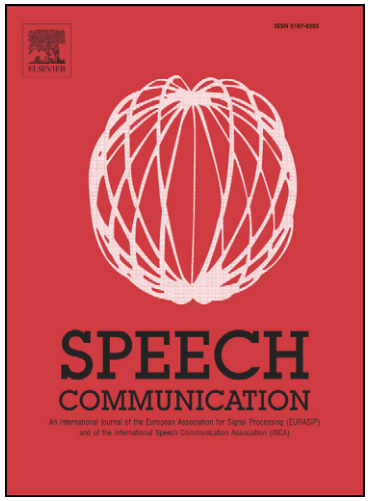
Modeling and interpolation of Austrian German and Viennese dialect in HMM-based speech synthesis

Accepted Manuscript



Modeling and Interpolation of Austrian German and Viennese Dialect in HMM-based Speech Synthesis

Michael Pucher, Dietmar Schabus, Junichi Yamagishi, Friedrich Neubarth, Volker Strom

PII: S0167-6393(09)00147-2
DOI: 10.1016/j.specom.2009.09.004
Reference: SPECOM 1836

To appear in: *Speech Communication*

Received Date: 5 May 2009
Revised Date: 22 September 2009
Accepted Date: 23 September 2009

Please cite this article as: Pucher, M., Schabus, D., Yamagishi, J., Neubarth, F., Strom, V., Modeling and Interpolation of Austrian German and Viennese Dialect in HMM-based Speech Synthesis, *Speech Communication* (2009), doi: 10.1016/j.specom.2009.09.004

Modeling and Interpolation of Austrian German and Viennese Dialect in HMM-based Speech Synthesis

Michael Pucher[a], Dietmar Schabus[a], Junichi Yamagishi[b], Friedrich Neubarth[c], Volker Strom[b]

[a] *Telecommunications Research Center Vienna (ftw.), Donau-City-Str 1, 3rd floor, 1220 Vienna, Austria*
[b] *The Centre for Speech Technology Research, University of Edinburgh, Informatics Forum, 10 Crichton Street, Edinburgh, EH8 9AB United Kingdom*
[c] *Austrian Research Institute for Artificial Intelligence (OFAI), Freyung 6/6, 1010 Vienna, Austria*

Abstract

An HMM-based speech synthesis framework is applied to both Standard Austrian German and a Viennese dialectal variety and several training strategies for multi-dialect modeling such as dialect clustering and dialect-adaptive training are investigated. For bridging the gap between processing on the level of HMMs and on the linguistic level, we add phonological transformations to the HMM interpolation and apply them to dialect interpolation. The crucial steps are to employ several formalized phonological rules between Austrian German and Viennese dialect as constraints for the HMM interpolation. We verify the effectiveness of this strategy in a number of perceptual evaluations. Since the HMM space used is not articulatory but acoustic space, there are some variations in evaluation results between the phonological rules. However, in general we obtained good evaluation results which show that listeners can perceive both continuous and categorical changes of dialect varieties by using phonological transformations employed as switching rules in the HMM interpolation.

Key words: speech synthesis, hidden Markov model, dialect, sociolect, Austrian German

1. Introduction

Statistical parametric speech synthesis based on hidden Markov models (HMMs) (Yoshimura *et al.*, 1999) has become established and well-studied, and has an ability to generate natural-sounding synthetic speech (Black *et al.*, 2007; Zen *et al.*, 2009). In recent years, the HMM-based speech synthesis systems have reached performance levels comparable to state-of-the-art unit selection systems (Fraser and King, 2007; Karaiskos *et al.*, 2008). In this method, acoustic features such as the spectrum, excitation parameters, and segment duration are modeled and generated simultaneously within a unified HMM framework. A significant advantage of this model-based parametric approach is that speech synthesis is far more flexible compared to conventional unit-selection methods, since many model adaptation and model interpolation methods can be used to control the model parameters and thus the characteristics of the generated speech (Yoshimura *et al.*, 2000; Yamagishi *et al.*, 2009a). In fact, these methods have already been applied to generating transitions between different speakers (Yoshimura *et al.*, 2000), different types of emotional speech, and different speaking styles (Tachibana *et al.*, 2005).

These techniques are also useful for achieving *varying* multi-dialect voices in text-to-speech (TTS) synthesis. They may be used for personalizing speech synthesis systems and have several potential benefits. For example, if the TTS system is used to provide an alternative voice output for patients who have progressive dysarthria (Creer *et al.*, 2009), some patients will desire a TTS system that has the same dialect as themselves.

However, it is not always feasible to prepare pronunciation dictionaries separately for every possible language variety in advance, since writing dictionaries is an extremely time-consuming and costly process. Often one variety is taken as a standard, and the linguistic resources such as pronunciation dictionaries are only available for this standard variety. Thus, to flexibly model as many varieties as possible, some acoustic and linguistic control based on this standard or typical dialect is required.

Although one might regard dialect control[1] as conceptually equivalent to the emotional control mentioned above, there is a significant difference in the requirements for the

Email address: pucher@ftw.at (Michael Pucher)

[1] In this paper we use the notion of 'dialect' in a broad sense as referring to non-standard language varieties. In the case at hand, it would be more accurate to speak of Viennese sociolect, since language varieties in Vienna are discerned by social criteria and not (or no longer) identified by association to a certain geographical region. We use the term 'dialect control' as shorthand for 'control of dialectal or sociolectal language variety'.

control of dialectal varieties. The speaker or emotional interpolation mentioned above implicitly assumes that the target models use the same pronunciation dictionary, and therefore phone strings, within the same language and linear interpolation is applied just to the relevant models, which results in acoustic transitions within the same phone or sub-word unit. For dialect control, we need to additionally consider linguistically-motivated transitions. In other words, we need to include not only the HMMs but also the pronunciation dictionary as targets of the interpolation process. That is, the HMMs to be interpolated may represent different phone sequences derived from different dictionaries. Moreover, these sequences may also consist of a different number of phones.

A major premise for dialect control is that dialects, as varieties of languages, form a "continuum" (Saussure, 1983): the varieties are related to one another in terms of being linguistically close, which makes it possible for us to hypothesize the existence of varieties on that continuum of fine-grained subtleties that lie between two different varieties already defined by linguistic resources. In addition to geographical transition of the dialect varieties, that is, regiolects, we may apply the same logic to other varieties of languages such as sociolects, which are categories of linguistic varieties defined by the social level of speakers.

The proposed dialect interpolation aims to produce synthetic speech in a phonetically intermediate variety from given models and dictionaries for adjacent typical varieties. For the phonetic control, we simply use linear interpolation of HMMs that represent the acoustic features similar to speaker or emotional interpolation. Since relations between articulatory and acoustic features are non-linear (Stevens, 1997), the phonetic control that can be achieved using acoustic features alone is noisy and might sometimes exhibit unexpected behavior. However it is worthwhile to investigate the basic performance of acoustic interpolation because proper acquisition of articulatory features requires specialized recording equipment such as electromagnetic articulography (EMA) (Schönle *et al.*, 1987) and also because phonetic knowledge such as vowel height or backness and place or manner of articulation can be used in clustering the acoustic HMMs via manually-defined linguistic questions.

A closer inspection of potential phonetic transitions between language varieties reveals several exceptional cases. From phonetic studies of Viennese dialects (Moosmüller, 1987) we know that some gradual transitions are well motivated (e.g., spirantization of intervocalic lenis plosives), while some other transitions between phones are strong markers for that specific variety, and thereby categorical. In the latter case, either the standard form of a given phone is produced, or its dialectal counterpart, with no possible in-between variants. One example of such a transition is the phone [aː] in the Standard Austrian German variety which is realized as [ɔː] in the Viennese dialect (mentioned in detail later in Table 5). For such a case, the use of interpolation (e.g., model interpolation between [aː] and [ɔː] phone HMMs) is not appropriate. For this reason, we introduce several knowledge-based switching rules that allow for overriding acoustic interpolation in such cases. Since it is known from psycholinguistics that continuous transitions between phones are often only perceived categorically (Liberman, 1970), the knowledge-based switching rules should improve the perception of dialects compared to acoustic interpolation alone. Hence, we include interpolations with and without switching rules in the subjective evaluation to measure the effect of the proposed dialect interpolation and switching rules.

In addition we investigate efficient clustering strategies for the dialect varieties in HMM-based speech synthesis. In general there are insufficient speech resources for non-standard dialect varieties. This situation might be even more severe for minor languages. Thus we compare several clustering algorithms for a practical case where the amount of speech data for dialects is limited, but there is sufficient speech data for the standard. We also include speech data from speakers that are able to speak standard and dialect.

This paper is organized as follows. Section 2 gives an overview of modeling strategies of HMM-based speech synthesizers for dialect varieties and an associated evaluation. In Section 3 we show how to generate speech that forms a continuous transition between one variety and another. The two varieties we are considering in this paper are Standard Austrian German and Viennese dialect. Apart from continuous interpolation of HMMs, we also define specific switching rules. We then present the results of a series of listening tests. Section 4 summarizes our findings and discusses remaining issues.

2. Acoustic modeling of dialect varieties in HMM-based speech synthesis

2.1. Overview of HMM-based TTS system

All TTS systems described here are built using the framework from the "HTS-2007/2008" system (Yamagishi *et al.*, 2009b, 2008), which was a speaker-adaptive system entered for the Blizzard Challenges in 2007 and 2008 (Karaiskos *et al.*, 2008). The HMM-based speech synthesis system, outlined in Fig. 1, comprises four main components: speech analysis, average voice training, speaker adaptation, and speech generation.

In the speech analysis part, three kinds of parameters for the STRAIGHT (Speech Transformation and Representation by Adaptive Interpolation of weiGHTed spectrogram) (Kawahara *et al.*, 1999) mel-cepstral vocoder (Tokuda *et al.*, 1991; Fukada *et al.*, 1992) with mixed excitation (Yoshimura *et al.*, 2001; Kawahara *et al.*, 2001) (i.e. a set including the mel-cepstrum, $\log F_0$, and band aperiodicity measures) are extracted as feature vectors for the HMMs. These features are described in (Zen *et al.*, 2007a). In the average voice training part, context-dependent multi-stream left-to-right multi-space distribution (MSD) hidden semi-Markov models (HSMMs) (Zen

Table 1: Phone sets used in the experiments, represented with IPA symbols. The coding for 'Austrian German' is in accordance with the phonetic analysis presented in (Muhr, 2007), the coding for 'Viennese dialect' reflects our own analysis. Phones in brackets indicate that these are not really members of the native set.

Category	Austrian German	Viennese dialect
vowel	a aː (ɐː) eː e̞ (e̞ː) iː i oː o̞ uː u yː y øː ø̞	a aː ɐ ɐː e eː ɛ ɛː i iː ɪ o oː u uː ʊ y yː øː œ œː
di-/monophthong/nasal	a͡ɛ a͡o o̞͡e (ãːː) (œ̃ː) (ɔ̃ː)	æː ɒː œː ɐ͡ɪ o͡ɪ u͡ɪ ãː ɐ̃ ɐ̃ː ĩː æ̃ː ɔ̃ː
r-vocalized	e̞ɐ e̞ːɐ iːɐ iɐ oːɐ o̞ɐ uːɐ uɐ yːɐ yɐ øːɐ ø̞ɐ	ɐɐ ɐːɐ e̞ɐ e̞ːɐ iɐ iːɐ o̞ɐ o̞ːɐ ʊɐ ʊːɐ (yːɐ) øːɐ
schwa	ə ɐ	ə ɐ
plosive	b d g p t k	b d g β ð ɣ p t k
fricative	f v s ʂ ʃ ʒ ç x h	f v s sː ʃ ç x h
liquid/nasal/glide	ʀ l m n ŋ j	ʀ l l̥ m m̥ n n̥ ŋ ŋ̊ j
silence/pause/glottis	*'sil' 'pau'* ʔ	*'sil' 'pau'* ʔ

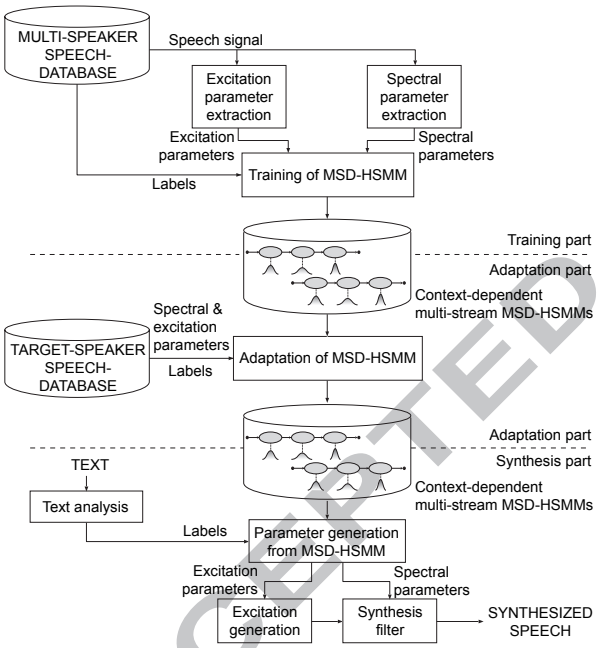


Figure 1: Overview of the HTS-2007 speech synthesis system, which consists of four main components: speech analysis and feature extraction, average voice training on multi-speaker speech database, speaker adaptation to a target speaker, and speech generation from the adapted models (Yamagishi *et al.*, 2009b, 2008).

et al., 2007b) are trained on multi-speaker databases in order to simultaneously model the acoustic features and duration. The phonetic and linguistic contexts we employ contain phonetic, segment-level, syllable-level, word-level, and utterance-level features as follows:

- preceding, current, and succeeding phones;
- acoustic and articulatory classes of preceding, current, and succeeding phones;
- the part of speech of the preceding, current, and succeeding words;
- the number of syllables in the preceding, current, and succeeding accentual phrases;
- the type of accent in the preceding, current, and succeeding accentual phrases;
- the position of the current syllable in the current accentual phrase;
- the number of accented syllables before and after the current syllable in the current phrase;
- the number of syllables in the preceding, current, and succeeding breath groups;
- the position of the current accentual phrase in the current breath group;
- the number of words and syllables in the sentence;
- the position of the breath group in the sentence;
- the specific language variety in the case of clustering of dialects (i.e. Viennese dialect or Standard Austrian German).

Phonesets used for Standard Austrian German and Viennese dialect are shown in Table 1. Austrian German and Viennese dialect have 58 and 75 phones, respectively. A set of model parameters (mean vectors and covariance matrices of Gaussian probability density functions (pdfs)) for the speaker-independent MSD-HSMMs is estimated using the feature-space speaker-adaptive training (SAT) algorithm (Anastasakos *et al.*, 1996; Gales, 1998). In the speaker adaptation part, the speaker-independent MSD-HSMMs are transformed by using constrained structural maximum *a posteriori* linear regression (Yamagishi *et al.*, 2009a).

In the speech generation part, acoustic feature parameters are generated from the adapted MSD-HSMMs using a parameter generation algorithm that considers both the global variance of a trajectory to be generated and trajectory likelihood (Toda and Tokuda, 2007). Finally, an excitation signal is generated using mixed excitation (pulse plus band-filtered noise components) and pitch-synchronous overlap and add (PSOLA) (Moulines

Table 2: Data sources used for training and adaptation of standard Austrian German (*AT*) and Viennese dialect (*VD*) HMM-based speech synthesis systems.

				Number of utterances	
Speaker	Gender	Age	Profession	AT utterances	VD utterances
HPO	M	≈ 60	actor	219	513
SPO	M	≈ 40	radio narrator	4440	95
FFE	M	≈ 40	engineer	295	–
BJE	M	≈ 50	actor	87	95
FWA	M	≈ 60	language teacher	87	95
CMI	M	≈ 35	singer	–	95

Table 3: Definitions of modeling approaches used. SD and SI refer speaker-dependent and speaker-independent modeling. DD, DI, DC, DN, and DM refer to dialect-dependent, dialect-independent, dialect clustering, dialect-adaptive training, and DC plus DN, respectively. × means negative and √ means positive for each factor.

			Data Dependency		Dialect	
Name	Target	# utt.	Speaker	Dialect	Clustering	Normalization
SD-DD (*AT*)	AT	219	√	√	×	×
SD-DD (*VD*)	VD	513	√	√	×	×
SD-DI	AT/VD	732	√	×	×	×
SD-DC	AT/VD	732	√	×	√	×
SD-DN	AT/VD	732	√	×	×	√
SD-DM	AT/VD	732	√	×	√	√
SI-DD (*AT*)	AT	5128	×	√	×	×
SI-DD (*VD*)	VD	892	×	√	×	×
SI-DI	AT/VD	6020	×	×	×	×
SI-DN	AT/VD	6020	×	×	×	√

and Charpentier, 1990). This signal is used to excite a mel-logarithmic spectrum approximation (MLSA) filter (Fukada *et al.*, 1992) corresponding to the STRAIGHT mel-cepstral coefficients and thus to generate the speech waveform.

2.2. Speech database for Austrian German and Viennese dialect

For training and adaptation of Austrian German and Viennese dialect voices, a set of speech data comprising utterances from 6 speakers was used. Table 2 shows details of the speakers and number of utterances recorded for each. Here *AT* stands for Standard Austrian German and *VD* for Viennese dialect. There are many differences between Standard Austrian German and Viennese dialect on many linguistic levels, which we have described previously in (Neubarth *et al.*, 2008). All speakers are male speakers, of which five are native speakers of the Viennese dialect. As we can see from this table, the data sets widely vary in terms of the number of utterances, and whether they contain speech data from standard, dialect, or both. This is simply because these speech data sources were recorded for different purposes: some were recorded for unit selection synthesis test voices (*BJE, CMI, FWA*), one data set was recorded for a small unit selection voice (*FFE*), one was recorded for a large unit selection voice (*SPO*), and one was recorded for the adaptation and interpolation experiments described here (*HPO*). Ideally we should use a well-balanced larger speech database having equal amounts of data from Standard Austrian German and Viennese dialect in terms of quantity and linguistic contexts mentioned in the previous section. However since such a well-balanced database is not available yet and there are always fewer resources for non-standard varieties, we explore the best modeling for both *AT* and *VD* from the available unbalanced database.

Our first goal was to evaluate which modeling approach works best to train Austrian German and Viennese voices for the speaker *HPO* since this speaker's data is phonetically balanced for both *AT* and *VD* and this enables the evaluation of several modeling strategies.

2.3. Modeling approaches

Table 3 defines the modeling approaches we used. SD and SI refer to speaker-dependent and speaker-independent modeling. Likewise we can consider dialect-dependent and dialect-independent modeling. For dialect-independent modeling, there are two possible approaches.

The first 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 HMMs. The second is to divide a set of speech data in both varieties uttered by one speaker into two subsets of speech data in different varieties uttered by two different pseudo speakers. In similar way to SAT estimation (Anastasakos *et al.*, 1996; Gales, 1998) where acoustic differences between speakers are normalized for better average voice model training, we can normalize acoustical differences between varieties and can train a more canonical dialect-independent model. We call this training procedure "dialect-adaptive training". DD, DI, DC and DN refer to dialect-dependent, dialect-independent, dialect clustering and dialect-adaptive training, respectively. DM refers to "DC plus DN". In the table, the first column gives a short name for each modeling method, the second column gives the target dialect of the adaptation, the third column gives the number of utterances available, the fourth and fifth columns show the dependency on speaker or dialect, in which $\times$ means negative and $\surd$ means positive for each factor, and the sixth and seventh columns show training with or without clustering of dialects and dialect-adaptive training.

In the clustering of dialects, a new question that distinguishes Viennese from Austrian German data is added to a set of questions for the decision-tree-based clustering (Young *et al.*, 1994) and minimum description length (MDL) based automatic node-splitting (Shinoda and Watanabe, 2000) is performed. Dialect 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 (Yamagishi *et al.*, 2008). In the clustering we observe that the question concerning the variety is used near the root of the decision trees. Figure 2 shows part of the constructed decision tree for the mel-cepstral parameters of the third state and the corresponding duration parameter clustering tree. "C-Vowel" means "Is the center phoneme a vowel?", "C-Fricative" means "Is the center phoneme a fricative?", "Is-Viennese-Dialect" means "Is the current utterance in Viennese dialect?", and so on. From this example, we can see that separate Gaussian pdfs for vowel and fricative models for the Viennese dialect are produced from those for Austrian German. We can also see that separate Gaussian pdfs are generated for Viennese vowel duration.

We applied model adaptation with *AT* and *VD* data to all models except the first two. The adaptation represents dialect adaptation in the SD-DI, SD-DC, SD-DN, and SD-DM systems. It represents speaker adaptation in the SI-DD (*AT* or *VD*) systems. It represents simultaneous adaptation of speaker and dialect in the SI-DI and SI-DN systems. Therefore we have 16 voices in total (8 Austrian German and 8 Viennese voices), where 14 are adapted voices and 2 are speaker- and dialect-dependent voices.

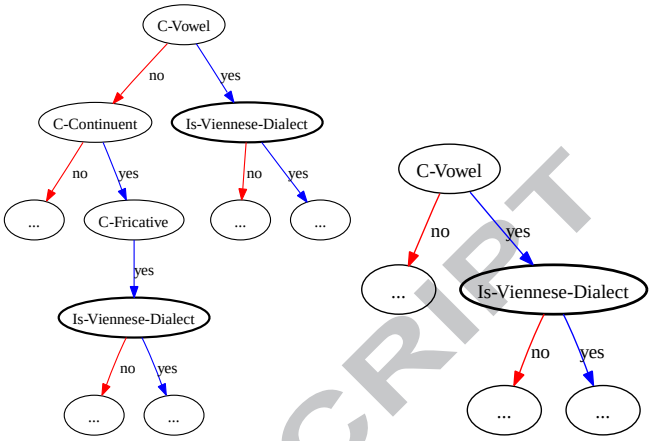


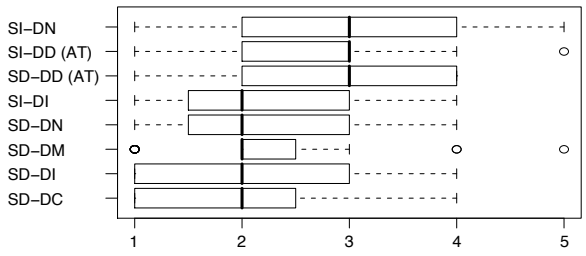
Figure 2: Dialect clustering results. The left figure shows a part of a decision tree built for mel-cepstral coefficients and the right figure shows a part of a decision tree built for state duration. Both are for SD-DM systems.

2.4. Experimental conditions

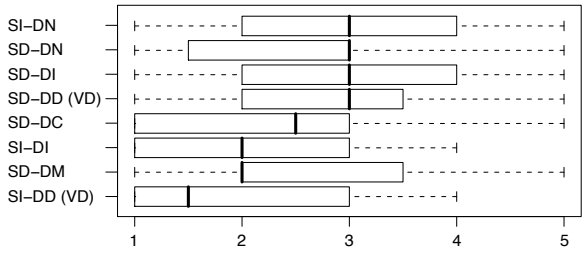
Speech signals were sampled at a rate of 16 kHz and windowed by an F_0-adaptive Gaussian window with a 5 ms shift. The feature vectors per frame consisted of 138-dimension vectors: 39-dimension STRAIGHT mel-cepstral coefficients (plus the zeroth coefficient), $\log F_0$, 5 band-filtered aperiodicity measures, and their dynamic and acceleration coefficients. We used 5-state left-to-right context-dependent multi-stream MSD-HSMMs without skip transitions. Each state had a single Gaussian pdf with a diagonal covariance matrix in each stream for continuous features (mel-cepstra and band-limited aperiodicity) and MSDs consisting of scalar Gaussian pdfs and discrete distributions in each stream for log F0 (Zen *et al.*, 2007b) as emission probabilities, and also a Gaussian pdf as a duration probability. For speaker adaptation, the transformation matrices were triblock diagonal corresponding to the static, dynamic, and acceleration coefficients.

2.5. Evaluation

In order to choose the best voice for each variety that is used in the interpolation experiments in Section 3.2, a listening evaluation was conducted with 40 subjects. The listening evaluation consisted of two parts: in the first part listeners were asked to judge the overall quality of synthetic speech utterances generated from several models using the different training strategies from Table 3. The evaluation method used a 5-point scale, where 5 means "very good" and 1 means "very bad". In the second part, after hearing a pair (in random order) of synthetic speech samples generated from the models, the listeners were asked which synthetic speech sample they preferred. The same

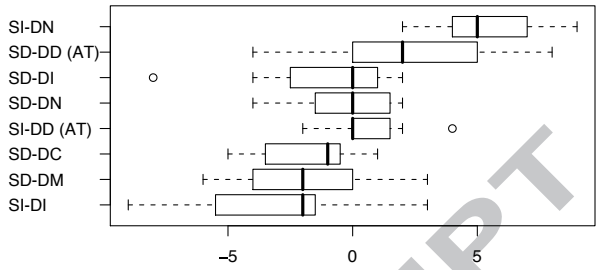


(a) Austrian German voices

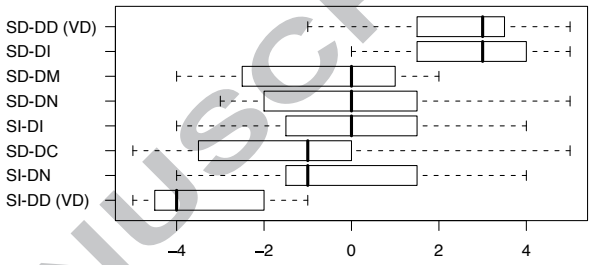


(b) Viennese dialect voices

Figure 3: Box-plots for 5-point scale evaluation for the overall quality for the AT (a) and VD (b) varieties. 5 means "very good" and 1 means "very bad".



(a) Austrian German voices



(b) Viennese dialect voices

Figure 4: Box-plots of pairwise comparison score for the AT (a) and VD (b) varieties. The data for one voice i comprise seven scores $s_j = w_{ij} - l_{ij}$, where $j \neq i$ and w_{ij} and l_{ij} are the numbers of comparisons won and lost, respectively, of voice i against voice j.

synthetic speech utterances were used for both the evaluation tests. A Mann-Whitney-Wilcoxon test was used to identify significant differences.

Figure 3 shows the results of the first part of the evaluation. For *AT*, there are three voices that are significantly better than other voices ($p < 0.05$), namely SI-DN, SI-DD (*AT*), and SD-DD (*AT*). For *VD*, the evaluation results for overall quality are less clear than those for the *AT* voices. Here we have only a clear loser with SI-DD, which is significantly worse than most other voices ($p < 0.05$) because of the low performance of the average voice model that was trained on a limited amount of *VD* speech data only. These results are simply due to the amounts of data used for training the HMMs, rather than linguistic issues. In general, training of average voice models requires $O(10^3)$ utterances (Yamagishi *et al.*, 2009a) but SI-DD (*VD*) has only about 900 utterances.

Figure 4 shows the evaluation results for the pairwise comparisons of *AT* and *VD* voices. For the *AT* voices, the SI-DN voice is significantly better than all others except SD-DD (*AT*) ($p < 0.05$). However the speaker- and dialect-dependent SD-DD (*AT*) voice is significantly better than only two other voices; thus, the SI-DN voice may be considered the best. This is an interesting result: although we have enough *AT* speech data (particularly compared to *VD* speech data), the simultaneous use of both *AT* and *VD* speech data leads to better performance. This good performance of the adapted models is consistent with previous results (Yamagishi *et al.*, 2009b). Furthermore we can see that the best training strategy is to divide utterances by a single speaker into standard (*AT*) and dialect (*VD*) utterances and treat them as two speakers in the SAT process, which is done in the SI-DN voice.

For *VD* there are two methods, namely SD-DD (*VD*) and SD-DI, that are significantly better than three other methods ($p < 0.05$). Since the speaker HPO has a relatively large amount of *VD* speech data but the amount of *VD* speech data from other speakers is very small, speaker-independent models do not perform well for *VD*.

From these results we chose SI-DN and SD-DD (*VD*) systems for the *AT* and *VD* voices, respectively. The *mixed variety* modeling approach is unfortunately not very successful, although we did observe some intuitively reasonable classes emerging from the clustering, such as a separate vowel cluster for the Viennese dialect. We believe that these problems are due to the limited amount of training data. We plan to repeat the experiments for the mixed dialect modeling when we have more balanced speech data available.

3. Dialect interpolation for HMM-based speech synthesis

In this section we add new phonological aspects to the model interpolation techniques for HMM-based speech synthesis, then apply this to dialect interpolation based on the concept of a dialect continuum. Specifically, we consider phonological rules which transform the standard variety to another variety. The rules between varieties

Table 4: *Minor shifts between Austrian standard and Viennese dialect.*

Phonological process	AT orthographic	gloss	AT IPA	VD IPA
tense vowels	**Bett**, offen	*bed, open*	bɐ̯t, ɒ̯fən	bet, ofm̩
monophthongs	**Deutsch**	*German*	dɒ̯͡ɐ̯etʃ	dæːtʃ
spirantization	**Leber**, sorgen	*liver, worry*	leːbɐ, sɔ̯ɐgən	leːβɐ, sʊɐɣn̩

Table 5: *Phonologically-manifested differences of the Viennese dialect.*

Phonological process	AT orthographic	gloss	AT IPA	VD IPA
input shift	**Schlag**, lieb	*cream, nice*	ʃlaːk, liːp	ʃlo̯ːk, lɪɐp
l-vocalization-1	viele, **Keller**	*many, basement*	fiːlə, kɐ̯elɐ	fyːlə, kœlɐ

Table 6: *Differences affecting the segmental structure.*

Phonological process	AT orthographic	gloss	AT IPA	VD IPA
l-vocalization-2	**Holz**, **Milch**	*wood, milk*	hɔ̯ɐ̯lts, milç	hɔ̯͡ɪts, myːç
schwa-deletion	**Hände**, liege	*hands, lie*	hɐ̯ndə, liːgə	hent, liːk
	Gewicht	*weight*	gəviçt	gviçt

Table 7: *Applying processes selectively for the German word "Gefahr" ('danger')*

	AT IPA	*Process*		VD IPA
from AT to VD		*schwa deletion:*	[kfaː]	
	[gəfaː]			[kfo̯ːɐ]
from VD to AT		*input shift /aː/*	[gəfo̯ːɐ]	

determine which target phones are to be interpolated and the interpolation modes. In-between variants are thus generated using HMM interpolation under phonological constraints.

Differences between several typical English dialects are well-researched and well-formalized (e.g., (Fitt and Isard, 1999)). Certain differences between the standard variety of Austrian German and the Viennese dialect can also be formalized in phonological terms. Note that we are not concerned about differences on higher linguistic levels such as morphology – these have to be dealt with by generating different inputs and no direct comparison may be applied to them. We will first give an overview of the formalized phonological processes between the standard variety of Austrian German and the Viennese dialect.

3.1. Phonological processes between the standard variety of Austrian German and the Viennese dialect

The phonological differences between the language varieties under consideration can be classified according to formal criteria that also have a significant impact on the way one can interpolate between the models associated with different phones or phone strings (cf. (Moosmüller, 1987; Neubarth *et al.*, 2008)):

1. **Minor shifts between Austrian standard and Viennese dialect** that are phonetically close and where these shifts are also observable in real life when people use different registers between the standard and some dialect variety (Table 4).
2. **Phonologically-manifested differences of the Viennese dialect** that are attributed to an 'input switch' between standard and dialect or differences that involve different phonological processes (Table 5).
3. **Differences affecting the segmental structure** by deleting or inserting phones from or into the phone string (Table 6).

The first set of differences involve vowels that only have a tense (closed, non-lowered) realization in the dialect variety, *monophthongization*, and the *spirantization* of intervocalic lenis plosives. The examples in Table 4 exemplify these processes. Crucially, the differences between the respective phones are gradual in a phonetic sense (Moosmüller, 1987). To model this group of processes and the transition between Austrian German and Viennese dialect, only an interpolation between phone models is necessary. Additionally, there are further common phonetically-motivated processes across word boundaries (hence post-lexical), which we did not consider in our experiments (assimilation of homorganic vowels, absorption of homorganic plosives, simplification of consonant clusters) (Moosmüller, 1987).

The second group of differences involve either different vowels (diachronically these phones have a different input base, so the notion *input shift* applies here), or different phonological processes apply to the input, while the segmental structure remains the same (Table 5). The term *l-vocalization(-1)* may be a little misleading here since the

phone /l/ is not vocalized itself, but rather remains unchanged as an onset. However, it still spreads the feature [round] onto the preceding vocalic segment, and there are good reasons to view it as akin to the second version of *l-vocalization* (see below). Since the segmental structure is the same, it is unproblematic to apply a gradual interpolation between the relevant models – at least in a technical sense. For this kind of difference one normally does not find intermediate stages of a gradual shift in real life; rather, these differences are used to signal the use of a different (dialect) variety of a language. They are taken to be strong dialect markers. Depending on their presence or absence in an utterance or word it is perceived as dialect or not (Moosmüller, 1987).

The third group of differences shown in Table 6 poses a more difficult technical challenge, since the segmental structure changes. Most prominently these are instances of *l-vocalization* in non-onset position, where the phone /l/ forms a secondary rising diphthong with the preceding (round) vowel or is not realized at all, and various instances of *schwa-deletion*.

These groups of phonological processes may be applied in a combined fashion in order to achieve more complex phonological transitions between standard and dialects. Table 7: *schwa-deletion* can be applied from the standard AT variety in order to indicate a slight approximation to the dialect without committing the speaker to strong dialect markers. Input shift for the vowel /a/ is always a strong dialect marker, but leaving the *schwa* pronounced indicates an approximation in the opposite direction, namely from dialect towards the standard. With this method it becomes possible also to model the direction of approximation between standard and dialect. In other words, it is possible to model a speaker of a certain variety who intends to speak another variety without fully committing him/herself to this variety.

3.2. Phonological constraints for HMM interpolation

For the first group mentioned in the previous subsection, we can straightforwardly apply HMM interpolation since they have the same number of phones in Austrian and Viennese. A good example is the distinction between di- and monophthongs in the Austrian Standard vs. Viennese dialect.

(1) **AT** d o͡e t ʃ
VD d æː t ʃ

For simplification of HMM-level processing we assumed all phone HMMs have the same number of states and applied state-level interpolation in these experiments. Duration models for HSMM can also be interpolated. If the models had a different number of states, we would need to perform a state alignment between the two phone HMM sequences, based on some criterion (e.g., Kullback-Leibler divergence).

For the second group, which does not have in-between variants, we utilize simple switching rules which disable

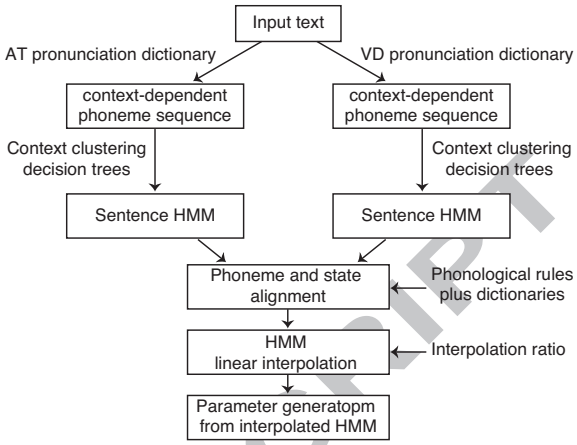


Figure 5: Flow of dialect interpolation

the HMM interpolation and switch the target phone for one variety to the other variety at some intermediate point (threshold). When such a threshold is given for the current phone, and the interpolation ratio for the utterance is below it, this phone is not interpolated, but rather the lower extreme point is used, as if the interpolation ratio were 0.0. If the interpolation ratio exceeds the threshold, the other extreme point (1.0) is used. Note that this is done phone by phone, so for neighboring phones it is possible that one is interpolated and the other is not.

This means that we can turn on or off the processes at a different point in the shifting continuum. Although we simply set this threshold to 0.5 in all our experiments, one could adjust this point for each phone individually.

For the third group (having words consisting of different numbers of phones in standard and dialect versions), we introduce a null phone [], which simply corresponds to a phone model with zero duration. Then, only the target phone's duration model is interpolated with the zero duration model.

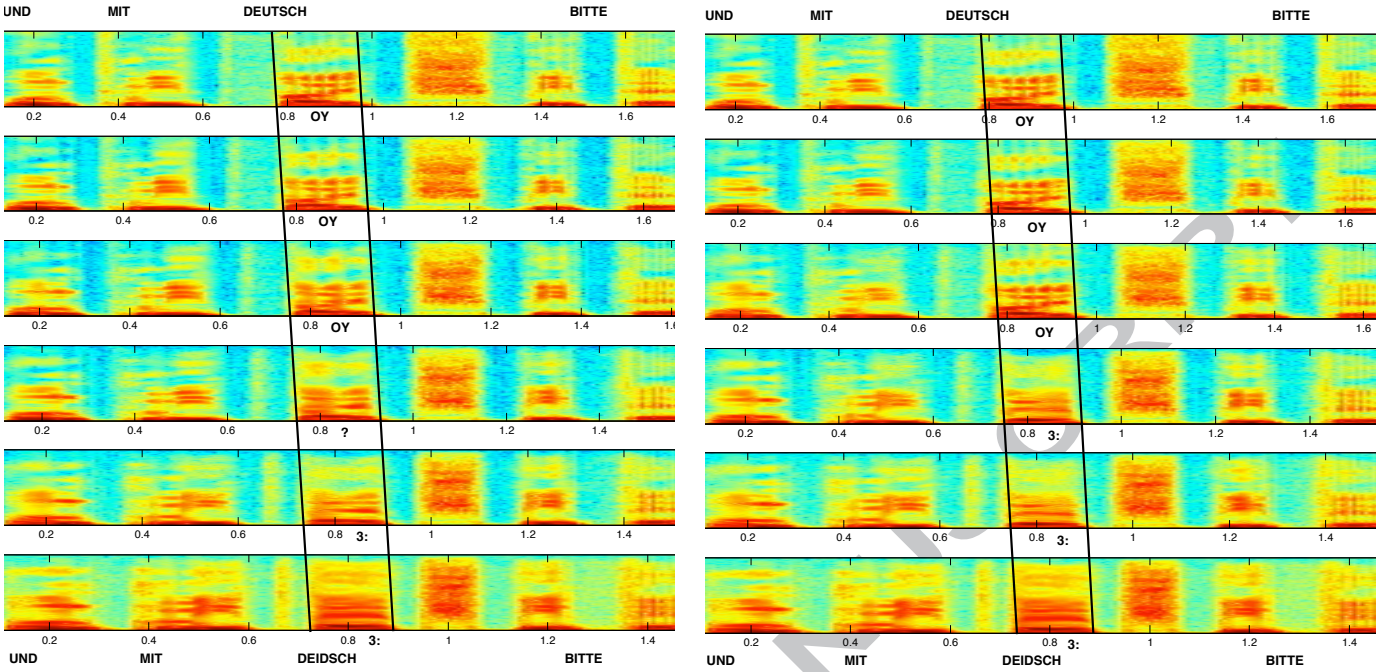
(2) **AT** g ə v i ç t
VD g [] v i ç t

The above example (2) shows the alignment for the phonological process of *schwa-deletion* (Table 6) where the missing ə is aligned to the null duration model [].

Although these three groups and their combinations are not enough to automatically and completely reproduce the VD variety from the standard AT variety in TTS systems, we believe that they are sufficient to answer our scientific questions and to form a basis for our next large-scale experiments.

3.3. HMM linear interpolation and its underlying issues

From the above examples it should be clear that we cannot perform offline interpolation on the level of HMMs, since the same phone HMM may have several interpolation modes depending on what kinds of word the phone HMMs belong to and what kinds of phonological groups



(a) without switching rules

(b) with switching rules

Figure 6: An interpolation example between Austrian German "Und mit Deutsch bitte" (*And with German please*) and Viennese "Und mit Deidsch bitte". Interpolation ratio between them increments from 0.0 to 1.0 in steps of 0.2.

the word belongs to. Hence the interpolation of HMMs must be done on-line at synthesis time. We have therefore chosen *interpolation between observations* for the HMM interpolation, which was also used in (Tachibana *et al.*, 2005) and is the simplest interpolation method described in (Yoshimura *et al.*, 2000).

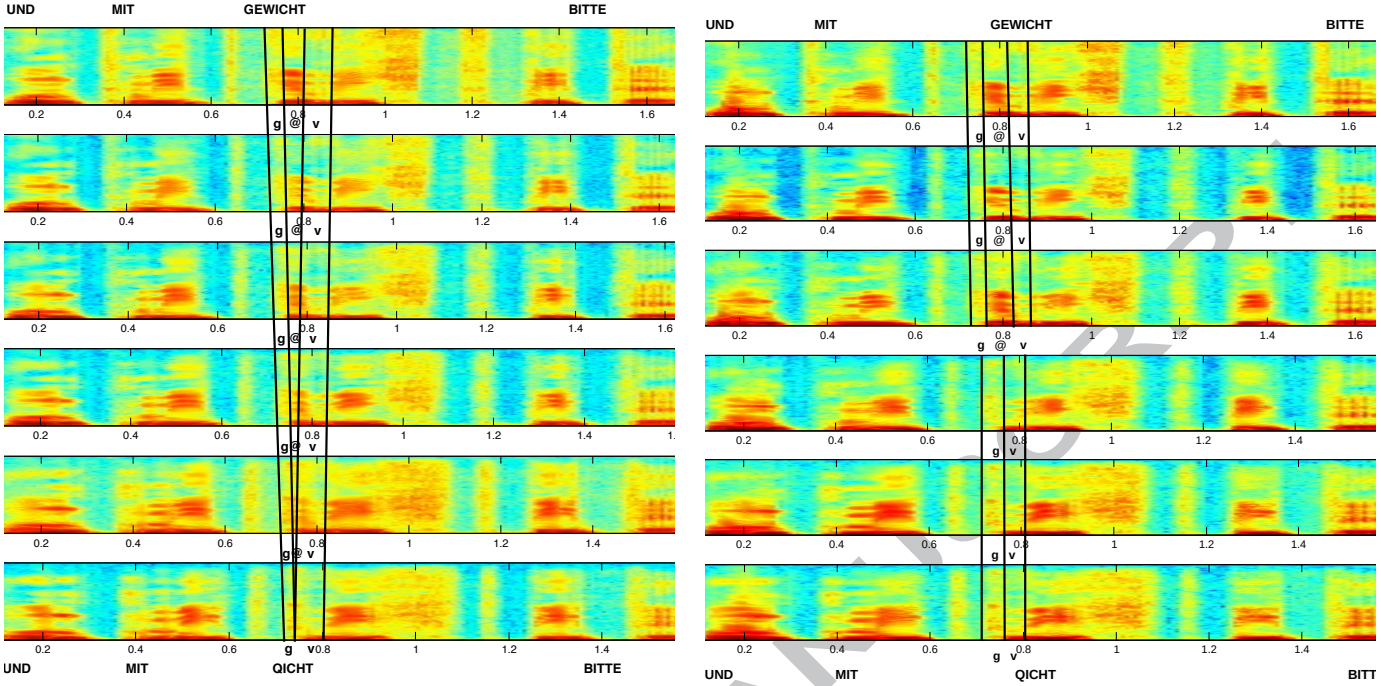
Figure 5 shows the overall procedure flow for dialect interpolation. First we convert a given text into two context-dependent phoneme label sequences based on AT and VD pronunciation dictionaries. Then by consulting the context clustering decision trees built for each state of each feature in the HMMs for AT and VD voices separately, the context-dependent phoneme label sequences are converted into two sentence HMMs having different state sequences. Each state has several Gaussian pdfs for each of the acoustic features and a single Gaussian pdf for its duration. A Gaussian pdf for state i is characterized by a mean vector $\boldsymbol{\mu}_i$ and a covariance matrix $\boldsymbol{\Sigma}_i$. The dimension of the mean vector may vary depending on the acoustic features. Then, based on the pronunciation dictionaries and phonological rules adopted, the two state sequences are aligned and linear interpolation between the sequences is applied. Let $\boldsymbol{\mu}_i^{AT}$ and $\boldsymbol{\mu}_i^{VD}$ be mean vectors of Gaussian pdfs for AT and VD voices, respectively, at aligned state i. Likewise $\boldsymbol{\Sigma}_i^{AT}$ and $\boldsymbol{\Sigma}_i^{VD}$ are their covariance matrices. In the interpolation above (Yoshimura *et al.*, 2000), the interpolated mean vector $\hat{\boldsymbol{\mu}}$ and covariance matrix $\hat{\boldsymbol{\Sigma}}$ at state i are calculated as follows:

$$\hat{\boldsymbol{\mu}}_i = w\boldsymbol{\mu}_i^{AT} + (1-w)\boldsymbol{\mu}_i^{VD} \tag{3}$$

$$\hat{\boldsymbol{\Sigma}}_i = w^2\boldsymbol{\Sigma}_i^{AT} + (1-w)^2\boldsymbol{\Sigma}_i^{VD} \tag{4}$$

where w is an interpolation ratio between AT and VD voices. After all the Gaussian pdfs for all the acoustic features and their duration are interpolated in a similar way, an optimal acoustic trajectory is generated from the interpolated HMM sequence.

One obvious issue is that the HMMs represent acoustic features rather than articulatory features. Since the relationship between articulatory and acoustic features is non-linear (Stevens, 1997), it would be preferable to use articulator positions for the phonetic transition. In fact one of the authors and colleagues have already proposed "articulatory-controllable" HMM-based speech synthesis (Ling *et al.*, 2008, 2009) based on this motivation. This would require the use of articulator positions; the current approach using only acoustic features is an approximation to this. Therefore it is expected that the current approach introduces some noise into the interpolation and may exhibit unexpected behavior from time to time. On the other hand, we emphasize that it is still worthwhile investigating the performance of such an acoustic interpolation, since proper acquisition of articulator positions requires specialized recording equipment. It is much easier to introduce phonetic knowledge such as vowel height or frontness and place or manner of articulation when clustering the acoustic HMMs via manually-defined linguistic questions.



(a) without switching rules

(b) with switching rules

Figure 7: An interpolation example between Austrian German "Und mit Gewicht bitte" (*And with weight please*) and Viennese "Und mit Qicht bitte" having different segmental structures. Interpolation ratio between them increments from 0.0 to 1.0 in steps of 0.2.

The success of the interpolation in the third group will also depend on whether the segment is the only vocalic portion of the syllable nucleus (as in the example *schwa-deletion* case above) or not. If it is the sole vocalic portion, intermediate stages may sound artificial because the vowel duration approaches zero and is thus too short to establish a phonetically-acceptable nucleus.

3.4. Interpolated examples

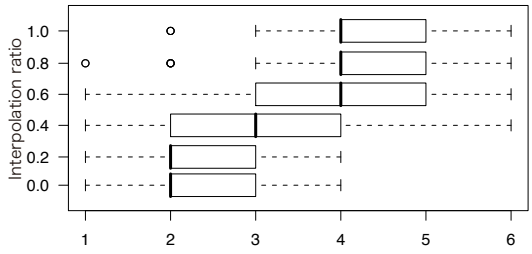
Figure 6 shows spectrograms of synthetic speech interpolated between the *AT* variety (top) and the *VD* variety (bottom) in interpolation ratio increments of 0.2. In Figure 6 (a) only the HMM linear interpolation was used, whereas in Figure 6 (b) a combination of the HMM interpolation and switching rules was applied. These samples can be downloaded from *http://dialect-tts.ftw.at*. In Figure 6 (a) we can see the continuous transformation from /OY/ [ɶ͡e] to /3:/ [æː]. Interestingly, while categorizing the sample utterances by experts, one intermediate stage was always classified as "undefined". This must be due to the nonlinear relation between articulatory and acoustic features. In the other setting (Figure 6 (b)) a switching rule governs the application of either model for the relevant phone. The upper three spectrograms were generated with a model from Austrian Standard /OY/, the other lines with a model from Viennese dialect /3:/. The remaining parts of the utterance are interpolated linearly. This results in appropriate categorical transitions of phones. Figure 7 shows the spectrogram of synthetic speech for the *schwa-deletion* case with and without switching rules. One can immediately see how the /@/ [ə] gradually disappears in Figure 7 (a). All the intermediate stages except for the penultimate one are judged as sounding natural. In the one exception, the duration of the /@/ segment is too short to be either classified as completely missing or present. In Figure 7 (b) we can see the categorical transition of *schwa-deletion* with switching rules, which delete /@/. Gradual changes are possible for a set of phonological processes like *monophtongization* or *input shifts*, but they produce gaps in the acoustic perception with other processes like *schwa-deletion*. Additional samples are shown in Appendix A.

3.5. Evaluation

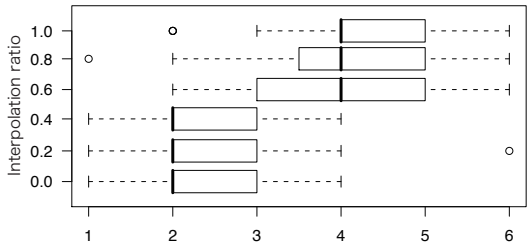
We designed a carrier sentence "Und mit ... bitte" (*And with ... please*) whose slot was filled with the words shown in bold in Tables 4–6. Each word represents a different process, with the exception of *l*-vocalization-1 and *s*chwa-deletion which are used twice. The phonetic transcription of the carrier sentence is provided in Example 5. This sentence has virtually no differences in different dialects.

(5) **AT/VD** ʔ u n t m i t ... b i t ə

For this evaluation we again used 40 listeners that had to answer two different questions after listening to synthesized interpolated prompts. In the first type of question, listeners were asked to give a rating as to what extent they would associate a given prompt with Viennese dialect or with standard Austrian German. For the rating,



(a) HMM interpolation only



(b) HMM interpolation plus switching rules

Figure 8: Box-plot for all utterances. Interpolation (vertical axis) ranges from 0.0 to 1.0 with or without switching rule. On the horizontal axis, 1 means strongly *VD*, 6 means strongly *AT*.

we used a scale from 1 (*'strongly Viennese'*) to 6 (*'strongly standard'*). Intermediate values were labelled *'Viennese'*, *'rather Viennese'* etc. In the second type of question, listeners were presented with two prompts and they were asked to judge how similar or different these were with respect to the differentiation between the dialect varieties. The first type of question is an *identification* task, the second type a *discrimination* task (Garman, 1990). The same Mann-Whitney-Wilcoxon test was used for finding significant differences.

Figure 8 shows the overall results for the identification task. In the figure, a ratio of 0.0 corresponds to the *VD* non-interpolated speech samples and 1.0 corresponds to the *AT* non-interpolated speech samples. The interpolation ratio between them increments in step of 0.2; figure (a) shows results without switching rules and figure (b) shows results with switching rules applied to the phonological process. Overall we can see that a gradual change was perceived for the interpolations without switching rule and a categorical change was perceived with the interpolations that applied a switching rule relatively. The gradual change is underpinned by the significant differences between 0.2 and 0.4, between 0.4 and 0.6, and between 0.6 and 0.8. The categorical change due to the switching rules is supported by the fact that there is a significant difference only between 0.4 and 0.6 ($p < 0.05$), and no significant differences between 0.2 and 0.4 or 0.6 and 0.8.

Figure 9 shows the result of the discrimination task (pairwise comparison), visualized using multi-dimensional scaling (MDS) (Cox and Cox, 2001). From this figure, we can confirm several findings from the identification task. HMM interpolation generates continuous transitions: the first dimension found by MDS (horizontal axis) corresponds to this. Adding the switching rule causes this continuous transition to become categorical: 0.0, 0.2, and 0.4 are clustered at the left side and 0.6, 0.8 and 1.0 are clustered at the right side. There is a wide gap between 0.4 and 0.6 when the switching rules are applied. In fact, since the switch threshold was set to 0.5, the switching rule is applied between 0.4 and 0.6. The second dimension found by MDS (vertical axis) is related to the switching rules. Distances between switched and non-switched interpolations are represented by this dimension. Interpolated samples using a ratio of 0.6 with and without the switching rules are far apart: these samples were judged by the listeners to sound different. This is consistent with our earlier funding that experts always classified one intermediate stage as an undefined phoneme.

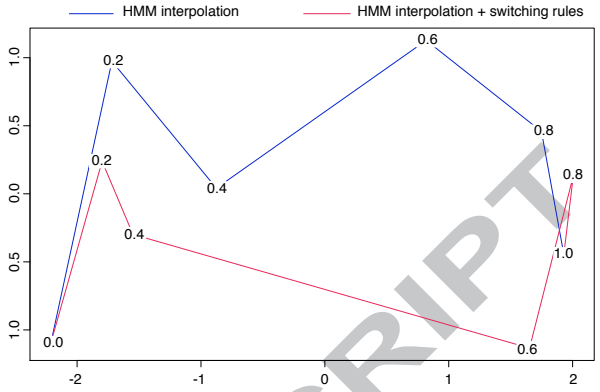
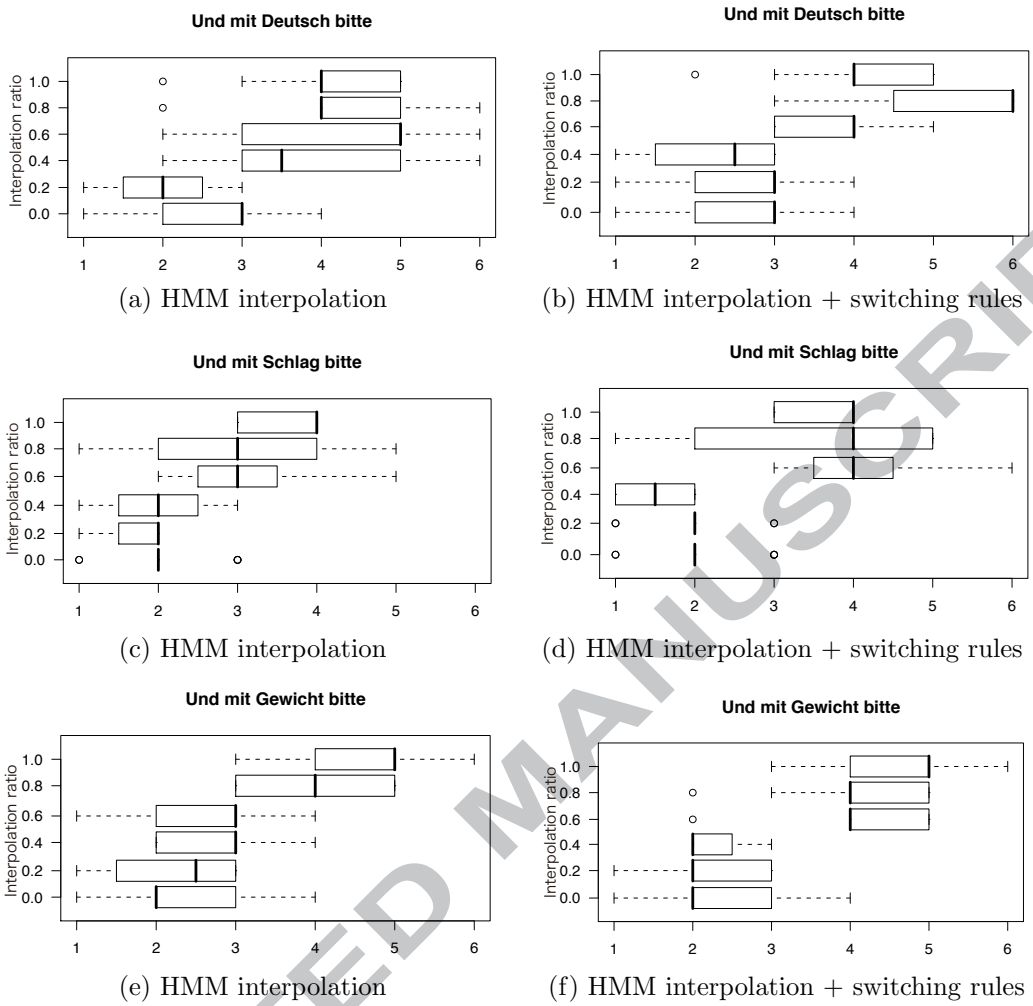


Figure 9: Evaluation of similarity in terms of dialect. Multi-dimensional scaling is used for 2D visualization of evaluation results.

Figure 10 shows the "Viennese-ness" ratings for three selected phonological processes, *monophthongization*, *input shift*, and *schwa-deletion* chosen from the three groups in Tables 4,5 and 6. We can clearly see the different behavior of these processes as dialect markers. The *monophthongization* process generates a relatively continuous transition between standard and dialect from both conditions. The *input shift* process generates a continuum between standard and dialect from the HMM interpolation, which does not match real phenomena, and generates a categorical shift with the switching rules. The *schwa-deletion* process creates a categorical shift at a certain point regardless of the use of switching rules. This means that a categorical change is perceived even if there is a continuous interpolation of the signal (Liberman, 1970).

4. Discussion and conclusion

The HMM-based speech synthesis framework has been applied to Austrian German and Viennese dialect. We have investigated and evaluated several training strategies for multi-dialect modeling such as dialect clustering and dialect-adaptive training. Although the speech database was unbalanced in terms of the amount of Austrian German and Viennese dialect speech data, such a situation



(a) HMM interpolation (b) HMM interpolation + switching rules

(c) HMM interpolation (d) HMM interpolation + switching rules

(e) HMM interpolation (f) HMM interpolation + switching rules

Figure 10: Box-plots for three different utterances chosen from three categories. Interpolation (vertical axis) ranges from 0.0 to 1.0 with or without switching rule. On the horizontal axis, 1 means strongly *VD*, 6 means strongly *AT*.

frequently occurs for non-standard varieties and so our results will apply to other dialects. For the *AT* variety, average voice models using dialect-adaptive training (where speech data uttered by a single speaker is divided into standard and dialect speaker data sets, and they are treated as different 'speakers' in the SAT process) achieve the best quality of synthetic speech. For the *VD* variety, speaker- and dialect-dependent modeling achieves the best quality. Although there was sufficient *AT* speech data, it did not help to improve the quality of the *VD* voice. We presume this is due to the linguistic differences between the AT and VD varieties.

In addition, we have bridged the gap between HMM-level processes and linguistic-level processes, by adding phonological processes to the HMM interpolation and applying it to dialect interpolation. We employed several formalized phonological rules between Austrian German and Viennese dialect as constraints for the HMM interpolation and verified their effectiveness in a number of perceptual evaluations. Since the HMM space used is not articulatory but simply acoustic, there are some variations in the effectiveness of each of the phonological rules. However, in general we obtained good evaluation results, which demonstrate that listeners can perceive both continuous and categorical changes of dialect variety in speech synthesised using phonological processes with switching rules in the HMM interpolation.

Our analysis results are obtained from relatively small-scale experiments designed to answer our scientific questions and to form a basis for our future large scale experiments. For large scale experiments on automatic dialect interpolation, we need to identify and employ additional phonological rules for each dialect. More sophisticated models that use articulatory features may also bring improvements, especially for consonant transformation.

Our future work will also focus on an interpolation method that applies switching rules hierarchically to introduce the notion of direction into our modeling. Furthermore we wish to extend the interpolation strategy from the approach that uses a null phone to more sophisticated modeling approaches that use a distance metric on HMMs and dynamic programming to align sequences of models.

Acknowledgements

The project "Viennese Sociolect and Dialect Synthesis" is funded by the Vienna Science and Technology Fund (WWTF). The Telecommunications Research Center Vienna (FTW) is supported by the Austrian Government and the City of Vienna within the competence center program COMET. OFAI is supported by the Austrian Federal Ministry for Transport, Innovation, and Technology and by the Austrian Federal Ministry for Science and Research. Junichi Yamagishi is funded by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement 213845 (the EMIME project). We thank Dr. Simon King and Mr. Oliver Watts of the University of Edinburgh for their valuable comments and proofreading. We also thank the reviewers for their valuable suggestions.

References

Anastasakos, T., McDonough, J., Schwartz, R., and Makhoul, J. (1996). A compact model for speaker-adaptive training. In *Proc. ICSLP-96*, pages 1137–1140.

Black, A., Zen, H., and Tokuda, K. (2007). Statistical parametric speech synthesis. In *Proc. ICASSP 2007*, pages 1229–1232.

Cox, T. and Cox, M. (2001). *Multidimensional Scaling*. Chapman and Hall.

Creer, S., Green, P., and Cunningham, S. (2009). Voice banking. *Advances in clinical neuroscience & rehabilitation*, **9**(2), 16 – 18.

Fitt, S. and Isard, S. (1999). Synthesis of regional English using a keyword lexicon. In *Proc. Eurospeech 1999*, volume 2, pages 823–826.

Fraser, M. and King, S. (2007). The Blizzard Challenge 2007. In *Proc. Blizzard 2007 (in Proc. Sixth ISCA Workshop on Speech Synthesis)*, Bonn, Germany.

Fukada, T., Tokuda, K., Kobayashi, T., and Imai, S. (1992). An adaptive algorithm for mel-cepstral analysis of speech. In *Proc. ICASSP-92*, pages 137–140.

Gales, M. (1998). Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech and Language*, **12**(2), 75–98.

Garman, M. (1990). *Psycholinguistics*. Cambridge University Press.

Karaiskos, V., King, S., Clark, R. A. J., and Mayo, C. (2008). The Blizzard challenge 2008. In *Proc. Blizzard Challenge Workshop*, Brisbane, Australia.

Kawahara, H., Masuda-Katsuse, I., and Cheveigné, A. (1999). Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds. *Speech Communication*, **27**, 187–207.

Kawahara, H., Estill, J., and Fujimura, O. (2001). Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT. In *2nd MAVEBA*.

Liberman, A. M. (1970). Some characteristics of perception in the speech mode. *Perception and its disorders*, **XLVIII**(11).

Ling, Z.-H., Richmond, K., Yamagishi, J., and Wang, R.-H. (2008). Articulatory control of HMM-based parametric speech synthesis driven by phonetic knowledge. In *Proc. Interspeech*, pages 573–576, Brisbane, Australia.

Ling, Z.-H., Richmond, K., Yamagishi, J., and Wang, R.-H. (2009). Integrating articulatory features into HMM-based parametric speech synthesis. *IEEE Trans. Speech, Audio & Language Process.*, **17**(6), 1171–1185.

Moosmüller, S. (1987). *Soziophonologische Variation im gegenwärtigen Wiener Deutsch*. Franz Steiner Verlag, Stuttgart.

Moulines, E. and Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, **9**(5-6), 453–468.

Muhr, R. (2007). *Österreichisches Aussprachewörterbuch Österreichische Aussprachedatenbank*. Peter Lang Verlag, Frankfurt.

Neubarth, F., Pucher, M., and Kranzler, C. (2008). Modeling Austrian dialect varieties for TTS. In *Proceedings of the 9th Annual Conference of the International Speech Communication Association (INTERSPEECH 2008)*, pages 1877–1880, Brisbane, Australia.

Saussure, F. D. (1983). *Course in General Linguistics*. Duckworth, London. (Original work published 1916).

Schönle, P. W., Gräbe, K., Wenig, P., Höhne, J., Schrader, J., and Conrad, B. (1987). Electromagnetic articulography: Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract. *Brain Lang.*, **31**, 26–35.

Shinoda, K. and Watanabe, T. (2000). MDL-based context-dependent subword modeling for speech recognition. *J. Acoust. Soc. Japan (E)*, **21**, 79–86.

Stevens, K. (1997). Articulatory-acoustic-auditory relationships. In W. J. Hardcastle and J. Laver, editors, *The handbook of phonetic sciences*, pages 462–506. Blackwell, Cambridge.

Tachibana, M., Yamagishi, J., Masuko, T., and Kobayashi, T. (2005). Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing,. *IEICE Trans. Inf. & Syst.*, **E88-D**(11), 2484–2491.

Toda, T. and Tokuda, K. (2007). A speech parameter generation algorithm considering global variance for HMM-based speech synthesis. *IEICE Trans. Inf. & Syst.*, **E90-D**(5), 816–824.

Tokuda, K., Kobayashi, T., Fukada, T., Saito, H., and Imai, S. (1991). Spectral estimation of speech based on mel-cepstral representation. *IEICE Trans. Fundamentals*, **J74-A**(8), 1240–1248. in Japanese.

Yamagishi, J., Zen, H., Wu, Y.-J., Toda, T., and Tokuda, K. (2008). The HTS-2008 system: Yet another evaluation of the speaker-adaptive HMM-based speech synthesis system in the 2008 Blizzard Challenge. In *Proc. Blizzard Challenge 2008*.

Yamagishi, J., Kobayashi, T., Nakano, Y., Ogata, K., and Isogai, J. (2009a). Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm. *IEEE Trans. Speech, Audio & Language Process.*, **17**(1), 66–83.

Yamagishi, J., Nose, T., Zen, H., Ling, Z.-H., Toda, T., Tokuda, K., King, S., and Renals, S. (2009b). A robust speaker-adaptive HMM-based text-to-speech synthesis. *IEEE Trans. Speech, Audio & Language Process.*, **17**(6), 1208–1230.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (1999). Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In *Proc. EUROSPEECH-99*, pages 2374–2350.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (2000). Speaker interpolation for HMM-based speech synthesis system. *Acoustical Science and Technology*, **21**(4), 199–206.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (2001). Mixed excitation for HMM-based speech synthesis. In *Proc. EUROSPEECH 2001*, pages 2263 °!'2266.

Young, S., Odell, J., and Woodland, P. (1994). Tree-based state tying for high accuracy modelling. In *Proceedings of ARPA Human Language Technology Workshop*, pages 307–312, New Yersey, USA.

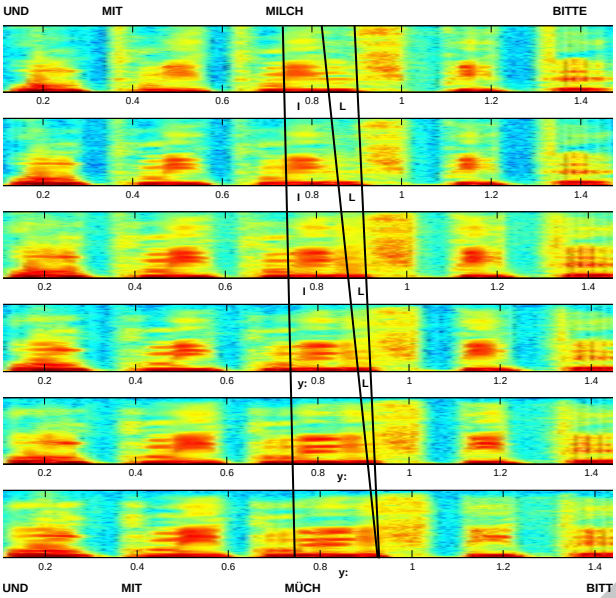
Zen, H., Toda, T., Nakamura, M., and Tokuda, K. (2007a). Details of Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005. *IEICE Trans. Inf. & Syst.*, **E90-D**(1), 325–333.

Zen, H., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (2007b). A hidden semi-Markov model-based speech synthesis system. *IEICE Trans. Inf. & Syst.*, **E90-D**(5), 825–834.

Zen, H., Tokuda, K., and Black, A. W. (2009). Statistical parametric speech synthesis. *Speech Communication*, **In Press**, –.

A. Additional interpolated examples

Figure 11 shows the *l-vocalization-2* process (Table 6) with and without switching rule. Without switching rule /L/ [l] gradually disappears and /I/ [ɪ] is gradually transformed into /y:/ [yː]. When a switching rule is applied /L/ is deleted.

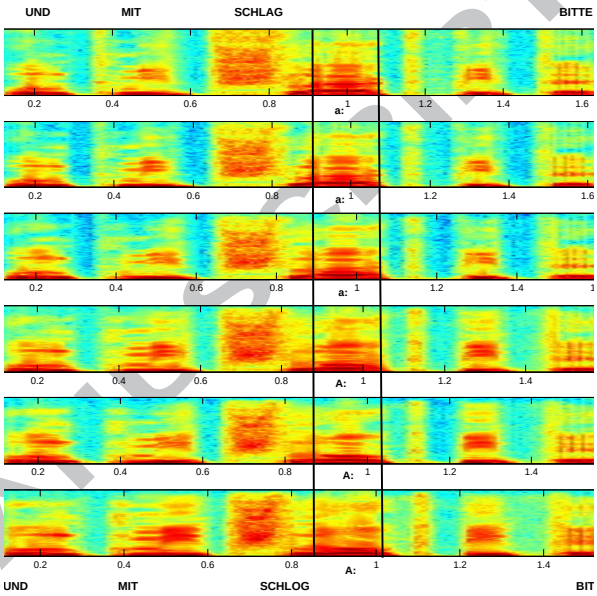


(a) without switching rules

(b) with switching rules

Figure 11: An interpolation example between Austrian German "Und mit Milch bitte" (*And with milk please*) and Viennese "Und mit Müch bitte" having different segmental structures. Interpolation ratio between them increments by 0.2.

Figure 12 shows the *input shift* process (Table 5), which is very similar to *monophtongization* (Table 4) as shown in Figure 6 where one vowel is transformed into another vowel. There is a continuous transformation when no switching rule is applied, whereas there is a categorical change when a switching rule is applied.



(a) without switching rules

(b) with switching rules

Figure 12: An interpolation example between Austrian German "Und mit Schlag bitte" (*And with cream please*) and Viennese "Und mit Schlog bitte". Interpolation ratio between them increments by 0.2.