Regionalizing Virtual Avatars
Towards Adaptive Audio-Visual Dialect Speech Synthesis

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

FTW Telecommunications Research Center Vienna
Vienna, Austria

{pucher,schabus,hofer}@ftw.at

Austrian Academy of Sciences (ÖAW), Acoustics Research Institute
Vienna, Austria

{nadja.kerschhofer,sylvia.moosmueller}@oeaw.ac.at

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 sociolects 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).*

HTK	IPA	#	HTK	IPA	#
s	s	207	t	t	204
d	d	179	n	n	171
m	m	115	k	k	98
h	h	84	g	ɡ	79
v	v	79	f	f	62
pf	pf	3	S	ʃ	49
N	n̥	42	l	l	41
b	b̥	31	ts	ts	27
ng	ŋ	19	p	p	17
w	β	14	L	l̥	12
X	x	11	c	c	10
RX	χ	9	j	j	7
R	ʀ	67	ks	ks	3

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

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).

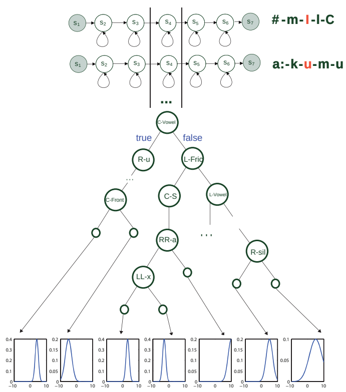

Figure 1: Shared decision-tree clustering.

Visual Modeling

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

- For visual synthesis we record face markers using an infrared-based marker tracking system (Figure 3, left).
- The recorded marker sequence is reduced to a lower dimensional space using *principal component analysis* (PCA).
- 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.

Adaptive Audio-Visual Dialect Synthesis*, https://portal.ftw.at/projects/avds/.*

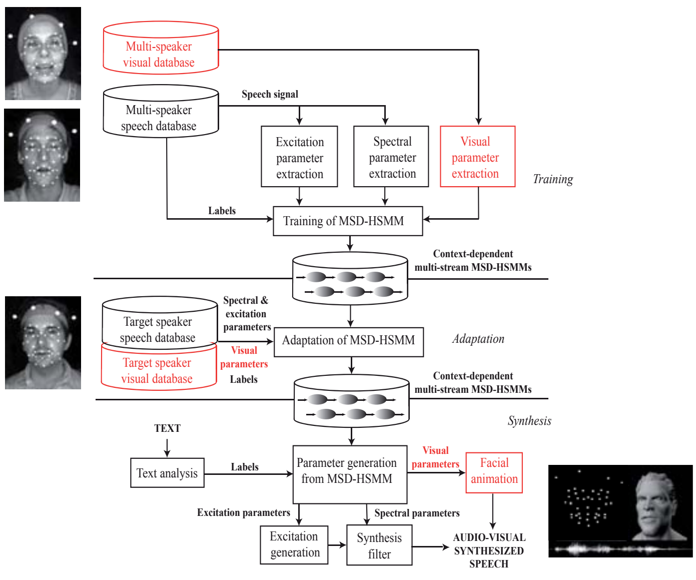

Figure 3: Adaptive HMM-based audio-visual speech synthesis.

Der Wissenschaftsfonds.
This work is funded by the Austrian Science Fund (FWF): P22890-N23

Competence Centers for Excellent Technologies

CogSys 2012, 5th International Conference on Cognitive Systems, Vienna.