MERGING METHODS OF SPEECH VISUALIZATION

Sascha Fagel
Technical University Berlin, Institute for Speech and Communication
sascha.fagel@tu-berlin.de

Abstract

The author presents MASSY, the Modular Audiovisual Speech SYnthesizer. The system combines two approaches of visual speech synthesis. Two control models are implemented: a (data based) di-viseme model and a (rule based) dominance model where both produce control commands in a parameterized articulation space. Analogously two visualization methods are implemented: an image based (video-realistic) face model and a 3D synthetic face model. Both face models can be driven by both the data based and the rule based articulation model.

The high-level visual speech synthesis generates a sequence of control commands for the visible articulation. For every virtual articulator (articulation parameter) the 3D synthetic face model defines a set of displacement vectors for the vertices of the 3D objects of the head. The vertices of the 3D synthetic head then are moved linearly by linear combinations of these displacement vectors to visualize articulation movements.

For the image based video synthesis an image database is searched for appropriate video frames. If no image with facial properties according to the control commands is found, the missing image is generated by deforming a neutral image. MPEG-4 facial definition parameters (FDPs) and additional points in the mouth opening area and around the lower jaw are defined in the neutral image as feature points. A two-dimensional displacement vector is defined for each feature point. For the image deformation a mesh of triangles connecting the feature points is used. The displacement vector of a point in a triangle is interpolated from the displacement vectors of the vertices. Hence, the video synthesis algorithm is capable to use either a database of appropriately annotated video frames or a single neutral image with specified feature points and displacement vectors. Other well known data (image) based audio-visual speech synthesis systems like MikeTalk and VideoRewrite concatenate pre-recorded video sequences. Parametric talking heads like baldi control a parametric face with a parametric articulation model. The presented system demonstrates the compatibility of parametric and data based visual speech synthesis approaches.

The 3D synthetic speech visualization of MASSY improves the intelligibility nearly as much as natural visible speech. The evaluation of its naturalness will follow. The quality of the image based visual speech synthesis yet has to be evaluated in terms of intelligibility and naturalness of the generated images and articulation movements. Although the articulation movements are modeled in a comparatively simple way (linear movements, quasi stationary phases of fixed relative duration and no phase shift between articulators) the audiovisual speech synthesis system is capable to produce the so called McGurk effect with both of the visualization methods. Yet the face models are “somewhat less far away from natural” as the 3D synthetic face is well designed in high resolution and the image based face model is derived from a photo. This leads to the assumption that the (static) appearance of the talking face may be more important for the visual information to be integrated into speech perception than the (dynamic) speech movement modeling – or at least more important than considered in the past.