Towards Adaptive Audio-Visual Dialect Speech Synthesis

Michael Pucher, Dietmar Schabus, Gregor Hofer, Nadja Kerschofer-Puhalo, Sylvia Moosmüller

ABSTRACT

The goal of our work is to investigate multimodal adaptation for audio-visual dialect speech synthesis. Human speech is multimodal and therefore we aim at modeling both the audio and visual signals jointly. Furthermore, in speech behavior we are confronted with intra-speaker variability (e.g. variability in dependence on different speech situations, speaking tasks or emotional states of the speaker) and inter-speaker variability (e.g. variability across sociolects and/or dialects). The second type of variation can be modeled by adapting average models of speakers with different dialects to a speaker of a specific dialect. Dialect is chosen as a source of variation between speakers to extend our previous work on Viennese sociolinguistics to other Austrian dialects and to conduct basic research on the audio-visual synthesis of dialects.

Corpus Design

• We record an audio-visual corpus of two Austrian varieties (Middle Bavarian and Southern Bavarian) for 8 speakers.

• Based on a phonetic analysis (Table 1) of the dialects we create a phonetically balanced recording script.

• This script will contain spontaneously uttered sentences and elicited sentences.

M. Pucher et al., Phone set selection for HMM-based dialect speech synthesis, DIALECTS 2011 (EMNLP 2011).

Acoustic Modeling

• The context-dependent quinphone acoustic models are clustered with the Standard shared decision-tree clustering (Figure 1) for hidden Markov models (HMMs).

• Each state is clustered by a separate tree using phonetic and prosodic features.

• The average voice models are trained with speaker adaptive training (SAT).

Table 1: Consonants for Bad Gossen dialect (Middle Bavarian).

<table>
<thead>
<tr>
<th>Phonemes</th>
<th>DIALECTS 2011 (EMNLP 2011)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05 0.1</td>
<td></td>
</tr>
<tr>
<td>0.15 0.2</td>
<td></td>
</tr>
<tr>
<td>0.3 0.4</td>
<td></td>
</tr>
</tbody>
</table>

Visual Modeling

• For visual synthesis we record face markers using an infrared-based marker tracking system (Figure 3, left).

• The recorded marker sequence is then used to animate a virtual head by NaturalPoint (http://www.naturalpoint.com).

• These reduced features are used to train a HMM similarly to the acoustic HMM training.

• At synthesis time a given text input is converted to a sequence of phone and context labels from which HMMs are selected.

• Then the HMM parameter generation algorithm is used to generate a sequence of marker points from the given HMM.

• This synthesized sequence is then used to animate an avatar (Figure 3, right).

D. Schabus et al., Simultaneous Speech and Animation Synthesis, SIGGRAPH 2011.

Adaptive Audio-Visual Modeling

• For adaptive audio-visual modeling we record a multi-speaker audio-visual database.

• Multiple speakers are used to train an average audio-visual model using speaker adaptive training (SAT) (Figure 3).

• Here it is possible to train audio and visual models jointly by combining them in one stream or by using a multi-stream model.

• Furthermore it is possible to train separate audio and visual models and combine them via a common duration model.

• At adaptation time audio-visual data of a certain speaker is used to adapt the average model.

• At synthesis time we generate a synchronized acoustic and visual sequence.

• The advantage of the adaptive approach is the possibility to use an average (background) model that is trained on a large amount of training data and only need a small amount of adaptation data from the target speaker.

• The adaptive approach has already been used successfully in acoustic speech synthesis.

• The flexibility of HMM modeling also allows for different interpolation methods that can be used to create transitions between models.


Figure 2: Recording, synthesis and animation of visual speech. Virtual head by NaturalPoint (http://www.naturalpoint.com).

Figure 3: Adaptive HMM-based audio-visual speech synthesis.