ISCA Archive
http://www.isca-speech.org/archive

FAAVSP - The 1st Joint Conference on
Facial Analysis, Animation, and
Auditory-Visual Speech Processing
Vienna, Austria,
September 11-13, 2015

Comparison of dialect models and phone mappings in HSMM-based visual dialect speech synthesis

Dietmar Schabus, Michael Pucher

FTW Telecommunications Research Center Vienna, Austria
schabus@ftw.at, pucher@ftw.at

Abstract

In this paper we evaluate two different methods for the visual synthesis of Austrian German dialects with parametric Hidden-Semi-Markov-Model (HSMM) based speech synthesis. One method uses visual dialect data, i.e. visual dialect recordings that are annotated with dialect phonetic labels, the other methods uses a standard visual model and maps dialect phones to standard phones. This second method is more easily applicable since most often visual dialect data is not available. Both methods employ contextual information via decision tree based visual clustering of dialect or standard visual data. We show that both models achieve a similar performance on a subjective pair-wise comparison test. This shows that visual dialect data is not necessarily needed for visual modeling of dialects if a dialect to standard mapping can be used that exploits the contextual information of the standard language.
Index Terms: visual speech synthesis, dialect

1. Introduction

Visual speech synthesis techniques have possible applications in computer games and films. Generating visual speech directly from audio data is nowadays a state-of-the-art technique in facial animation in the computer games industry [1]. In this paper we investigate visual dialect text-to-speech synthesis where we generate an acoustic and visual signal of a certain speaker from given text.

We evaluate two different methods for the visual synthesis of Austrian German dialects with parametric Hidden-Semi-Markov-Model (HSMM) based speech synthesis. One method uses visual dialect data, i.e. visual dialect recordings that are annotated with dialect phonetic labels, the other methods uses a standard visual model and maps dialect phones to standard phones. This second method is more easily applicable since most often visual dialect data is not available.

By comparing these two methods we analyze if it is necessary to use visual dialect data for visual synthesis. It is clear that there exists a "visual dialect" to a certain extent, since there are phonemes in the dialect that are not existing in the standard. As can be seen in Figure 1 there are many phones that are exclusive dialect phones for the case of the Austrian German dialects that are used in this study. Within the whole synthesis pipeline and subjective evaluation it might however still happen that the differences between dialect and standard visual phones are too small to result in perceivable quality differences. The results of this paper can be extended to other languages with a similar relation between dialect and standard.

The mapping between dialect and standard is performed on the level of phones, but in the synthesis stage the full contextual information of the standard language visual model is taken into account.

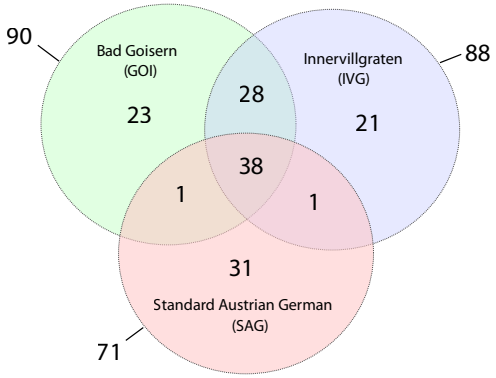

Figure 1: Dialect (GOI, IVG) and Standard Austrian German (SAG) phone set.

This paper is organized as follows: in Section 2 we describe the corpus and in Section 3 we present audiovisual modeling. Section 4 describes the visual dialect modeling method and Section 5 presents the evaluation. In Section 6 and 7 we discuss and conclude the paper.

2. Corpus

The Goisern and Innervillgraten Dialect Speech (GIDS) Corpus [2] is a collection of audiovisual speech recordings for research purposes. It consists of a total of 7068 sentences spoken by eights speakers from two Austrian villages, Bad Goisern (BG) and Innervillgraten (IVG). For each speaker, about two thirds of the recorded sentences are in the speaker's respective dialect and the rest is in Regional Standard Austrian German (RSAG). The dialect of Bad Goisern in the Salzkammergut region belongs to the (South)-Central Bavarian dialects, and the dialect of Innervillgraten in the East Tyrol region belongs to the Southern Bavarian dialect family as shown in Figure 2.

Standard Austrian German (SAG) refers to the variety spoken by the upper social classes of the big cultural centers located predominantly in the Middle Bavarian region. Since the IVG and GOI speakers were genuine dialect speakers, meaning that they were raised in the respective dialect and learned SAG only in school, SAG spoken by these speakers contained also regional features. Therefore, the SAG variety produced by the GOI and IVG speakers is referred to as regional standard Austrian German (RSAG). A detailed analysis of the dialects can be found in [3].

After a careful phonetic analysis we compiled sets of phonetically balanced sentences (656 for IVG and 665 for GOI)

with respect to the phone set established for the dialect, the frequency of occurrence of each phone in the data, and the context specific variation of phones. The utterances of the recording script were extracted from a larger corpus of material consisting of 18-20 hours of recordings for each dialect with at least 10 speakers per dialect. These sentences consisted of spontaneous speech (elicited with key words) and translation tasks. We created a lexicon of words occurring in the script. The script was divided into a training and testing part. In the final audio-visual recordings we recorded 2 male and 2 female speakers per dialect, i.e., 8 speakers in total.

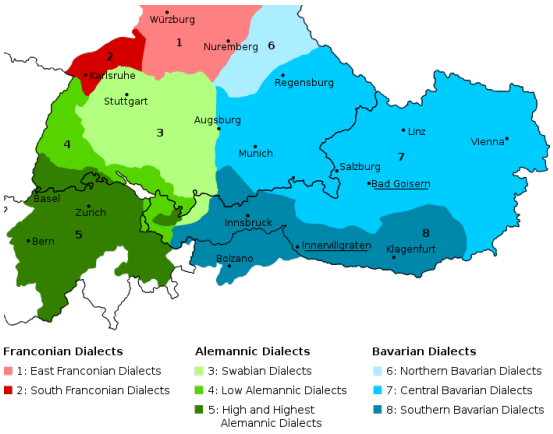

Figure 2: Upper German dialects

The recordings consist of optical 3D facial motion tracking data, captured with a NaturalPoint OptiTrack Expression system,[1] the greyscale video data also recorded by the same system, and studio quality audio.

For each of the recorded utterances, the corpus contains, a RIFF wave audio file, facial marker data in the form of a matrix stored as a text file, a gray scale video from the optical system, the sentence of the utterance in plain text, a text file listing the phones spoken in the utterance including begin and end times of all phones, and a quin-phone full-context label file.

3. Audiovisual modeling

For audiovisual modeling we train acoustic and visual models within the Hidden-Semi-Markov-Model (HSMM) parametric speech synthesis framework.

For audio-only modeling, we apply the state-of-the- art CSTR/EMIME HTS system [4] without modifications. For visual-only modeling, we use the same system but with only one feature stream for the visual Principal Component Analysis (PCA) space features. In order to obtain the same frame rate as the audio features (5 ms frame shift, i.e., 200 frames per second), we have up-sampled (interpolated) the visual features from their native 100 frames to 200 frames per second. Similar to the cepstral features, they are also augmented by their dynamic features and the models are clustered using the same set of questions. This results in a speaker-dependent text-to-visual speech system, like we have investigated in previous work [5].

For audiovisual synchronization we use the audio duration copy as presented in [6]. In this method we use the audio duration model for both audio and visual synthesis. This is equivalent to replacing the visual duration models and trees with the ones obtained from audio training. The advantage here is the tighter synchronization, a possible disadvantage is that a new duration model is forced upon the visual system which might not match the visual feature models.

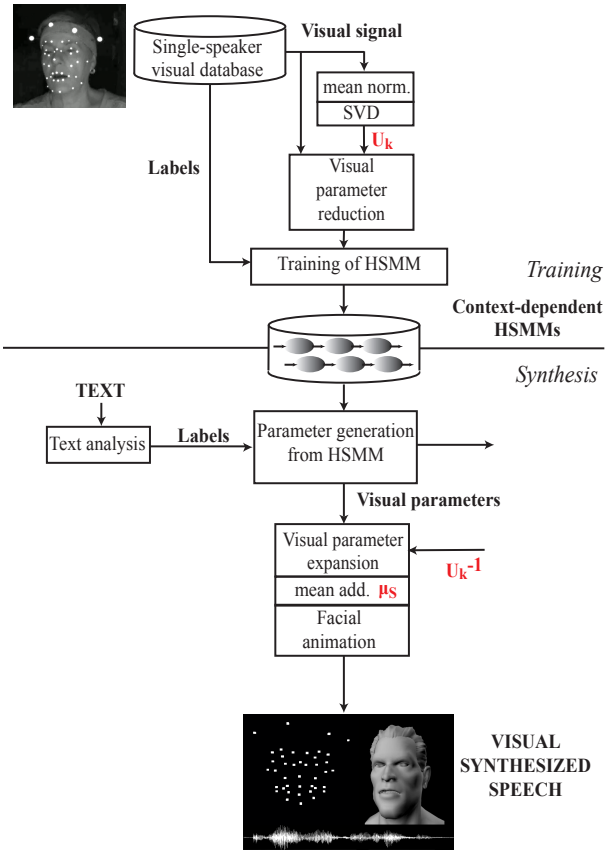

Figure 3: Speaker dependent visual modeling system.

The system for speaker dependent visual modeling is shown in Figure 3. It consists of a training and synthesis step. Before training the visual features are reduced with PCA. We use 30 dimensional reduced features, which was found to be optimal for visual modeling [7]. After synthesis the visual features are projected back into the full dimensional space with dimensionality 123 (x, y, z coordinates of markers), which are then used to drive a talking head.

4. Visual dialect modeling

If synchronous acoustic and visual speech recordings are available for a given speaker in the desired language variety, it is straightforward to produce an audiovisual synthesizer from this data. In that case, creating an audiovisual "voice" for a dialectal variety is no different from creating a voice in the standard variety. However, we want to investigate the following restricted scenario: from a given speaker, synchronous acoustic and visual speech data is available in the standard variety (Standard Austrian German in our case), and from the same speaker also dialectal speech data is available (Bad Goisern (BG) or Innervillgraten (IVG) dialect in our case), but only in the acoustic modality and not in the visual modality. Can the HMM-based visual speech model of that speaker, which was trained using speech in standard Austrian German only, be used to generate dialectal

[1] http://www.naturalpoint.com/optitrack/

visual speech? In particular, can this be done without transforming the visual model, i.e., can the standard models be used for generating visual dialect trajectories of sufficient quality?

As Figure 1 shows, there are 39 phones for both dialects which also appear in the standard. For these phones, it is reasonable to assume that they can be adequately produced by the standard visual voice model, also for use together with acoustic dialectal speech. However, the Bad Goisern and Innverillgraten dialect speech recordings contain 51 and 49 phones, respectively, which do not appear in the standard. These are unknown symbols to the standard visual models; practically speaking, none of the questions in the clustering trees will match an input label with such a phone.

The fine and precise distinction between phones we make in the transcriptions of our data is however purely acoustically motivated. For visual speech, it can be argued that there are groups of two or more phones, which are equivalent in terms of facial speech motion, although they are acoustically distinct. The concepts of visemes [8, 9, 10] and phoneme equivalence classes [11] are based on this kind of argumentation. Carrying this idea over to our scenario means that we might be able to generate visual dialect speech motion from standard visual speech models, if we find a visually equivalent or sufficiently similar phone in the standard phone set for each dialect phone.

The simplest way to achieve this is to manually define a mapping from the 51 resp. 49 dialect-only phones to adequate standard phones, based on phonetic knowledge. Other methods based on acoustic and/or visual similarities computed on the data could also be investigated; for the experiments in this paper, however, we have simply defined a mapping from each dialect-only phone to the "most similar" standard phone, according to our judgment.

Hence, in our scenario of having from a speaker acoustic dialect speech data on the one hand and visual standard speech data on the other hand, with the goal of generating audiovisual dialect speech, we use the following strategy: we train an acoustic voice model on the acoustic dialect data and a visual standard voice on the visual standard data. At synthesis time, given a label sequence for dialect speech, we generate acoustic speech from the acoustic model in the normal way. Then, we apply the manually defined phone mapping to the input label sequence, which results in a label sequence that contains phones from the standard only. Using this new label sequence, we synthesize visual speech from the visual model. In order to ensure synchrony between the two generated sequences, we apply the duration-copy method as described in Section 3, i.e., the state durations predicted by the acoustic dialect model are used for both acoustic and visual synthesis. Within the HSMM-based synthesis framework we can use the full visual contextual information of the standard language by using such a phone mapping.

5. Evaluation

In order to assess the success of the method described in the previous section, we have carried out a subjective perceptual experiment comparing audiovisual speech generated using that method to audiovisual speech generated in the regular fashion, where visual dialect data is used to train a visual dialect model.

We used recordings from one speaker for each of the two dialects, both female. For both methods, the acoustic speech is generated from the same acoustic dialect speech model, which was trained using 623 and 618 utterances for the Bad Goisern and Innervillgraten dialect, respectively. This means that the generated acoustic speech is identical for the phone mapping and the baseline method.

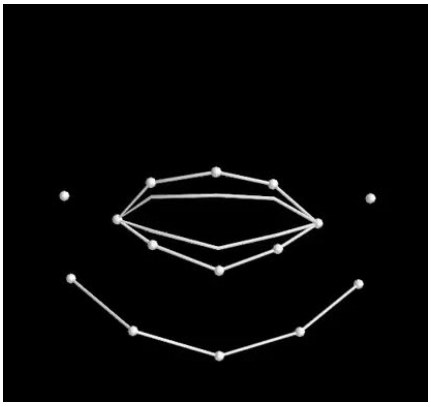

Figure 4: Rendering of synthesized facial maker data as shown during the subjective evaluation. Supporting lines for the outer lip contour and the chin are added between the respective markers. The inner lip contour is added automatically based on six points which are defined on a fixed distance from their corresponding outer lip points.

For the baseline method, visual speech is generated from a visual dialect model which was trained using 200 dialect utterances. For the phone mapping method, visual speech is generated from a visual standard model which was trained using 200 standard utterances, with the phone mapping being applied at synthesis time.

Using the two methods, we generated audiovisual speech for 30 dialect utterances per speaker which were not part of the training data. 10 test subjects were asked to judge in pairwise comparisons, which of the two presented audiovisual speech stimuli had better agreement between the acoustic speech and the visual speech motion. "No preference" was also an option. 30 utterances in two dialects give rise to 60 comparisons in total. In the evaluation, each of these 60 comparisons was judged by 5 different subjects, i.e., we have 300 judgements in total and each subject saw 30 comparisons.

We decided to present visual speech motion in the form of synthesized point movement of the lower part of the face, rather than applying the point data to a 3D head via a retargeting procedure. Hence, the test subjects saw renderings of the synthesized facial markers with supporting lines, as shown in Figure 4. One reason for this is that the retargeting procedure is ruled out as an influencing factor. As a second reason, we believe that differences between two given sequences are easier to see on the marker data than on the final head animation.

The results are shown in Figure 5 as overall results (top), results per dialect (middle) and results per listener (bottom). Overall, i.e., of all 300 decisions made in the experiment, 98 times the baseline method was preferred, 106 times the phone mapping method was preferred, and 96 times no preference was stated.

6. Discussion

In our evaluation we saw that the phone mapping used produces visual synthesis of similar quality as the models trained from visual dialect data. There might however be more subtle visual dialect differences between some phones, which are however

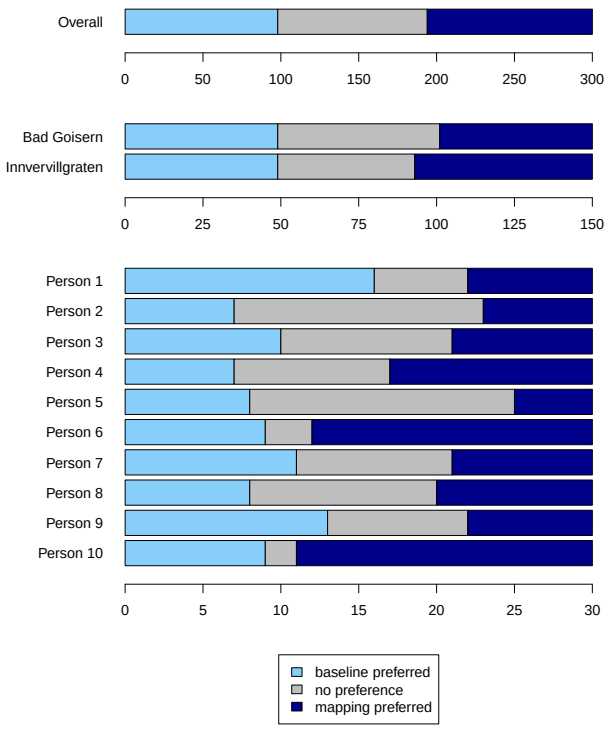

Figure 5: Results of the subjective experiment: overall results (top), results split per dialect (middle), results split per listener (bottom).

not captured by our evaluation. Such subtle differences can be found with a different evaluation approach that focuses on the most different samples within the space of possible samples.

In general it is very difficult to measure differences in visual modeling, since untrained listeners as the ones that were asked for this study, have difficulty to spot small differences between videos of marker sequences as those shown in Figure 4.

Furthermore we have to restrict our results to the investigated dialect-standard pair or pairs that are similar in terms of overlapping phones or visual differences. There might be other standard-dialect pairs where visual differences between the varieties are much larger and visual dialect models therefore more beneficial.

Differences between visual dialect and standard might also be more important when using a different modeling framework than HSMM-based visual modeling. By training a large visual decision tree for the standard language and employing this tree together with the phone mapping for dialect phones we are able to select a sequence of models for the dialect that closely follows the visual movements in the dialect.

The results might also change if an acoustic dialect model of a speaker is combined with a visual standard model of a different speaker. In this study the acoustic and visual models were trained from data of the same speaker to restrict the investigation to intra-speaker variation between dialect and standard.

7. Conclusion and future work

We have shown that the visual modeling of dialects in the HSMM framework can be done successfully with a mapping between standard and dialect, or with the usage of dialect specific training data. As the first approach is less time consuming it will be the preferred method for many applications.

Although our result is negative with respect to the benefits of using visual dialect data, it has to be kept in mind that our result depends on several preconditions such as having the same standard and dialect speaker (i.e. only intra-speaker no inter-speaker variability), having similar differences between standard and dialect as in the used Austrian German dialects, using the HSMM-based modeling framework, and evaluating on a random sample of utterances. Changing any of these preconditions might show that visual dialect speech data is beneficial after all.

In the future we want to investigate how the phone mapping can be determined automatically given acoustic and/or visual similarity measures derived from audiovisual recordings. We also want to investigate how the usage of models from different speakers influences the quality of visual dialect models. Furthermore we want to evaluate other audio-visual synchronization strategies for visual dialect modeling like joint audiovisual modeling.

8. Acknowledgements

This work was supported by the Austrian Science Fund (FWF): P22890-N23 and P23821-N23. The Competence Center FTW Forschungszentrum Telekommunikation Wien GmbH is funded within the program COMET – Competence Centers for Excellent Technologies by BMVIT, BMWFJ, and the City of Vienna. The COMET program is managed by the FFG.

9. References

[1] SpeechGraphics, "Speech Graphics - Audio-driven facial animation," http://www.speech-graphics.com/, 2015.

[2] FTW, "GIDS - Goisern and Innervillgraten audiovisual dialect speech corpus," http://cordelia.ftw.at/gids/.

[3] M. Toman, M. Pucher, S. Moosmüller, and D. Schabus, "Unsupervised interpolation of language varieties for speech synthesis," *Speech Communication*, 2015 (accepted).

[4] J. Yamagishi and O. Watts, "The CSTR/EMIME HTS system for Blizzard challenge 2010," in *Blizzard Challenge Workshop*, Kansai Science City, Japan, 2010.

[5] D. Schabus, M. Pucher, and G. Hofer, "Speaker-adaptive visual speech synthesis in the HMM-framework," in *Proc. INTERSPEECH*, Portland, OR, USA, 2012, pp. 979–982.

[6] ——, "Joint audiovisual hidden semi-Markov model-based speech synthesis," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 8, no. 2, pp. 336–347, April 2014.

[7] ——, "Objective and subjective feature evaluation for speaker-adaptive visual speech synthesis," in *Auditory-Visual Speech Processing, AVSP 2013, Annecy, France, August 29 - September 1, 2013*, 2013, pp. 37–42. [Online]. Available: http://www.isca-speech.org/archive/avsp13/av13_037.html

[8] C. G. Fisher, "Confusions among visually perceived consonants," *Journal of Speech, Language, and Hearing Research*, vol. 11, pp. 796–804, dec 1968.

[9] T. Chen, "Audiovisual speech processing," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 9–21, jan 2001.

[10] D. W. Massaro, M. M. Cohen, M. Tabain, J. Beskow, and R. Clark, "Animated speech: Research progress and applications," in *Audiovisual Speech Processing*, G. Bailly, P. Perrier, and E. Vatikiotis-Bateson, Eds. Cambridge University Press, 2012, pp. 309–345.

[11] L. E. Bernstein, "Visual speech perception," in *Audiovisual Speech Processing*, G. Bailly, P. Perrier, and E. Vatikiotis-Bateson, Eds. Cambridge University Press, 2012, pp. 21–39.