Evaluation of a virtual speech cuer

Guillaume Gibert, Gérard Bailly, Frédéric Elisei
Institut de la Communication Parlée, UMR CNRS n°5009, INPG/Univ. Stendhal, 46 av. Félix Viallet 38031 Grenoble Cedex, France

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 - in 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.

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* such as [u] vs. [y] for French). Indeed, even the best speechreaders do not identify more than 50 percent of phonemes in nonsense syllables (Owens and Blazek 1985) or in words or sentences (Bernstein, Demorest et al. 2000). Cued Speech (CS) was designed to complement speechreading. Developed by Cornett (1967; 1982) and adapted to more than 50 languages (Cornett 1988), 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 with a hand shape). Numerous studies have demonstrated the drastic increase of intelligibility provided by CS compared to speechreading alone (Nicholls and Ling 1982) and the effective facilitation of language learning using CS (Leybaert 2000; Leybaert 2003). A large amount of work has been devoted to CS perception but few works have been devoted to CS synthesis. We describe here a multimodal text-to-speech system driving a virtual CS speaker and its first evaluation by deaf users.

The multimodal text-to-speech system

The multimodal text-to-speech system developed in the framework of the ARTUS project converts a series of subtitles into a stream of animation parameters for the head, face, arm and hand of a virtual cuer and an acoustic signal. The control, shape and appearance models of the virtual cuer (see

Proceedings of ISCA Tutorial and Research Workshop on Experimental Linguistics, 28-30 August 2006, Athens, Greece.

Figure 1) have been determined using multiple multimodal recordings of one human cuer. The different experimental settings used to record our target cuer and capture its gestures and the complete text-to-Cued speech system are described in our JASA paper (Gibert, Bailly et al. 2005).

-	[bɔ̃]	[ʒu]	[r]	-

Figure 1: Chronogram for the word “Bonjour” pronounced and cued by our virtual speaker.

Evaluation

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. The first evaluation campaign was dedicated to segmental intelligibility and the second one to long-term comprehension.

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 (Peckels and Rossi 1973): the minimal pairs do not here test acoustic phonetic features but gestural ones. A list of CVC word pairs 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). 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 all vocalic substrates, 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.

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 (in fact the head accomplished on

average 16.43% of the head/hand contact distance). We did not modify segmental nor suprasegmental settings that could enhance articulation.

Subjects Height subjects were tested. They are all hearing impaired people who have practised French CS (FCS) since the age of 3 years.

Results Mean intelligibility rate for "lipreading" condition is 52.36%. It is not different from haphazard way of response that means minimal pairs are not distinguishable. 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 different $F(1,3134)=7.5$, $p<0.01$ and lower than for the "lipreading" one.

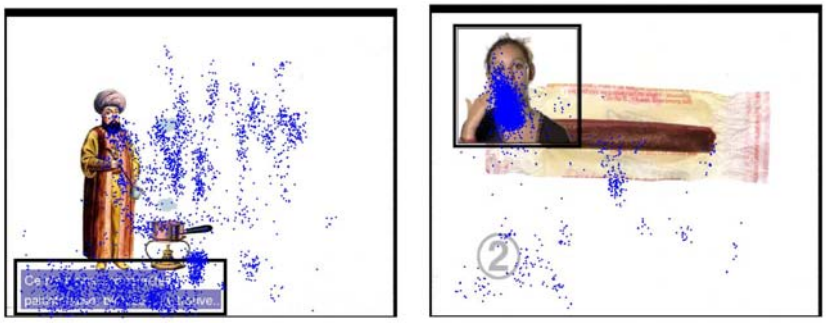

Figure 2: Eye gaze for one subject captured during the comprehension test using an eye tracker system: (left) teletext, (right) video superimposed.

Comprehension

To evaluate the global comprehension of our system, we asked the same 4 subjects to watch a TV program where subtitles were replaced by the incrustation of the virtual cuer. Ten questions were asked. The results show all the information is not perceived. On average, the subjects replied correctly to 3 questions. The difficulties of the task (proper names, high speaking rate) could 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.

Conclusions

The observation and recordings of CS in action allow us to implement a complete text-to-Cued Speech synthesizer. 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 used. Discourse segmentation and part of speech emphasis by prosodic cues (not yet implemented) is expected to enlighten this effort.

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 25 subjects who took part to the evaluation.

References

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

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

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

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): 375-384.

Gibert, G., G. Bailly, D. Beautemps, F. Elisei and R. Brun (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.

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

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

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

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

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.