

2D Audiovisual Text-to-Speech Synthesis for Human-Machine Interaction in Dutch

Wesley Mattheyses, Lukas Latacz & Werner Verhelst

Vrije Universiteit Brussel, Dept. ETRO-DSSP

Belgium

{wmatthey, llatacz, wverhels}@etro.vub.ac.be

Speech has always been the most important means of communication between humans. Therefore, using speech in machine-human communication can help in increasing the naturalness of the communication between a computer system and a user. Systems that can make a machine pronounce any given input text are referred to as text-to-speech systems. To further enhance the communication, a talking head can be added to the text-to-speech synthesis, since the addition of this synthetic visual speech mode will improve the intelligibility of the artificial speech (Pandzic *et al.*, 1999). Furthermore, users will perceive this multimodal speech communication as more natural and they will feel more positive and confident if they can see the (artificial) person that is talking to them. In this paper we propose an audiovisual text-to-speech synthesis system for Dutch that is able to create both the target auditory and the target visual speech by using a same audiovisual database, which makes it possible to maximize the intermodal coherence in the audiovisual output signal.

Classical talking heads are constructed by a 3D model from which the polygons vary in accordance with the target phoneme sequence. However, similar to what happened in the research domain of audio-only text-to-speech systems, a trend is noticeable from these model-based synthesis techniques towards datadriven strategies, where the system uses a pre-recorded database to construct the output signal. The most important reason for this is that datadriven synthesizers are capable to produce a more natural output: datadriven visual synthesis makes it possible to attain a quasi photorealistic synthetic visual speech signal. In contrast with the 2D talking heads found in the literature (Bregler *et al.*, 1997; Cosatto & Graf, 2000; Ezzat *et al.*, 2002), our synthesis strategy creates both the auditory and the visual mode of the speech together, which allows us to successfully transfer the original multimodal coherence present in the speech from the audiovisual database to the target output signal. To achieve this, the unit selection technique (Hunt & Black, 1996), which can be found in state-of-the-art auditory text-to-speech systems, was extended to work in the audiovisual domain. This means that from the continuous speech in the database, the system will select an appropriate set of multimodal segments that will be concatenated to construct the output speech signal. This strategy has the advantage that the final output will consist of original combinations of auditory and visual speech fragments, which will maximize the audiovisual correlation in this synthetic signal and thus minimize quality degradations caused by audiovisual coarticulation effects (*e.g.*: the McGurk effect, McGurk & MacDonald, 1976). The synthesizer is provided with a phonetically labeled audiovisual database, from which both the auditory and the visual mode are analyzed beforehand in order to construct several sets of metadata. This metadata contains information about the waveforms and about the video frames in the database and will be used by the system's selection algorithms. The selection of a particular audiovisual segment is based on target cost functions, which indicate how well this segment matches the target speech, and join cost functions which indicate how well two consecutive segments can be concatenated without the creation of disturbing artifacts. To use with our multimodal unit selection technique, cost functions are needed for both the audio track and the video track, since the selection of an audiovisual unit will depend on its multimodal cost. As a target cost we use the phonetic correctness of the audiovisual segment, together with an evaluation of the extended phonetic context (Latacz *et al.*, 2007). Auditory join costs are calculated by using several properties like pitch-, energy- and spectral similarity across the segment joins. Visual join costs are determined by tracking the position and the appearance of the mouth and the other facial parts through the database, together with extra information like the amount of visible teeth present in the video frames. After selection, the audio tracks are concatenated using a pitch synchronous overlap-add technique (Mattheyses *et al.*, 2006), while the video tracks are concatenated by using a morphing technique (Wolberg, 1990) to smooth the joins. During this multimodal concatenation process, the amount of audiovisual asynchrony is kept as small as possible in order to retain the original multimodal correlation.

The proposed synthesis strategy results in the creation of an artificial audiovisual speech signal that exhibits a high coherence between its auditory and its visual mode. This assures a natural perception of the multimodal speech signal: observers truly believe that the person displayed in the visual mode could indeed have been the source of the presented auditory speech signal. Furthermore, a suitable combination of auditory and visual cost functions leads to the selection of an appropriate set of multimodal segments from which a smooth output signal can be constructed.

REFERENCES

- Bregler, C., Covell, M. & Slaney, M., 1997. Video Rewrite: Driving Visual Speech with Audio. *Association for Computing Machinery's Special Interest Group on Graphics and Interactive Techniques*, 353–360.
- Cosatto, E. & Graf, H.P., 2000. Photo-realistic talking-heads from image samples. *IEEE Transactions on multimedia*, 2, 152–163.
- Ezzat, T., Geiger, G. & Poggio, T., 2002. Trainable videorealistic speech animation. *Association for Computing Machinery's Special Interest Group on Graphics and Interactive Techniques*, 21, 388–398.
- Hunt, A. & Black, A., 1996. Unit selection in a concatenative speech synthesis system using a large speech database. *International Conference on Acoustics, Speech and Signal Processing*, 373–376.
- Latacz, L., Kong, Y. & Verhelst, W., 2007. Unit Selection Synthesis Using Long Non-Uniform Units and Phoneme Identity Matching. *6th ISCA Workshop on Speech Synthesis*, 270–275.
- Mattheyses, W., Latacz, L., Kong, Y.O. & Verhelst, W., 2006. A Flemish Voice for the Nextens Text-To-Speech System. *Fifth Slovenian and First International Language Technologies Conference*.
- McGurk, H. & MacDonald, J., 1976. Hearing lips and seeing voices. *Nature*, 264 746–748.
- Pandzic, I., Ostermann J., & Millen D., 1999. Users Evaluation: Synthetic talking faces for interactive Services. *The Visual Computer*, 15, 2330–2340.
- Wolberg, G., 1990. *Digital image warping*. IEEE Computer Society Press.