Toward a speaker-independent visual articulatory feedback system

Atef Ben Youssef, Thomas Hueber, Pierre Badin, Gérard Bailly, Frédéric Elisei
GIPSA-lab UMR 5216/CNRS/INP/UJF/U.Stendhal, Grenoble, France

Context

Several studies tend to show that visual articulatory feedback is useful for phonetic correction, both for speech therapy and "Computer Aided Pronunciation Training" (CAPT) [1]. In [2], we proposed a visual articulatory feedback system based on a 3D talking head used in "an augmented speech scenario", *i.e.* displaying all speech articulators including the tongue and velum. In the proposed system, the clone is automatically animated from the audio speech signal, using acoustic-to-articulatory inversion techniques based on statistical approaches. However, this system was speaker-dependant and thus could not be used in realistic situations. We report here on latest developments of this system and present a first approach that allows its potential use by any speaker.

Talking head

The talking head developed at GIPSA-lab consists of individual three-dimensional models of various speech organs of the same speaker, built from MRI, CT and video data. The "jaw/lips/face model" is controlled by 5 parameters (*jaw height*, *jaw advance*, *lip protrusion, upper and lower lip heights*). The "jaw / tongue model" is controlled also by 5 parameters (*jaw height*, *tongue body* and *tongue dorsum* which respectively control the front-back and flattening-arching movements of the tongue, *tongue tip vertical / horizontal* which control the shape of the tongue tip). No velum model is used in this study. The developed talking head with different types of displays is presented in figure 1. As shown in [3], this 3D clone can be efficiently animated from a 2D EMA data stream (for the same speaker): the information provided by the location of the EMA coils is sufficient to inverse the articulatory models of the talking head, *i.e.* to estimate the control parameters that provide the best fit between the modeled 3D surfaces and the coils' measured coordinates.

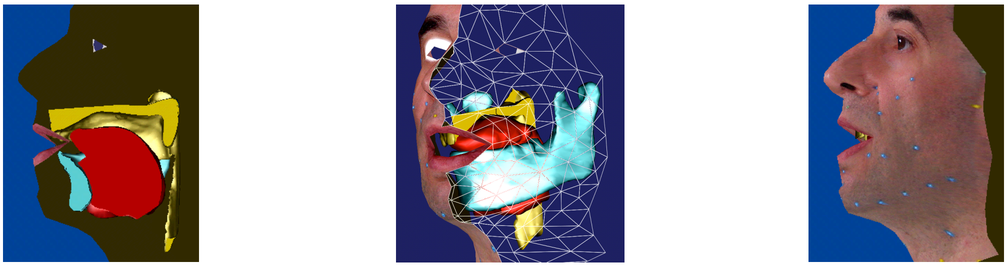

Figure 1: Talking head for different types of displays. Left: "augmented 2D view", middle: "augmented 3D view", right: "complete face in 3D with skin texture"

HMM-based acoustic-to-articulatory inversion

In our visual articulatory feedback system, acoustic-to-articulatory inversion is used to recover trajectories of the EMA coils from the audio speech signal in order to finally animate the talking head. The inversion method (described in [4]) is based on the joint modeling of speech audio data and corresponding articulatory trajectories using multistream Hidden Markov Models (HMM). The inversion is then performed in two stages: (1) a HMM-based recognition stage during which the most likely phonetic sequence is determined from the acoustic signal; (2) a HMM-based synthesis stage during which the articulatory trajectories are inferred from the decoded phonetic sequence and the state durations of the associated HMMs.

This inversion method was evaluated on a corpus consisting of 17 minutes of multimodal speech data (including sentences, VCV and French CVC words). Articulatory activity was recorded synchronously with the audio signal using the Carstens 2D EMA system (AG200). Six coils were respectively glued on the tongue tip, tongue blade, tongue dorsum, upper lip, lower lip and jaw. EMA data were low-pass filtered at 20 Hz. The acoustic wave of each recorded item was parameterized by 12 Mel-frequency cepstral coefficients (MFCC) with their energies and first derivatives. For each phonetic class, acoustic and articulatory streams were modeled by left-to-right, continuous, multistream HMM with 3 emitting states. In order to take coarticulation into account, we adopted a context-dependant modeling strategy by adding information about left and right contexts to phone models. A data-driven state tying procedure was used to deal with the lack of training data for some contexts. All the procedures involving HMM manipulations were done using the HTK and HTS toolkits. The accuracy of the inversion was measured by calculating the root mean square error (RMSE) between the measured and recovered positions of the EMA coils. A jackknife (leave-one-out) technique was employed in order to increase the statistical relevance of our measurements. With a mean phonetic recognition accuracy of 90 %, we obtained an inversion RMSE of 1.6 mm (cf. [4] for more details).

Speaker adaptation

Compared to other approaches (based on ANN or GMM for instance), the mapping between acoustic and articulatory modalities is not performed at the feature level, but at the phonetic level. Based on this consideration, we investigated the possibility to drive the clone by directly decoding the “learner’s voice” (in a CAPT situation) at this level. Because the accuracy of the inversion process strongly depends on the performance of this decoding stage, we added a speaker adaption front-end module to our system. This additional stage results in making the models of the reference speaker (*i.e.* the speaker used to build the talking head) compatible with the learner’s voice (and also with a different acoustic environment). To build the adaptation database, the learner is asked to record a few sentences. Corresponding audio signals are automatically segmented at the phonetic level using a forced-alignment procedure and initial acoustic models. A set of transformation that reduces the mismatch between the initial models and learner adaptation data is then determined using Maximum Likelihood Linear Regression (MLLR). Preliminary tests run on a few speakers showed that only 2 minutes of speech are usually enough to efficiently adapt the initial HMMs of the reference speaker and to achieve a mean recognition accuracy of 85 %.

Conclusion

This paper presents a visual articulatory feedback system based on HMM-based acoustic-to-articulatory inversion and a 3D talking head giving access to internal articulators. A speaker adaption stage was included in the processing chain as a first step toward a speaker-independent system. This approach must now be evaluated in a realistic language training situation. A demonstration of the system at the conference is planned.

Acknowledgements

The authors would like to acknowledge Christophe Savariaux for its implication in the recording of EMA data. This work is supported by the French National Research Agency (ANR), under contract number ANR-08-EMER-001-02 (*Artis* project).

References

[1] Bailly, G., Badin, P., Beautemps, D., and Elisei, F., "Speech technologies for augmented communication," in *Computer Synthesized Speech Technologies: Tools for Aiding Impairment*, Mullennix, J. and Stern, S., Eds.: IGI Global, Medical Information Science Reference, 2010, pp. 116-128.

[2] Badin, P., Ben Youssef, A., Bailly, G., Elisei, F., and Hueber, T., "Visual articulatory feedback for phonetic correction in second language learning,", L2SW, Tokyo, Japan, 2010.

[3] Badin, P., Tarabalka, Y., Elisei, F., and Bailly, G., "Can you 'read' tongue movements? Evaluation of the contribution of tongue display to speech understanding," *Speech Communication*, vol. 52, pp. 493-503, 2010.

[4] Ben Youssef, A., Badin, P. & Bailly, G. (2010). Acoustic-to-articulatory inversion in speech based on statistical models. In AVSP2010, 9th International Conference on Auditory-Visual Speech Processing (K. Sekiyama, S. Sakamoto, A. Tanaka, S. Tamura & C.T. Ishi, editors), vol., pp. 160-165. Hakone, Kanagawa, Japan.