Phone set selection for HMM-based dialect speech synthesis

Michael Pucher
Telecommunications Research Center (FTW)
Vienna, Austria
pucher@ftw.at

Nadja Kerschhofer-Puhalo
Acoustics Research Institute (ARI)
Vienna, Austria
nadja.kerschhofer@oeaw.ac.at

Dietmar Schabus
Telecommunications Research Center (FTW)
Vienna, Austria
schabus@ftw.at

Abstract

This paper describes a method for selecting an appropriate phone set in dialect speech synthesis for a so far undescribed dialect by applying hidden Markov model (HMM) based training and clustering methods. In this pilot study we show how a phone set derived from the phonetic surface can be optimized given a small amount of dialect speech training data.

1 Introduction

In acoustic modeling for dialect speech synthesis we are confronted with two closely related major problems[1], (1) to find an appropriate phone set for synthesis and (2) to design a recording script with sufficient phonetic and prosodic coverage. In HMM-based synthesis, we can use the training process of the voices itself to analyze the used phone set and to try to optimize it for synthesis.

2 Corpus and phone set design

Goiserian, the dialect of Bad Goisern in the most southern part of Upper Austria, is a local dialect of the Middle Bavarian/Southern Bavarian transition zone. The target variety for speech synthesis described here demonstrates the typical problems related to scarcity of data. While several varieties of the central and northern part of Upper Austria are quite well described, detailed descriptions of the varieties in this region do not exist. Lacking a lexicon, a phonological description, orthographic rules or a transcription system, a speech corpus and an appropriate phone set have to be created. Our current project aims at audio-visual dialect synthesis, which is based on a systematic description of speech data collected from spontaneous speech, word lists and translated sentences by 10 speakers of the same dialect. Although it would be ideal to use conversational speech data for dialect speech synthesis (Campbell, 2006) we decided to use a hybrid approach for our full corpus where we plan to collect a set of prompts from conversational dialect speech, which will be realized by the dialect speakers.

The speech data for the first preliminary study presented here consists of 150 sentences and colloquial phrases spoken in Goiserian by a female speaker who can be described as a conservative speaker of the original basic dialect of the region. The prompts were translated spontaneously by the speaker from Standard German into Goiserian and contain typical phonetic and phonological characteristics of local Bavarian varieties in multiple occurrences.

3 Voice building

The data was originally recorded at 96kHz, 24 bit and was downsampled to 16kHz, 16 bit for synthesis and voice building. A preliminary phone set (PS1) was created on the basis of a fine phonetic transcription including sub-phonemic details (e.g. nasalization of vowels before nasals "VN"). Phones occurring less than twice were substituted prior to voice training with phonetically similar phones or representatives of the same phoneme. This leaves us with a set of 72 phones (see Table 1 and 2).

[1] Apart from additional problems that have to do with text analysis, orthography, and grapheme-to-phoneme conversion.

Proceedings of EMNLP 2011, Conference on Empirical Methods in Natural Language Processing, pages 65–69,
Edinburgh, Scotland, UK, July 27–31, 2011.

The TRA voice was trained with a HMM-based speaker-dependent system. Given the limited amount of training data (150 prompts) and to be able to analyze the decision trees we only used the current, 2 preceding, and 2 succeeding phones as features.

HTK	IPA	#	HTK	IPA	#
s	s	207	t	t	204
d	d	179	n	n	171
m	m	115	k	k	98
h	h	84	g	g	79
v	v	79	f	f	62
r	r	61	S	ʃ	49
N	ɳ	42	l	l	41
b	b	31	ts	ts	27
ng	ŋ	19	p	p	17
w	β	14	L	ɭ	12
X	x	11	c	c	10
RX	χ	9	j	j	7
R	ɾ	6	ks	ks	3
pf	pf	3			

Table 1: Consonants (27) in phone set PS1 for training (72 phones) (Blue = not in PS2).

Based on a phonetic transcription of the training corpus, flat-start forced alignment with HTK was carried out. Stops are split into two parts, one for the closure and one for plosion plus burst. Additionally, we applied forced alignment using pronunciation variants[2], which is the preferred method when building a voice for dialect synthesis using a larger corpus (Pucher, 2010). With this method it is not necessary to have a phonetic transcription of the recordings. Given our small corpus, this method produced several errors ([tsvoa] / [tsvai], [tsum] / [tsun] etc.) which led us to use the standard alignment method from a transcription of the corpus. After the transcription we had to correct severe alignment errors. These errors are simple to find since several segments within the utterance are affected.

From this corpus we selected 5 prompts containing only phonemes that appear at least more than 3 times in the rest of the corpus. This leaves us with a training corpus of 145 prompts and a 5 prompt test set. For the subjective evaluation, the entire re-synthesized corpus was used to show us how well the used phone set covers the data.

[2]In a previous project on Viennese dialect synthesis, 33% of the lexicon entries are pronunciation variants.

HTK	IPA	#	HTK	IPA	#
a	a	138	aa	aː	10
A	ɒ	80	AA	ɒː	3
AN	ɒ̃	80	Ai	ɒi	3
AuN	ɒ̃u	7			
e	e	100	ee	eː	9
ei	ei	22	eiN	ẽi	10
E	ɛ	20	EE	ɛː	11
EN	ɛ̃	4	EiN	ɛ̃i	6
i	i	175	ii	iː	7
iN	ĩ	6			
o	o	45	oo	oː	3
ou	ou	4	Ou	ɔ	4
u	u	20	U	ʊ	15
UN	ʊ̃	3			
q	ø	9	qY	øʏ	3
QY	œʏ	4			
y	y	9	yy	yː	3
Y	ʏ	4			
eV	ə	11	aV	ɐ	89
ai	ai	24	aiN	ãi	9
au	au	24	ea	eɐ	7
eaN	ẽɐ	4	ia	iɐ	30
oa	oɐ	16	oaN	õɐ	9
Oi	ɔi	6	oi	oi	26
ua	uɐ	21	ui	ui	6

Table 2: Vowels (33) and diphtongs (12) in phone set PS1 for training (72 phones) (Blue = not in PS2, Red = not in PS2 and PS3, green = not in PS3).

The 145 prompts were then used for training a speaker-dependent HMM-based synthetic voice. Figure 1 shows the architecture of the HMM-based speaker dependent system (Zen, 2005). For synthesis we used the full-context label files of the corpus without duration information. By that text analysis is not necessary for synthesis. Our implicit assumption is that the letter-to-sound rules and text analysis produce exactly the string of phones from the transcription. In this way we can evaluate the acoustic modeling part separately, independently from text analysis.

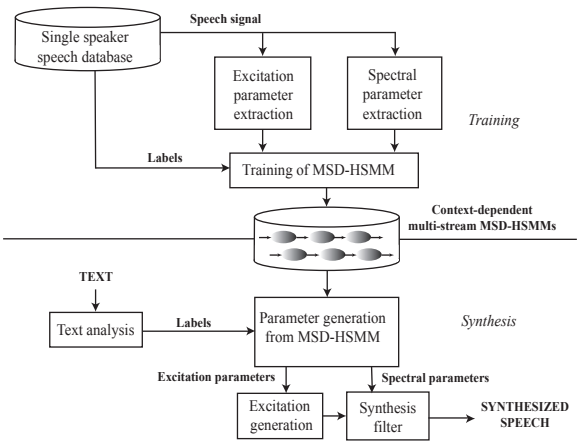

Figure 1: HMM-based speaker dependent speech synthesis system.

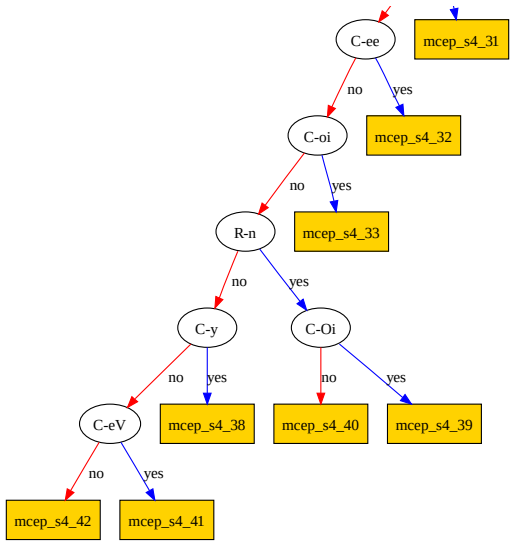

Figure 2: Part of the mel-cepstrum clustering tree for the 3rd state of the HMM.

4 Voice analysis

To be able to analyze the decision trees we used phone features only. The HMM-based voice consists of a mel-cepstrum, duration, F0, and an aperiodicity model. In a first step we defined the phones that are not used for modeling, or are used for a certain model only.

Figure 3 shows those phones that are not used for clustering of the different models. This may be due to their rare occurrence in the data (3-4 times) or due to possible inconsistencies in their phonetic realization. The F0 model is not shown since all phonemes were used in the F0 tree in some context.

To define other possible phone sets we decided to substitute the phones only occurring in the F0 model but not in the other 3 models, namely the mel-cepstrum, duration, and the aperiodicity model. We therefore merged “Ai”, “AuN”, “EN”, “ks”, “L”, “Ou”, “qY”, “yy” with their phonetically most similar equivalents (e.g. syllabic “L” with “l”, “ks” with “k”+“s”, or a nasalized “EN” or “AuN” before nasals with the non-nasal phone) and thus obtained a new smaller phone set (PS2), which was used for training a second voice model.

Another possible set of phones (PS 3) is defined by merging long (VV) and short (V) vowels of the same quality, namely “ii”, “yy”, “ee”, “EE”, “aa”, “AA”, “oo” with their short counterpart. From a linguistic point of view, the phonemic status of vowel duration as a primary feature in Austrian German is a controversial issue. While differences in length do exist at the phonetic surface, these differences are not necessarily of phonemic relevance (Moosmüller, 2007; Scheutz, 1985). We obtain thus a third phone set (PS3) by merging long and short vowels.

Model	#	#C	#L	#LL	#R	#RR
Mel-cep.	42	38	2	0	1	0
Aperiod.	36	31	0	3	0	1
F0	196	54	37	38	30	36
Duration	83	32	14	9	14	13

Table 3: Number of models and questions in mel-cepstrum, aperiodicity, F0, and duration model for central HMM state.

4.1 Mel-cepstrum and aperiodicity model

The mel-cepstrum model contains a separate model for each phone that is used in the cepstral clustering. In Figure 2 this is shown with the model “mcep_s4_32”, which is used in case that the current phone is an “ee” (C-ee) and with the model “mcep_s4_33”, which is used in case that the current phone is an “oi”. These two models are special models which only cover certain phones. The only effect of the clustering is that some phones are not modeled separately, resulting in an unbalanced tree.

However there is one instance of context cluster-

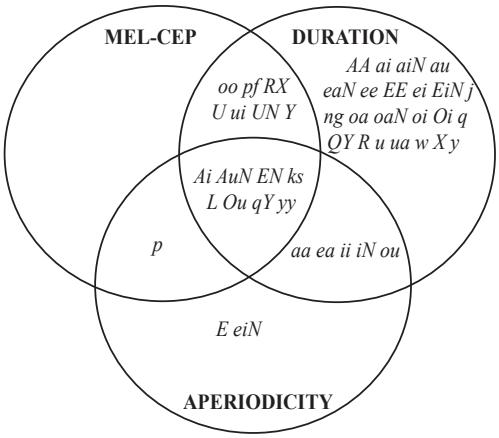

Figure 3: Phones that were not used for clustering in the trees for mel-cepstrum, duration, and aperiodicity in any context (current, 2 preceding, and 2 succeeding phones) and any of the 5 states.

ing in the central state of the mel-cepstrum HMMs. If the right phone is an "n" (R-n) there are two different models used ("mcep_s4_39", "mcep_s4_40"), depending on whether the current phone is an "Oi" (C-Oi) or not (Figure 2).

All phones that are not modeled through a separate model are modeled by the model at the end of the tree (model "mcep_s4_42").

The aperiodicity model is very similar to the mel-cepstrum model, as can be seen in Table 3 and Figure 3.

4.2 F0 and duration model

The F0 model uses all phones as shown in Figure 3 and is the most complex model in terms of context questions as can be seen from Table 3.

The duration model contains the lowest number of phone related questions as shown by Figure 3 but is still more complex than the spectrum related models in terms of context-dependent questions as shown in Table 3. Similarly to the F0 model, it is rather difficult to analyze this model directly.

5 Voice evaluation

After the analysis of the voice that was trained with our basic phoneset PS1 we defined two new phonesets PS2 and PS3. These phonesets were used to train additional voice models for the same speaker. With these voice models, we synthesized our small set of 5 test sentences. To evaluate the suitability of the phonesets for the training data, we re-synthesized the training corpus of 145 prompts.

In a pair-wise comparison test of the 150 prompts we evaluated the three voice models in a subjective listening test with three expert listeners. The experts listened to a set of prompts, each prompt synthesized with two different voice models. They were asked to compare them and to decide which prompt they would prefer in terms of overall quality, or whether they would rate them as "equally good".

PS1	PS2	PS3
56	102	105

Table 4: Number of winning comparisons per phone set (PS1-PS3).

Table 4 illustrates that both approaches to reduce and redefine the phoneset (PS2, PS3) improved the overall quality estimation considerably compared to the initial phoneset PS1.

6 Conclusion

One major challenge for speech synthesis of so far undescribed varieties is the lack of an appropriate phoneset and sufficient training data. We met this challenge by deriving a phoneset directly from the phonetic surface of a very restricted corpus of natural speech. This phone set was used for voice training. Based on the outcome of the first voice training we reconsidered the choice of phones and created new phone sets following 2 approaches: (1) removing phones that are not used in the clustering, and (2) a linguistically motivated choice of phone substitutions based on clustering results. Both methods yielded a considerable improvement of voice quality. Thus, HMM-based machine learning methods and supervised optimization can be used for the definition of the phoneset of an unkown dialect. Our future work will elaborate this method with dialect speech training corpora of different size to show whether it can be applied to adaptive methods involving multiple-speaker training. The consideration of inter- and intra-speaker variation and style shifting will be a crucial question for further study.

References

Nick Campbell. 2006. *Conversational speech synthesis and the need for some laughter.* IEEE Transactions on Speech and Audio Processing, 14(4), pages 1171-1178.

Michael Pucher, Friedrich Neubarth, Volker Strom, Sylvia Moosmüller, Gregor Hofer, Christian Kranzler, Gudrun Schuchmann and Dietmar Schabus. 2010. *Resources for speech synthesis of Viennese varieties.* In Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC), Valletta, Malta.

Sylvia Moosmüller. 2007. *Vowels in Standard Austrian German. An Acoustic-Phonetic and Phonological Analysis.* Habilitationsschrift, Vienna.

Hannes Scheutz. 1985. *Strukturen der Lautveränderung.* Braumüller, Vienna.

Heiga Zen and Tomoki Toda. 2005. *An Overview of Nitech HMM-based Speech Synthesis System for Blizzard Challenge 2005.* In Proceedings of the 9th European Conference on Speech Communication and Technology (INTERSPEECH), Lisboa, Portugal.