PHONE MAPPING AND PROSODIC TRANSFER IN SPEECH SYNTHESIS OF SIMILAR DIALECT PAIRS

Michael Pucher[1]*, Carina Lozo*[2]*, Sylvia Moosmüller*[1]

[1]*Acoustics Research Institute, Austrian Academy of Sciences, Vienna, Austria*
[2]*Department of Linguistics, University of Vienna, Austria*
{michael.pucher,sylvia.moosmueller}@oeaw.ac.at,carina.lozo@gmail.com

Abstract: In this paper we describe a phone mapping based method that can be used to synthesize a new dialect with an existing dialect model of a similar dialect. The method only uses transcriptions of original dialect data, which are then mapped onto the phones in the model. We use prosodic transfer of original duration and F0 to evaluate how the basic mapping model can be improved. We show that the prosodically enhanced models can outperform the basic model in a pairwise comparison task and can also achieve a slightly higher score on dialect authenticity. The goal of the proposed systems is to realize a dialect synthesis system with a small amount of symbolic training data that can come from transcribed dialect utterances or from the literature.

1 Introduction

For authentic dialect synthesis, we require a high-quality speech corpus of phonetically transcribed dialect utterances. The collection of such corpora is a time consuming task, due to the non-standard nature of dialects. In this paper, we evaluate a phone mapping method that allows us to synthesize a dialect of which no training corpus is available by using a trained acoustic Hidden-Markov-Model (HMM) of a similar dialect. As our target dialect we choose a dialect of southwestern Styria (STY). The source dialect HMM was trained on data of the dialect of Innervillgraten (IVG) [1]. We chose IVG to be the source language for the synthesis, since both dialects belong to the South Bavarian dialect group and therefore share a similar phone set. Both dialects also overlap regarding characteristic phones such as the retroflex lateral [ɭ]. We aim to use the similarity of these dialects to realize an authentic synthesized output for the STY dialect. As regards the phone mapping, each STY phone is mapped to an IVG phone, which defines a context-free mapping between the two phone sets. The mapping is created manually but can in principle also be derived automatically with a vector distance approach, if phones are defined by feature vectors. The phone set of STY was determined via a short recording of 14 sentences of an authentic dialect speaker from southwestern Styria. The phonetic transcriptions of these 14 sentences are also used as test sentences for the synthesizer. In the evaluation, the original recordings are compared with the synthesized ones to measure the dialect authenticity of the synthesizer. For the evaluation we synthesize three different versions of the test sentences. The first version is fully synthesized, the second with the original STY speaker duration, and the third with the original STY speaker duration and F0. In this way, we want to investigate the influence of STY dialect prosody on the synthesizer quality and authenticity. To evaluate the quality and the authenticity of the synthesized dialect with our 14 test sentences, we choose listeners who are either from this region, or closely familiar with this dialect. To evaluate if the synthesized output is recognized as Styrian at all, Austrians who are not strongly acquainted with STY in general will evaluate the synthesized output as well. With our method, we achieve the synthesis of STY exclusively by using the already created synthesis system for IVG and a phone mapping between STY and IVG. By this means, we can prepare the ground for further studies and experiments involving poorly digitalized dialects.

2 Corpora and voices

Ten dialect speakers, gender balanced, were recruited for the IVG corpus. The recordings consisted of spontaneous speech, reading tasks, picture naming tasks, and translation tasks from SAG into the dialect.

Table 1 – Phone sets in IPA symbols.

Category	(R)SAG	IVG	STY
Vowels (monoph.)	ɑ ɑː ɒ ɒː ɐ e ɛ eː i ɪ iː ɔ o oː øː æː ɐ œœː ə u ʊ uː ʏ y yː	ɑ ɑː ɒ ɒː e ɛ eː ɛː i ɪ iː ɔ o ɔː oː ɐ œə u ʊ uː ʏ y yː	ɑ ɑː ɒɛ e eː i ɪ ɔ œøœː ɐ ə ʉ ʊ u ʏ y
Vowels (monoph.) nasalized	ɒ̃ː ɔ̃ː ɐ̃ː œ̃ː		
Diphthongs	aɪ̯ ɒːɐ̯ ɑːɐ̯ ɒɪ̯ ɑʊ̯ ɛɐ̯ ɛːɐ̯ ɪɐ̯ iɐ̯ iːɐ̯ ɔɐ̯ ɔːɐ̯ oːɐ̯ ʏɐ̯ øːɐ̯ œɐ̯ ʊɐ̯ uːɐ̯ ʊːɐ̯ ʏɐ̯ yːɐ̯	aɐ̯ aɛ̯ ɑɔ̯ ɑɔɐ̯ ɑoɐ̯ eɑ̯ ɛɑ̯ eɛ̯ eɛ̯ ɛe̯ ɛi̯ ɛɐ̯ iɛ̯ ɪi̯ iɐ̯ ɔɑ̯ oɑ̯ ɔɛ̯ oɛ̯ ɔo̯ ɔɐ̯ ɔʊ̯ ɛɔ̯ ɐʊ̯ ui̯ ʊi̯ ʊu̯	ɑɔ̯ aɛ̯ ɑɛ̯ ɛɪ̯ ɛɐ̯ eɪ̯ ɪɐ̯ oɐ̯ ɔɐ̯ ɔɑ̯ ɔʊ̯ ʊɑ̯
Diphthongs nasalized			
Plosives (stops)	b̥ d̥ g̥ k ʔ p t	b̥ b d̥ d ɖː g̥ g k kʰ ʔ t tʰ	b̥ b d̥ d g̥ g k kʰ t tʰ
Nasal stops	m n ŋ	m n ŋ	m n ŋ
Fricatives	ç x f h s ʃ v z ʒ	β ð ç x f ɣ h s ʃ v	β ç x f ɣ h ɦ s ʃ v z
Affricates		b̥͡f kˣ͡ks t͡ʃ ts	b̥͡f kˣ͡ ts
Approximants	j	j	
Trill	r	ʀ ʀˣ	
Lateral approx.	l	l ɭ	l ɭ

From these recordings, 660 phonetically balanced sentences were selected and a phone set was created for each dialect. For the recording of the 660 phonetically balanced dialect sentences both the audio and the orthographic script, based on Standard German, of the samples to be collected were presented to the dialect speakers, who were then asked to repeat the individual samples in the IVG dialect. In addition, these speakers also read a corpus of SAG sentences. The speaker selection and recording process for IVG has been described in detail in [2]. In this paper we use the data from the male IVG speaker CSC, who, at the time of the recording, was 47 years old. 656 IVG dialect sentences were recorded from this speaker.

Sound samples were recorded at 44100 Hz, 16 bits/sample. The training process was also performed using these specifications. Cutting and selection was performed manually. Noise cancellation and volume normalization was applied to the recordings. Synthesized samples used in the evaluation were also volume normalized. A 5 ms frame shift was used for the extraction of 40-dimensional mel-cepstral features, fundamental frequency and 25-dimensional band-limited aperiodicity [3] measures. Speaker-dependent models were trained for the evaluations using the HSMM-based speech synthesis system published by the EMIME project [4].

The data of the STY speaker was obtained in fall 2016 in Zelko, a small village in the southwestern region of Styria, belonging to the political district of Groß St. Florian. For the recording, an autochtonic male dialect speaker, aged 75, was selected. The speaker spent all his life in this village, he reported no significant absences from this place. The recording took place at the speaker's house. The old farmhouse with only one big room presented an acoustic challenge. We decided to bring nub foam and prepared a small cubicle to improve the acoustic conditions. Due to the thick walls and small windows of the house, we were able to achieve a satisfactory recording setting. The recording session was split into two parts, the first session consisted of spontaneous conversational speech, which included a biographical narrative, questions of the speaker's language attitudes, and a picture naming task. The second session focused on the elicitation of dialectal elements. Based on the speakers request, we changed the originally planned reading task into a repetition task. A final translation task from Standard Austrian German into the dialect closed the recording session. From the 48 sentences of the repetition task, we chose 14 to

Table 2 – mapped phones

STY	ø	ʉ	ɑɛ̯	ɛɪ̯	eɪ̯	ɪɐ̯	oɐ̯	ɦ	z
IVG	œ	ʊ	aɛ̯	ɛi̯	ɛi̯	iɐ̯	ɔɐ̯	h	s

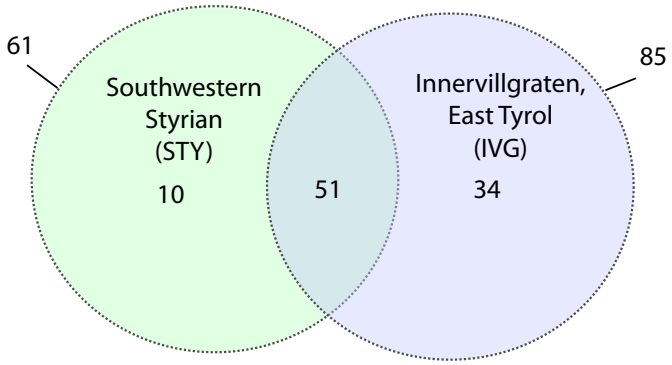

Figure 1 – Phone sets overlap

be later transcribed into IPA and SAMPA, following a phone-level segmentation. The sentences for the repetition task were gathered from the IVG reading corpus, see [1]. As mentioned above, we decided that a dialect from the same dialect group as STY would benefit the experiment because of the expected overlaps in their phone sets. If we had chosen a dialect from another dialect group the mapping process would have been a lot more time-consuming. The dialect of Bad Goisern, for example, belongs to the Middle Bavarian dialect group and would have left us with only 41 overlaps by a total amount of over 92 phones.

3 Phone Mapping

The examined 14 sentences allowed us to extract a small phone set (see Table 1), followed by a comparison between the two phone sets STY and IVG. On the basis of the phonetic transcription of 14 sentences, we determined 61 STY phones, 51 of them were also present in the IVG corpus, leaving only 10 phones without an IVG equivalent. Each of these phones were manually assigned to the most similar ones in the IVG phone set, meaning in the case of vowels, the mapped phone must be in a similar constriction location and must not differ in more than two distinctive features (quality, dorsality, height, or roundedness) from the STY vowel. For example, the close-mid front rounded vowel [ø] was mapped to the open-mid front vowel [œ], differing only in quality. As most of the mapped STY phones are diphthongs which differ in terms of tenseness from the IVG phones, we decided to map diphthongs in a similar way to vowels. For IVG equivalents to STY diphthongs, we chose IVG diphthongs with the highest agreement in constriction location and the distinctive features mentioned above. By this approach, only the diphthong [ʊɑ̯] remained unassigned. Since the diphthong [ʊɑ̯] could not be mapped satisfactorily, we considered breaking the diphthong up into its parts [ʊ] and [ɑ] and synthesize those two parts separately. Assimilation processes rendered the occurrence of voiced consonants like [ɦ] or [z]. In these cases, we mapped the voiceless cognates of the IVG phoneset.

4 Synthesis

For synthesizing the three different versions of our prompts, the phonetic transcription of the STY utterances were transformed into phonetic transcriptions of IVG utterances using the mapping process described in Section 3. The mapped phones are shown in Table 2.This phonetic transcription which also contains syllable and word boundaries, was then transformed into full-context labels that can be used with the HMM-based synthesis system. The whole process emulates the function of a speech synthesis front-end that performs text normalization, Letter-to-Sound (LTS) conversion, and utterance/full-context label building. Since we do not have such a front-end ready for the IVG voices we use this approach.

Using the full context labels and a speaker dependent voice of the IVG speaker CSC that we have developed, we synthesize the first set of prompts (*syn*). The second set of prompts (*syn_dur*) uses the phone duration information from the original STY speaker. In this way we can synthesize a sample that has exactly the same length as the original speaker, and each phone has almost the same length as the original speaker's phones. The HMM state level durations are, however, taken from the HMM, since we do not have a state alignment of the speaker. We could get a state alignment of the STY speaker by using the Viterbi algorithm with the IVG model, but then we would have to merge this with our manual transcription. As a third test set we additionally used the original speaker's F0 values during synthesis (*syn_dur_f0*). Here we also used the original speaker's phone durations, since otherwise it would have been necessary to perform a warping of original and synthesized F0 curves. To change the F0 to a similar range as measured for the IVG speaker, we modified the STY speaker's F0 curves $F0_{orig}$ as follows:

$$F0_{syn\_dur\_f0} = F0_{orig} + (\frac{\sum F0_{syn\_dur}}{N} - \frac{\sum F0_{orig}}{N}).$$

N is the number of F0 values, e.g. frames in the utterance, i.e. this adaptation was done on the utterance level. We also tried a more sophisticated approach that also takes the variance into account as in [5], but this introduced some artifacts on the utterance level. Additionally, we used the voicing decision from the synthesized F0 curve. The full algorithm for changing F0s is described in ALgorithm 1.

Algorithm 1 Algorithm for computing F0.

$N = \text{length}(F0_{syn\_dur}) = \text{length}(F0_{orig})$
for $f0_{syn\_dur}$ in $F0_{syn\_dur}$ **do**
  if $f0_{syn\_dur}$ is not 0.0 **then**
    $f0_{syn\_dur\_f0} = f0_{orig} + (\frac{\sum f0_{syn\_dur}}{N} - \frac{\sum f0_{orig}}{N})$
    if $f0_{syn\_dur\_f0} < 0.0$ **then**
      $f0_{syn\_dur\_f0} = 0.0$
    end if
  else
    $f0_{syn\_dur\_f0} = 0.0$
  end if
end for

The F0 algorithm was applied to only take over the F0 dynamics from the original speaker and to keep the F0 range within the range of the synthetic speaker, such that the comparison of synthesized samples is easier. There was a significant difference in F0 between the original and the synthesized voice; the original voice having a higher F0 than the synthesized one.

5 Evaluation

For the evaluation we had 10 listeners from different regions of Austria. From age 22 to 66, 6 female and 4 male listeners. We chose the participants according to their dialect familiarity. Seven of the listeners were considered as dialect speakers, since they were born and raised in rural regions of Austria. Three of them were raised in the particular region of southwestern Styria where our STY speaker comes from, two listeners were born and raised in Upper Styria. The remaining three dialect speakers had their main place of residence in Carinthia and Tyrol until adulthood. One listener can be considered as a standard speaker, since, in contrast, his/her parents have an academic background and he/she was raised in Vienna.

Table 3 – Word-Error-Rates (WER) in %

method	*orig*	*syn*	*syn_dur*	*syn_dur_f0*
WER	32.6	46.8	51.2	51.5

The evaluation consisted of an intelligibility test, a Mean Opinion Score (MOS) test on dialect authenticity, and a pair-wise comparison of the voices. For all tests we also included the original samples, such that we had $14 * 4 = 56$ samples in total. For the first part the listeners had to write the perceived content of audio samples into a text field. They were only allowed to listen to each sample once. The evaluation started with this part, because the listeners must not know the prompts for the intelligibility test. Each listener heard each of the 14 prompts once, with one of the synthesis conditions ((*orig, syn, syn_dur, syn_dur_f0*)). Table 3 shows the Word-Error-Rates (WER) from the intelligibility part of the evaluation for each method. The original is more intelligible than the synthesized samples, while there are only small differences in the synthesized versions. It has to be kept in mind that these are dialect sentences, which are more difficult to understand than standard ones. This can be seen on the high WER of the original sentences.

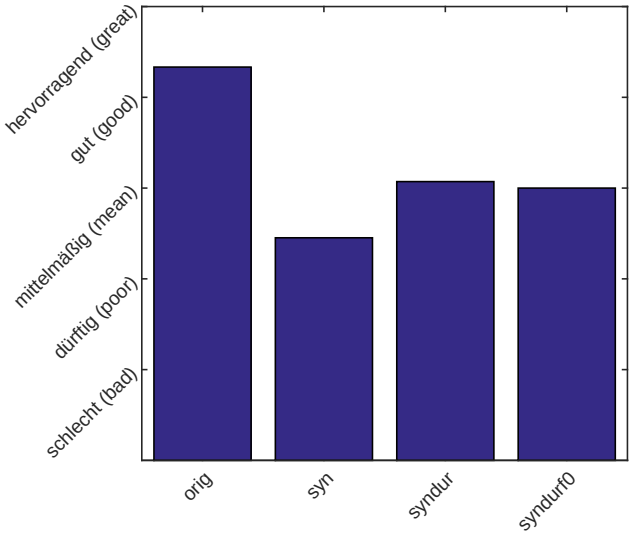

Figure 2 – Mean Opinion Score (MOS) evaluation.

In the second part each listener had to score all audio samples, both of the synthesized and the original dialect. For the evaluation the listeners rated each sample on an ordinal scale (1 - "schlecht (bad)", 2 - "dürftig (poor)", 3 - "mittelmaßig (mean)", 4 - "gut (good)", to 5 - "hervorragend (great)"). They were asked to evaluate the dialect authenticity of the samples. Figure 2 shows the average scores for each method. The highest mean after the original samples was given to the *syn_dur* samples (mean = 3.307). As expected, the *orig* audio samples were considered the best in terms of authenticity. We can see that the synthesized samples that use prosodic transfer (*syn_dur*, *syn_dur_f0*) are in general slightly better than the synthesized samples without (*syn*).

Figure 3 shows the results of the pairwise comparison, the third part of the evaluation. Here we have plotted the random variable that gives the pairwise comparison score between two methods for each listener. Figure 3 shows that the original samples were always rated better than the other samples, i.e. they won each comparison. The basic synthetic examples (*syn*) have won the comparisons in 30% of cases on average. The synthetic samples that used prosodic transfer of the original duration (*syn_dur*) have won the comparisons in 50% of cases on average. The best method was the synthesis method that used prosodic transfer of the original duration and F0 (*syn_dur_f0*). This method won the comparisons in 70% of cases on average. According to a two-sample t-test ($p < 0.001$) the *orig* method is different from all other methods, the *syn* method is different from the *syn_dur* and *syn_dur_f0* methods, while the *syn_dur* and *syn_dur_f0* methods are not different.

6 Conclusion

In this paper we showed how phone mapping and prosodic transfer can be used for the synthesis of a new dialect. The simple phone mapping based system (*syn*) could achieve a slightly higher intelligibility (46.8% WER), but a lower MOS score of "dürftig (poor)" compared to the systems that also used

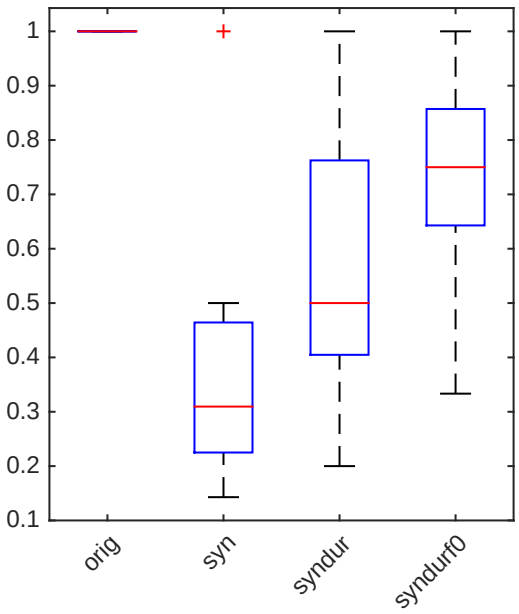

Figure 3 – Results of pairwise comparison.

prosodic transfer. In the pairwise comparison task there was a significant difference between the phone mapping based system and the systems that also use prosodic transfer. The systems using prosodic transfer are however not full synthesis systems, since the duration and F0 values were not synthesized but just extracted from the original speaker. Thus, the basic phone mapping based system can be seen as a way to build a full moderate quality synthesizer for a new dialect. In terms of dialect authenticity and intelligibility however, all proposed synthesis systems need further improvement. This improvement could come from the phone mapping, which needs a better handling of diphthongs, or from other adaptive approaches that need little adaptation data.

We could also see that although both dialects have a large number of overlapping phones, the contextual differences between the dialects as well as speaker differences lead to a decrease in synthesis quality. This holds for the phone mapping based models and the prosodic transfer models, and also shows the difficulty of synthesizing a new dialect without using acoustic data of the new dialect.

References

[1] TOMAN, M., M. PUCHER, S. MOOSMÜLLER, and D. SCHABUS: *Unsupervised and phonologically controlled interpolation of Austrian German language varieties for speech synthesis*. *Speech Communication*, 72, pp. 176–193, 2015.

[2] TOMAN, M. and M. PUCHER: *Structural KLD for cross-variety speaker adaptation in HMM-based speech synthesis*. In *SPPRA*. Innsbruck, Austria, 2013.

[3] KAWAHARA, H., I. MASUDA-KATSUSE, and A. CHEVEIGNÉ: *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, pp. 187–207, 1999.

[4] YAMAGISHI, J. and O. WATTS: *The CSTR/EMIME HTS system for Blizzard challenge 2010*. In *Proceedings of the Blizzard Challenge Workshop*, pp. 1–6. Kansai Science City, Japan, 2010.

[5] QIAN, Y., J. XU, and F. K. SOONG: *A frame mapping based HMM approach to cross-lingual voice transformation*. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5120–5123. IEEE, 2011.