ISCA Archive
http://www.isca-speech.org/archive
8th ISCA Speech Synthesis Workshop • August 31 – September 2, 2013 • Barcelona, Spain

Multi-variety adaptive acoustic modeling in HSMM-based speech synthesis

Markus Toman, Michael Pucher, Dietmar Schabus

Telecommunications Research Center (FTW), Vienna, Austria

{toman,pucher,schabus}@ftw.at

Abstract

In this paper we apply adaptive modeling methods in Hidden Semi-Markov Model (HSMM) based speech synthesis to the modeling of three different varieties, namely standard Austrian German, one Middle Bavarian (Upper Austria, Bad Goisern), and one South Bavarian (East Tyrol, Innervillgraten) dialect. We investigate different adaptation methods like dialect-adaptive training and dialect clustering that can exploit the common phone sets of dialects and standard, as well as speaker-dependent modeling. We show that most adaptive and speaker-dependent methods achieve a good score on overall (speaker and variety) similarity. Concerning overall quality there is no significant difference between adaptive methods and speaker-dependent methods in general for the present data set.

Index Terms: speech synthesis, dialect, voice modeling, adaptation

1. Introduction

Speech synthesis is an important part of human-machine communication systems. At present, speech synthesis systems are mostly restricted to standard varieties, which implies a strong limitation on possible applications.

The dialect or accent of a speaker is an important part of the persona of a voice-based user interface since "there is no such thing as a voice user interface with no personality" [1]. Perception of sociolect and dialect influence our evaluation of speakers' attributes like competence, intelligence, and friendliness. Persona is defined as the "standardized mental image of a personality or character that users infer from the applications voice and language choice" [1], where speech synthesis is an essential part of a spoken dialog system's persona.

To build speech synthesis systems that are able to use a range of different varieties it is important that we have methods that allow for a quick development of these voices. Methods based on adaptation are therefore a natural choice. Furthermore, we can exploit the fact that varieties (dialects and sociolects) and standards have an overlapping phone set. This overlap is illustrated in Figure 1 for the varieties of German we consider in this paper. It can be seen that 38 phones are shared across all three varieties. The small phone set overlap reflects the fact that there is a number of dialect phones that are characteristic and mark differences between standard and variety.

The modeling of accented speech data has received some interest in the last years [2, 3, 4] but the modeling of dialects that differ significantly from the standard language in terms of phonetics and the lexicon is still not widely investigated. This is of course also due to the lack of resources and the difficulty to acquire them.

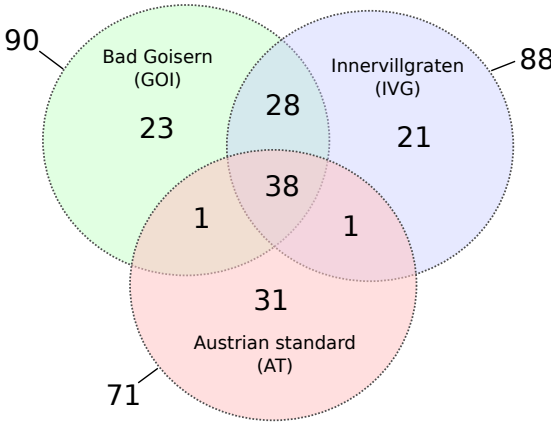

Figure 1: *Overlapping phone sets.*

In [5] it was shown how adaptive dialect modeling methods can be applied to the modeling of two different varieties, namely standard Austrian German and Viennese dialect. Here we show advanced adaptive modeling methods for varieties and evaluate these methods with three Austrian German varieties, namely standard Austrian German and two dialects.

In [6] we investigated cross-variety speaker adaptation between standard Austrian German and the dialects of Innervillgraten and Bad Goisern. This method is based on work of [7] and findings of the EMIME project [8].

2. Speech data and recording

We have recorded and annotated phonetically balanced speech data in different Austrian varieties from Innervillgraten (IVG), Bad Goisern (GOI) [9] and standard Austrian German (AT). In this paper we focus on the modeling of these varieties. The dialects in Austria can be divided into Middle-Bavarian, South-Bavarian, and Alemannic dialects. To cover different regions with as many speakers as possible, we decided to model one Middle-Bavarian and one South-Bavarian dialect in addition to the standard. To restrict the possible variance in the data, we restricted the recordings to a small village in each region: Bad Goisern in Upper Austria for the Middle-Bavarian dialect family and Innervillgraten in East Tyrol for the South-Bavarian dialect family. Initial linguistic studies exist for both dialects [9, 5] but no phone set, corpus, recording script, or synthesizer was available for them.

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 fre-

Table 1: *Dialect modeling approaches.*

			Data Dependency	
Name	Target	# utt.	Speaker	Dialect
SD-DD (AT)	AT	198	✓	✓
SD-DD (IVG)	IVG	618	✓	✓
SD-DD (GOI)	GOI	622	✓	✓
SI-DD (AT)	AT	1790	×	✓
SI-DD (IVG)	IVG	1236	×	✓
SI-DD (GOI)	GOI	1244	×	✓
SI-SN	AT/IVG/GOI	4270	×	×
SI-SDN	AT/IVG/GOI	4270	×	×
SI-SDNC	AT/IVG/GOI	4270	×	×
DHN	AT/IVG/GOI	4270	×	×

quency 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 recordings we recorded 4 speakers (2 male, 2 female) for each dialect. Here we only train models with the male speakers. For our training we have 4 dialect speakers (2 IVG and 2 GOI speakers) where we have dialect and standard data for each speaker, and 1 standard speaker.

The speakers had to fulfill the following linguistic criteria

- "Native speaker", i.e., raised within the dialect
- Consistent application of characteristic phonological processes (e.g., assimilations, deletions)
- Lexical knowledge and morpho-syntactic competence

For recording the dialect data we used a setting where the speaker can hear the utterance he/she is supposed to say and at the same time see an orthographic transcription of the utterance. This is not necessary when an orthographic standard is available and the speakers know how to produce speech from the standard transcription. With this approach we aim to minimize the linguistic variation between the orthographic transcription and the actual spoken utterances of a speaker. Nevertheless, there is still some variation due to fact that speakers may forget what they heard or attempt to correct the reference utterance in case of disagreement.

3. Modeling approaches

Our adaptive speech synthesis system [12] is based on the HSMM-based speech synthesis system (HTS) published by the EMIME project [8]. We use different sets of decision tree questions for each variety. These are partially handcrafted as well as automatically generated from our phone set definitions.

Table 1 defines the modeling approaches that we investigated. SD and SI refer to Speaker-Dependent and Speaker-Independent modeling, DD and DI refer to Dialect-Dependent and Dialect-Independent modeling and SN, SDN, SDNC and DHN refer to Speaker-Normalization, Speaker-Dialect-Normalization, Speaker-Dialect-Normalization with dialect Clustering and Dialect-Hierarchical-Normalization training, respectively. For dialect-dependent modeling, we train average models for each dialect. For dialect-independent modeling, we consider the following approaches: In SI-SN, we train a single model using data from all speakers. SI-SDN means to divide a set of speech data in two varieties uttered by a single speaker (able to speak both varieties) into two subsets of speech data uttered by two different pseudo-speakers (Figure 2). In this example, for speaker 1 AT and IVG recordings exist. Speaker 1 will then be treated as two different speakers, one AT and one IVG speaker. The idea of SI-SDNC is to add dialect information as a context for sub-word units and perform decision-tree-based clustering of dialects in the training of the HSMMs.

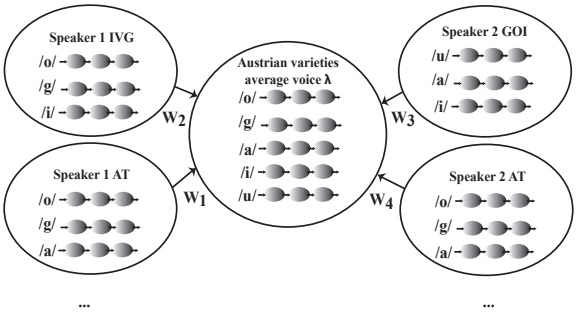

Figure 2: *Speaker-dialect-normalization - SI-SDN.*

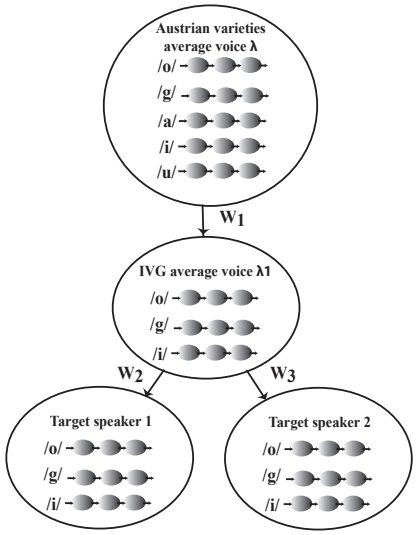

Figure 3: *Dialect-hierarchical normalization - DHN.*

In the clustering of dialects, new questions that identify the variety of an utterance (Is_ivg, Is_goi, Is_at) are added to a set of questions for the decision-tree-based clustering and minimum description length (MDL) based automatic node-splitting [13] is performed. Variety is treated as a clustering context together with other phonetic and linguistic contexts and it is included in the single resulting acoustic model. Note that a decision tree was constructed independently for each combination of state index and acoustic parameter (mel-cepstrum, $\log F_0$, band aperiodicity) and duration. The same idea has been reported for multi-accented English average voice models [14].

In the resulting clusterings, we observe that the first ques-

Table 2: *Occurrences of variety questions in decision trees.*

	# of occurrences				
Feature	State 1	State 2	State 3	State 4	State 5
mel-cepstral	76	139	53	54	67
log F0	28	62	70	44	33
bndap	23	24	37	28	29
duration			70		

tion concerning the variety is used near the roots of the decision trees. Figure 4 shows this part of the constructed decision tree for the mel-cepstral parameters of the third (middle) state and Figure 5 the corresponding duration parameter clustering tree. These are the top-most occurrences of variety questions in the trees and they appear on level 4 in the mel-cepstral-state-3 tree and on level 5 in the duration tree.

Overall occurrences of variety questions in mel-cepstral, logF0 and duration decision trees can be seen in Table 2. It can be seen that variety class questions are relevant in all states. In this example, "Is_ivg" means "Is the current utterance in Innervillgraten dialect?" and "Is_goi" means "Is the current utterance in Bad Goisern dialect?". This means that after these questions, separate Gaussian pdfs are produced for the different dialects. We also observed the labels which have been used to train each single pdf. In SDN only 928 pdfs where estimated using data from a single variety and 1620 pdfs using data from more than one variety. For SDNC, 2431 pdfs were estimated using single variety data and 322 pdfs using data from multiple varieties. This also shows the effect of the the variety questions on the clustering.

In addition to SD-DD, SI-DD, SI-SN, SI-SDN and SI-SDNC, which were already applied for AT and Viennese Dialect (VD) data in the past [5], we also apply dialect-hierarchical normalization (DHN) in this paper. In DHN, a general dialect-independent voice model is trained first, from which then specific dialect-dependent voice models are adapted. Finally, speaker-specific voice models are adapted from these, as shown in Figure 3. Furthermore, we extend this previous work to three different varieties.

We applied model adaptation with AT, IVG, and GOI data to all models. Therefore we have 30 voices in total, where 25 are adapted voices and 5 are speaker- and dialect-dependent voices [1].

4. Evaluation

To assess the quality of the synthetic voices resulting from the different modeling approaches described in Section 3, we have carried out a subjective evaluation with 21 test listeners (8 female, 13 male, aged 20 to 55, mean age 28.95). For each of the three varieties, we have held out 10 test utterances from the training data, in order to allow comparison also to recorded samples, and synthesized each of them using all of the methods for each of our five speakers. Comparing any two models for each (speaker, utterance)-combination gives rise to 1050 comparisons in total, which we distributed among our 21 listeners such that each listener heard each (speaker, utterance)-combination once and each method-pair two to three times.

[1] Synthesis samples on http://userver.ftw.at/~mtoman/ssw2013/m

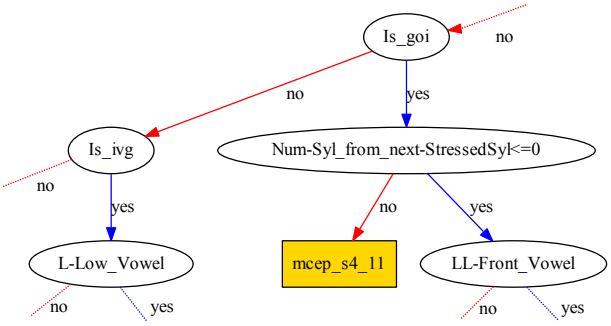

Figure 4: *Dialect clustering results for state 3 of mel-cepstral decision tree.*

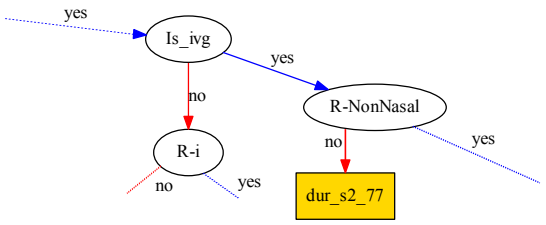

Figure 5: *Dialect clustering results for duration decision tree.*

For each of their 50 comparisons, the listeners heard a recorded reference sample and two samples from two different methods, where all three samples contained the same utterance from the same speaker. After listening to each of the three sound files as many times as they liked, they were asked to tell which of the two samples they felt to be more similar to the reference sample. *Recorded* was also added as a method, i.e., in some comparisons the reference sample and one of the samples in question actually contained the same (recorded) signal. There was also a "tied" option (both samples equally similar to the reference).

The results are given in Table 3, where we have counted the number of "won" comparisons and the number of "ties" for each method pair. In the last column, the symbol "$*$" indicates statistical significance of the preference scores according to Bonferroni-corrected Pearson's χ^2-tests of independence with $p < 0.001$. Even with a relaxed significance threshold of $p < 0.05$, only one additional significance appears (indicated by "$(*)$" in Table 3), but due to the large number of ties in this case (30), we do not consider this a meaningful difference between the two methods SI-SDN and SI-SN.

Additionally, we asked the listeners to specify the degree of similarity for the "winning" method (or both methods, in case of a tie), by choosing one of the five options "very similar", "similar", "no opinion", "different" and "very different". The results are given in Figure 6 as frequency bar plots, where 1 means "very similar" and 5 means "very different".

5. Analysis

We have investigated different adaptive modeling approaches for multi-variety modeling. For the pair-wise comparison of methods we see no significant differences between adapted and speaker-dependent methods, with the exception of dialect-hierarchical training (DHN), which is worse than all other meth-

Table 3: *Subjective pair-wise comparison scores.*

Compared methods			wins	ties	sig.
DHN	:	recorded	1 : 49	0	∗
DHN	:	SD-DD	6 : 26	18	∗
DHN	:	SI-SDN	7 : 26	17	∗
DHN	:	SI-SDNC	5 : 34	11	∗
DHN	:	SI-DD	5 : 34	11	∗
DHN	:	SI-SN	7 : 27	16	∗
recorded	:	SD-DD	50 : 0	0	∗
recorded	:	SI-SDN	49 : 0	1	∗
recorded	:	SI-SDNC	48 : 0	2	∗
recorded	:	SI-DD	46 : 3	1	∗
recorded	:	SI-SN	49 : 0	1	∗
SD-DD	:	SI-SDN	10 : 15	25	
SD-DD	:	SI-SDNC	10 : 15	25	
SD-DD	:	SI-DD	10 : 19	21	
SD-DD	:	SI-SN	19 : 13	18	
SI-SDN	:	SI-SDNC	14 : 8	28	
SI-SDN	:	SI-DD	12 : 15	23	
SI-SDN	:	SI-SN	16 : 4	30	(∗)
SI-SDNC	:	SI-DD	13 : 15	22	
SI-SDNC	:	SI-SN	18 : 15	17	
SI-DD	:	SI-SN	10 : 14	26	

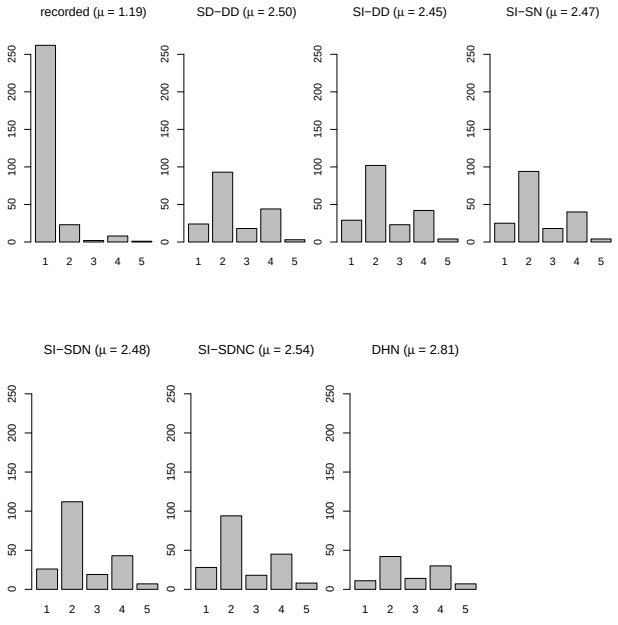

Figure 6: *Frequencies of similarity votes for each of the seven methods to the recorded reference sample as evaluated in the subjective listening test. 1 means very similar, 5 means very different.*

ods (Table 3). Concerning the adaptive methods this can be due to the small amount of speakers for some models, but for SI-SDN for example we had 10 pseudo-speakers in the average voice model. Furthermore improvements with adaptive modeling for Austrian German and Viennese were reported [5] with similar data sets.

Another reason could be that the phone overlap between the different varieties is not large enough for applying adaptive modeling directly. The larger phone overlap of 77% between Austrian German and Viennese supports this hypothesis.

For the varieties used for the full average voice, the phone set overlap was 26% (AT ∩ IVG ∩ GOI). For the variety pairs the phone set overlap was 33% (AT ∩ IVG, AT ∩ GOI) and 59% (GOI ∩ IVG). This suggests a pre-clustering of the data prior to training and adaptation, which is dependent on larger amounts of training data. It also shows that the distance between variety and standard in terms of phonetic overlap can be quite different for different varieties.

Concerning the overall similarity of synthesized samples to original ones we saw that we can achieve a satisfying modeling of overall similarity with all modeling methods except DHN (Figure 6). Assuming that overall similarity factors into variety similarity and speaker similarity, we can conclude that dialects and speakers can be modeled successfully.

Even if we see no significant differences between adaptive and speaker-dependent modeling with this data set, we would still favor the adaptive approach since it has shown its advantage in other experiments [5] and it does never decrease the quality (except for DHN). Furthermore the adaptive approach gives us additional possibilities for applications due to the common decision tree structure in the modeling of fast speech [15] or dialect interpolation [5] for example. The analysis of phone set overlaps points to a threshold that shows when it is possible to exploit the full potential of the adaptive approach.

6. Conclusion and future work

In this paper we have shown adaptive modeling methods for dialects. We have described our data selection and recording approach and have shown that speaker-dependent and adaptive approaches are able to model the overall similarity between synthetic and recorded speech. Although we found no significant differences between adaptive and speaker-dependent methods the adaptive approach is still beneficial for applications like fast speech and dialect interpolation. Furthermore we have built corpora for the Bad Goisern (GOI) and Innervillgraten (IVG) dialect and synthesized these dialects for the first time. These synthesizers can be applied in many fields like tourism and language learning. The corpora are an important step in dialect preservation.

In future work we want to include the (already available) data from female speakers and perform gender-dependent/independent modeling. Furthermore we want to investigate pre-clustering techniques that can be applied to small data sets.

It remains an open question how much overlap we need between varieties to fully exploit the adaptive approach ad how we should measure this overlap. As our GOI and IVG corpora have a larger phone set overlap with each other than either of them does with the AT corpus, building a combined average model of GOI and IVG could further assist the analysis. This model could then be compared with speaker-dependent models and dialect-dependent average models to further investigate the impact of phone set overlap on the final speech quality.

7. Acknowledgements

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

8. References

[1] M. H. Cohen, J. P. Giangola, J. Balogh, Voice User Interface Design, Addison-Wesley, 2004.

[2] R. Dall, C. Veaux, J. Yamagishi, S. King, "Analysis of speaker clustering strategies for HMM-based speech synthesis", INTERSPEECH 2012.

[3] C. Wutiwiwatchai, A. Thangthai, A. Chotimongkol, C. Hansakunbuntheung, N. Thatphithakkul, "Accent level adjustment in bilingual Thai-English text-to-speech synthesis", ASRU 2011, 295-299.

[4] M. Wester, R. Karhila, "Speaker similarity evaluation of foreign-accented speech synthesis using HMM-based speaker adaptation", ICASSP 2011, 5372-5375.

[5] M. Pucher, D. Schabus, J. Yamagishi, F. Neubarth, V. Strom, "Modeling and interpolation of Austrian German and Viennese dialect in HMM-based speech synthesis", Speech Communication, 52(2):164-179, 2010.

[6] M. Toman, M. Pucher, "Structural KLD for cross-variety speaker adaptation in HMM-based speech synthesis", in Proc. SPPRA, Innsbruck, Austria, 2013.

[7] Y.-J. Wu, Y. Nankaku, and K. Tokuda, "State mapping based method for cross-lingual speaker adaptation in HMM-based speech synthesis", in Proc. INTERSPEECH, Brighton, United Kingdom, 2009, pp. 528-531.

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

[9] M. Pucher, N. Kerschhofer-Puhalo, D. Schabus, S. Moosmüller, G. Hofer, "Language resources for the adaptive speech synthesis of dialects", Proc. of SIDG 2012, Vienna, Austria.

[10] H Scheutz, "Deutsche Dialekte des Alpenraums", 2009, `http://www.argealp.org/atlas/data/atlas.html`.

[11] H. Scheutz, S. Aitzetmüller, Peter Mauser, "Drent und herent. Dialekte im salzburgisch-bayerischen Grenzgebiet. Mit einem sprechenden Dialektatlas auf CD-ROM", EuRegio Salzburg, 2007.

[12] J. Yamagishi, T. Kobayashi, "Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training", IEICE Trans. Inf. & Syst., vol. E90-D, no. 2, pp. 533543, Feb. 2007.

[13] K. Shinoda, T. Watanabe, "Acoustic modeling based on the MDL principle for speech recognition", Eurospeech-97, 99-102.

[14] J. Yamagishi, H. Zen, Y.-J. Wu, T. Toda, K. Tokuda, "The HTS-2008 System: Yet Another Evaluation of the Speaker-Adaptive HMM-based Speech Synthesis System", Proc. Blizzard Challenge 2008.

[15] M. Pucher, D. Schabus, J. Yamagishi, "Synthesis of fast speech with interpolation of adapted HSMMs and its evaluation by blind and sighted listeners". INTERSPEECH 2010, Makuhari, Japan, pp. 2186-2189.