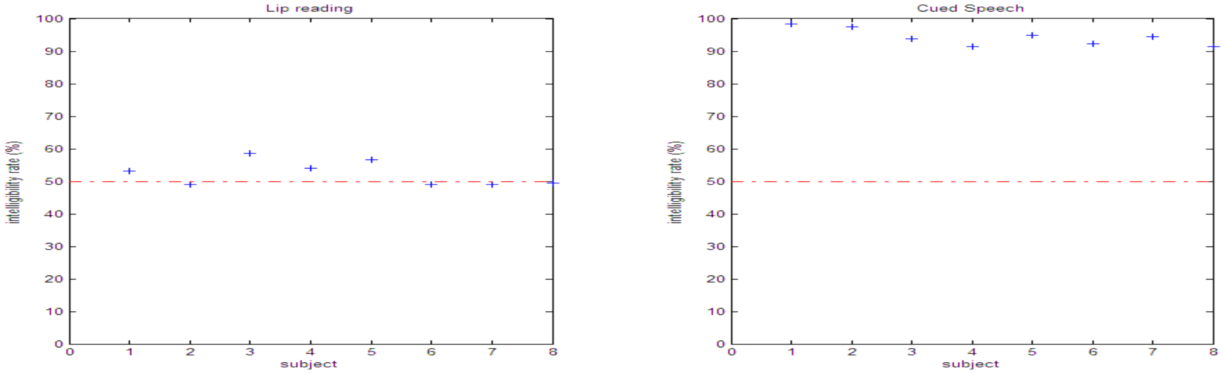

Figure 8: Mean intelligibility rate for each subject for "lipreading" and "Cued Speech" conditions.Whereas intelligibility rate is not different from haphazard way (red dot line) for the "lipreading" condition, it is up to 94% for the "Cued Speech" condition.

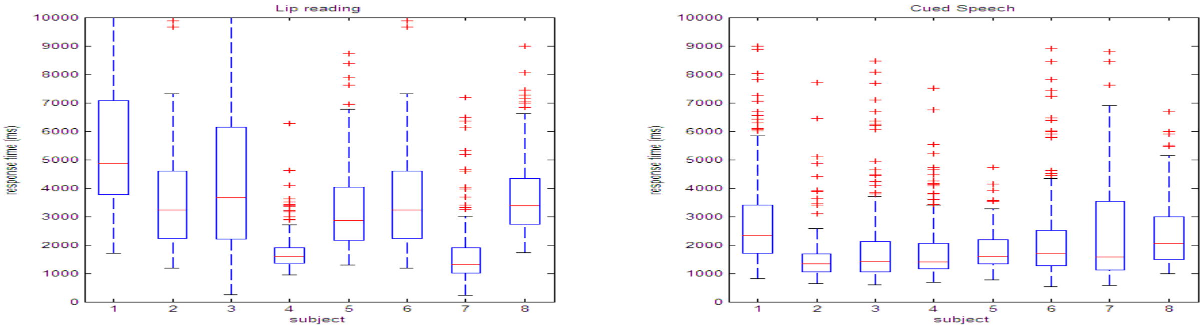

Figure 9: Response time for each subject for "lipreading" and "Cued Speech" conditions. Even there is variability between subjects, the "lipreading" response time is doubled compared to the "Cued Speech" response time.

3. Evaluation

A first series of experiments are been conducted to evaluate the intelligibility of this virtual cuer with skilled deaf users of the French cued speech. This first evaluation campaign is dedicated to segmental intelligibility and the second one to comprehension of content.

3.1. Segmental intelligibility

This test was conducted to assess the contribution of the cueing gestures in comparison with lip reading alone.

Minimal pairs. The test mirrors the Modified Diagnostic Rime Test developed for French by Peckels and Rossi [17]: the minimal pairs do not here test acoustic phonetic features but gestural ones. A list of valid French CVC word pairs has thus been developed that test systematically pairs of consonants in initial positions that differ almost only in hand shapes (ex: [bal] vs. [mal], [siR] vs. [tiR]): we choose the consonants in all pairs of 8 subsets of consonants (see Figure 2) that are highly visually confusable [18]. The vocalic substrate was chosen so as to cover all potential hand positions while the final consonant was chosen so that to avoid rarely used French words or proper names, and test our ability to handle coarticulation effects. Due to the fact that minimal pairs cannot be found in valid French CVC words, we end up with a list of 196 word pairs.

Conditions. Minimal pairs are presented randomly and in both order. The lipreading-only condition is tested first. The cued speech condition is then presented in order to be able to summon up cognitive resources for the most difficult task first (see the lower reaction times for CS stimuli in Figure 9).

Stimuli. In order to avoid a completely still head, head movements of the lipreading-only condition are those produced by the text-to-cued speech synthesizer divided by a factor of 10. No attempt is made to modify segmental or suprasegmental settings that enhance articulation. Although some subjects have cochlear implants that enhance their speech decoding capacity, sound generation was turned off.

Subjects. Eight subjects were tested. They are all hearing impaired people who have practised French Cued Speech since 3 years old.

Results. Mean intelligibility rate for "lipreading" condition is 52.36% (see Figure 8). It is not different from haphazard way of response that means minimal pairs are not distinguishable. Mean intelligibility rate for "Cued Speech" 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 "Cued Speech" task is easier: the response time is significantly different (one factor repeated measure ANOVA (F(1,3134)=7.5, p<0.01)) and lower than for the "lipreading" one (see Figure 9).

3.2. 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 superimposition of the virtual cuer (see Figure 1). Ten questions were asked. The results show all the information is not perceived. On average, the subjects correctly replied to 3 questions. The difficulties of the task (proper names, high speaking flow …) could explain these results.

We conducted further experiments using a Tobii® eye tracker system in order to understand previous results. We asked 4 other deaf people to watch a video consisted of 2 parts: on the first part, the video of a real cuer and on the second part the video of our virtual cuer. There is no significant difference on the time spent on the mouth area for both modalities F(1,6)=0.22, p=0.65. Then, we asked the same 4 deaf people to watch the TV program that was half subtitled and half commented by the virtual cuer (see Figure 10). The results show deaf people spend 56.36% of the time on the teletext and 80.70% on the overlay area for the cuer with a significant difference F(1,6)=9.06, p<0.05. A control group consisting of 16 normal-hearing subjects watches the same audiovisual sequence that was entirely subtitled. They spend 40.14% of the time reading teletext. There is no significant difference with the deaf subjects.

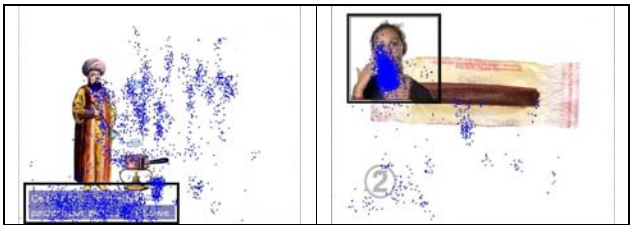

Figure 10: Eye gaze captured during the comprehension test using an eye tracker system for the subtitled (left) and video superimposed (right) broadcast audiovisual sequence.

4. Conclusions

The recordings and analysis of the performance of a real FCS speaker allow us to implement a complete text-to-cued speech synthesizer. The results of the preliminaries perceptive tests show that significant linguistic information with minimal cognitive effort is transmitted by our system. This series of experiments must be continued on more subjects. More evaluation and modelling work is required to quantify and reduce the cognitive effort devoted to CS decoding.

5. Acknowledgements

The authors thank Martine Marthouret, Marie-Agnès Cathiard, Denis Beautemps and Virginie Attina for their help and comments on building the perceptive tests. We also want to thank the 28 subjects who took part to the evaluation.

6. References

[1] 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.

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

[3] Bailly, G., 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., and Nguyen, P. (2006) *ARTUS : synthèse 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*. in *Handicap*. Paris. p.265-270.

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

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

[6] Brentari, D. (1999) *A prosodic model of sign language phonology*. Boston, MA: MIT Press.

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

[8] 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.

[9] Cornett, R.O. (1982) *Le Cued Speech*, in *Aides manuelles à la lecture labiale et perspectives d'aides automatiques*, F. Destombes, Editor. Centre scientifique IBM-France: Paris.

[10] 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.

[11] Elisei, F., Bailly, G., Gibert, G., and Brun, R. (2005) *Capturing data and realistic 3D models for cued speech analysis and audiovisual synthesis*. in *Auditory-Visual Speech Processing Workshop*. Vancouver, Canada

[12] Gibert, G., Bailly, G., Beautemps, D., Elisei, F., and Brun, R. (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): p.1144-1153.

[13] Leybaert, J. (2000) *Phonology acquired through the eyes and spelling in deaf children.* Journal of Experimental Child Psychology, **75**: p.291-318.

[14] Leybaert, J. (2003) *The role of Cued Speech in language processing by deaf children: an overview.* in *Auditory-Visual Speech Processing.* St Jorioz - France. p.179-186.

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

[16] 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.

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

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

[19] 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.