

# Photorealistic 2D Audiovisual Text-to-Speech Synthesis using Active Appearance Models

Wesley Mattheyses and Werner Verhelst\*  
Vrije Universiteit Brussel, Dept. ETRO-DSSP,  
Interdisciplinary Institute for Broadband Technology IBBT, Brussels, Belgium

**Keywords:** audiovisual text-to-speech synthesis, unit selection, active appearance models

## 1 Introduction

Audiovisual text-to-speech (AVTTS) synthesizers are capable of generating a synthetic audiovisual speech signal based on an input text. A possible approach to achieve this is model-based synthesis, where the talking head consists of a 3D model of which the polygons are varied in accordance with the target speech. In contrast with these rule-based systems, data-driven synthesizers create the target speech by reusing pre-recorded natural speech samples. The system we developed at the Vrije Universiteit Brussel is a data-based 2D photorealistic synthesizer that is able to create a synthetic visual speech signal that is similar to standard 'newsreader-style' television recordings.

## 2 Approach

### 2.1 AV Unit Selection

Our system uses a pre-recorded database containing natural sequences of AV speech from a single speaker. The synthesis strategy is based on the well-known unit selection technique [Hunt and Black 1996]. In order to create the synthetic speech, the system searches in the speech database for a series of suitable segments to match with the target phoneme sequence. The concatenation of these segments results in the target speech signal. We extended this technique to the audiovisual domain: our system selects from the speech database audiovisual segments, consisting of an original combination of auditory and visual speech. This selection is based on both target costs and join costs, which take into account acoustic as well as visual properties of the speech. The major benefit of our approach is the fact that the output speech consists of concatenated original combinations of auditory and visual speech. This implies that the synchrony and the coherence between the output auditory and the output visual speech mode will be maximal. It has been shown that such a high coherence is crucial to achieve a high perceived output quality. Indeed, it is not only necessary that the target text sequence is uttered correctly in both the auditory and the visual mode, we also have to achieve the impression that the displayed talking head could have been the producer of the auditory signal that the user hears [Mattheyses et al. 2009].

---

\*e-mail:wmatthey,wverhels@etro.vub.ac.be

### 2.2 AAM Modeling

2D active appearance models (AAMs) [Edwards et al. 1998] are statistical models that are able to project a set of similar images into a model-space. In addition, a trained AAM makes it possible to generate a new image from a set of AAM model parameters that is given as input. AAMs model two different aspects of an image: the shape and the texture. The shape of an image is defined by a set of landmark points that indicate the position of certain objects that are present in each training image. The texture of an image is determined by its pixel values, which are sampled over triangles defined by the image's landmark points. We applied such an AAM to project every frame from the AVTTS system's speech database on a set of shape- and texture parameters. The unit selection strategy (see section 2.1) now results in the selection of original combinations of waveforms and AAM-parameter trajectories. The AAM-based representation of the visual speech information has the benefit that the shape and the texture properties of the visual speech can be easily treated separately in the selection and concatenation steps. We also designed a technique to further differentiate the processing of the visual speech information among the different shape/texture aspects themselves. For instance, a normalization technique has been designed that is able to remove undesired variations from the visual speech contained in the system's speech database. Furthermore, the AAM-based synthesis approach makes it possible to effectively smooth the concatenated visual speech (by avoiding and removing video concatenation artifacts) while the strength of the visual articulations is unaffected. This is important since it is often seen that the smoothing of the visual mode results in visual speech that appears to be 'mumbled': the visual articulations are too much softened to match with the articulations present in the accompanying auditory speech (i.e., the unimodal smoothing resulted in a decrease of the audiovisual coherence). Sample syntheses produced by our system using the LIPS2008 audiovisual database [Theobald et al. 2008] can be found at <http://www.etro.vub.ac.be/Research/DSSP/DEMO/AVTTS/>.

## References

- EDWARDS, G., TAYLOR, C., AND COOTES, T. 1998. Interpreting face images using active appearance models. In *Int. Conf. on Face and Gesture Recognition*, 300–305.
- HUNT, A., AND BLACK, A. 1996. Unit selection in a concatenative speech synthesis system using a large speech database. In *International Conference on Acoustics, Speech and Signal Processing*, 373–376.
- MATTHEYSES, W., LATA CZ, L., AND VERHELST, W. 2009. On the importance of audiovisual coherence for the perceived quality of synthesized visual speech. *EURASIP Journal on Audio, Speech, and Music Processing SI: Animating Virtual Speakers or Singers from Audio: Lip-Synching Facial Animation*.
- THEOBALD, B.-J., FAGEL, S., BAILLY, G., AND ELISEI, F. 2008. Lips2008: Visual speech synthesis challenge. In *Interspeech '08*, 1875–1878.