Adaptive speech synthesis of Albanian dialects

Michael Pucher[1], Valon Xhafa[2], Agni Dika[2], and Markus Toman[1]

[1] Telecommunications Research Center Vienna, Austria
www.ftw.at
pucher@ftw.at,toman@ftw.at
[2] Department of Computer Engineering, University of Prishtina, Kosova
www.uni-pr.edu/
valon.xhafa@uni-pr.edu,agni.dika@uni-pr.edu

Abstract. In this paper, we show how adaptive modeling within the statistical parametric speech synthesis framework can be applied to Albanian dialects. We develop speaker dependent voices for the Tosk and Gheg dialect and adapt models for the Gheg dialect from the Tosk models. We show that the adapted Gheg models outperform the speaker dependent Gheg model on an intelligibility and dialect classification task. Furthermore we show that the speaker dependent Tosk model outperforms a formant based synthesizer on an intelligibility, dialect classification and pair-wise comparison task. This formant based synthesizer is the only publicly available synthesizer for Albanian at the moment. We also show that our Gheg and Tosk synthesizers are as intelligible as natural speech. The method where one dialect is modeled through adaptation of a closely related other dialect can be applied to language varieties in general, where the background variety and adapted variety can be chosen based on pragmatic considerations like speaker or data resource availability.

Keywords: speech synthesis, Albanian, adaptation, dialect

1 Introduction

Adaptive parametric HMM-based speech synthesis [1, 2] allows for the usage of a background model or average model to improve synthetic voice quality with small amounts of adaptation data. The authors of [3] applied the adaptive framework to the synthesis of Austrian German dialects. In this paper, we use an Albanian Tosk dialect background model to improve an Albanian Gheg dialect adapted voice.

The adaptive approach that works with small amounts of data is especially interesting for languages such as Albanian where not so many language resources are available. Today the only open-source available synthesizer for Albanian is a formant-based synthesizer that is still in a "provisional" development stage [4]. The synthesizers developed for this study are based on open-source components [4, 5] and we plan to release open-source synthesizers for Albanian.

The method we use here can be used for any variety pairs of phonetically closely related varieties (dialects, sociolects, and accents). We can use the variety were it is easier to collect a larger amount of data to train a background model. Then we can adapt the variety where it is more difficult to obtain large amounts of data. It can be

difficult to obtain data when speakers are difficult to find, or language resources like lexica, grapheme-to-phoneme rules, or texts for recording scripts are not available. In our case we choose the Tosk dialect, which is also the basis for Standard Albanian, as background model since it was easier to obtain the large amount of text data that is needed to select a recording script by solving the respective set-cover problem. Furthermore, we could use a slightly modified grapheme-to-phoneme conversion from an existing synthesizer for Tosk [4].

2 The Albanian language

The Albanian language belongs to the family of Indo-European languages. In the tree of languages, the Albanian language does not share any descent connection with other member languages of this family and is presented as a separate branch that grows from the root of the tree. This classification is based on phonological, morphological and other features [6].

There are two basic dialect forms of the Albanian language: Gheg and Tosk [7, 8]. Furthermore there are also mixtures between Gheg and Tosk dialects, like Arbëresh and Arvanitika [7]. In countries where Albanian is spoken the official language is Standard Albanian, based on the Tosk dialect [9]. It is used in institutions, newspapers, and books. Gheg is spoken in more informal settings. For these reasons we adapted Gheg from Tosk, using the Tosk language in our background model. The Tosk dialect has 7 vowels and 29 consonants. The Gheg dialect uses long and nasal vowels, which are absent in Tosk [7].

Albanian almost has a one-to-one correspondence between letters and phones, which make grapheme-to-phoneme conversion easier, compared to some other languages.

3 Grapheme-to-phoneme conversion

Grapheme-to-phoneme (G2P) rules for the Tosk dialect were taken from an existing speech synthesizer [4]. Some errors of the G2P rules had to be corrected for Tosk and additional rules had to be introduced for Gheg as shown in Table 1. In general the G2P problem is relatively simple to solve for Albanian since there is an almost one-to-one mapping between letters and phones.

c → ts [ts]	ô→ nO
ll → L [l]	û→ nU
dh → D [d]	â → nA [ɑ̃]
sh → S [ʃ]	ê → nE [ɛ̃]

Table 1. *Additional G2P rules for Tosk and Gheg (left) and Gheg (right).*

The additional rules for Tosk and Gheg shown in Table 1 are context independent, i.e. the characters “dh” are replaced everywhere by “D”. Table 1 also shows the International Phonetic Alphabet (IPA) symbols in brackets. Due to phonological differences

between Tosk and Gheg we need to introduce additional rules for Gheg. The Gheg dialect has the nasal vowels “ê” and “â”. These nasal vowels are mapped to “nA” and “nE” in our phone set.

4 Recording script

First we collected a large amount of sentences from books, newspapers and other sources to have a database from which we can select a set of sentences that fulfills specific phonetic criteria. Sentences with proper nouns were removed from the corpus to avoid any problems with irregular pronunciation in the training data. Furthermore, we excluded sentences that are longer than 20 words, which would make the reading during the recordings too difficult. This corpus without sentences containing proper nouns and long sentences consisted of 16041 sentences.

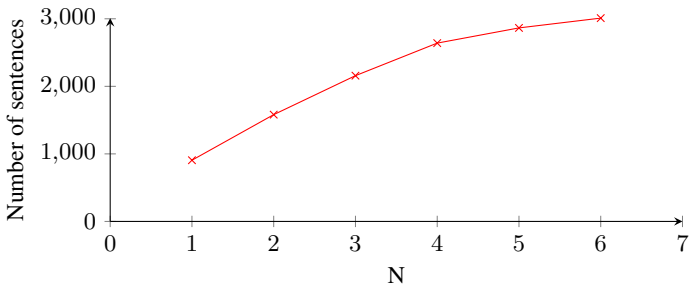

Fig. 1. Selected sentences for different minimum number of diphones for Tosk.

For selecting the recording script from the collected corpus we used the modified G2P rules from Section 3 to phonetically transcribe the selected 16041 sentences. Then we used a greedy algorithm to select the smallest set that contains all diphones N times, where N was in the range of $1-6$. Figure 1 shows the number of selected sentences for different number of N for the Tosk dialect. This algorithm always adds the sentences to the set that contains the largest number of new diphones. We used a greedy algorithm for this, since the set-cover problem is known to be NP-complete [10]. Our greedy algorithm will not necessarily find the smallest set of sentences containing all diphones N times.

5 Recording

The recording was done in a semi-professional soundproof recording room. We recorded 3010 sentences from one female Tosk speaker (containing all Tosk diphones 6 times) and 412 sentences from one female Gheg speaker (containing all Gheg diphones 1 time). We recorded more material from the Tosk speaker since we wanted to use Tosk as background language. To have good quality recordings, we recorded a small amount

of the corpus every day for three weeks. The microphone used was a vocal Sennheiser microphone with a pop filter. After recording, low frequency noise was removed. The recording of a whole recording session was automatically split into utterance size audio files.

6 Voice building

We built a speaker dependent model from the data of a Tosk female speaker and a Gheg female speaker. Adapted models for the Gheg speaker were trained through adaptation from the Tosk model. For the speaker dependent and adaptive training we used the training scripts from HTS 2.3-alpha [5]. Clustering questions included phone identity, phonetic, and articulatory features for current, previous two, and following two phones as well as word features. A flat-start model was used for forced alignment of the training data. The overall quality of the adapted voice could be improved by using more speakers in the background/average voice model [1, 2, 11].

7 Evaluation

We evaluated the different synthetic and natural voices through a subjective listening test that consisted of an intelligibility test and a pair-wise comparison to evaluate the quality of the voices.

7.1 Design

For the evaluation we compared the methods for Tosk synthesis and the methods for Gheg synthesis. All used voices are shown in Table 2. For each of the dialects we had one speaker.

10 listeners participated in the evaluation with age between 19 and 23. We had 5 male and 5 female listeners.

Table 2. *Voices used in the evaluation.*

Dialect / Method	Recorded	Speaker-dependent	Adapted	Formant-based
Tosk	✓	✓		✓
Gheg	✓	✓	✓	

In the first part of the evaluation each listener had to listen to each of the 20 prompts for Tosk and 20 prompts for Gheg (40 samples in total). Listeners were asked to write down what they have heard (intelligibility test) and if they think it was the Tosk or Gheg dialect. Figure 2 shows a screenshot of the application that was used for this part of the evaluation.

In the second part of the evaluation listeners had to listen to pairs of prompts and had to judge which one of the prompts is better in terms of overall quality.

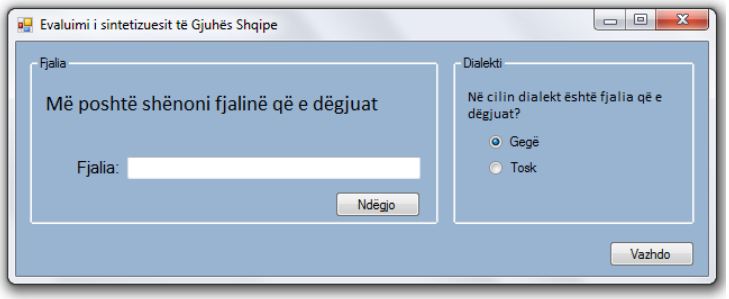

Fig. 2. Application used for evaluation of Word-Error-Rate (WER) and dialect rating.

7.2 Results

Word-Error-Rate (WER): Table 3 shows the Word-Error-Rate results for the different methods that were computed by finding the minimum number of edit operations that are necessary to convert the correct transcript into the transcribed version divided by the total number of words in the correct transcript.

Table 3. *Word-Error-Rate (WER) for the different methods in %.*

Error / Method	Tosk rec.	Tosk spk. dep.	Tosk formant	Gheg rec.	Gheg spk. dep.	Gheg adapt.
WER	4.7	7.8	23.5	8.2	9.9	11.9

For the Tosk synthesizers we can see that the formant based synthesizer is worse than the HMM-based synthesizer and that there is only a small difference between HMM-based synthesizer and recorded speech. This shows that the HMM-based synthesizer outperforms the formant synthesizer on the intelligibility task.

For the Gheg synthesizers there are only small differences between the speaker dependent and adapted version.

Dialect rating: Table 4 shows the Dialect Classification Error (DCE) that was computed as the ratio of wrongly classified samples and total samples.

For the Tosk dialect we can see that the formant based synthesizer had the worst performance in terms of modeling the dialect accurately. 16.1% of the samples were wrongly classified as Gheg dialect samples. The speaker dependent HMM-based voice has similar performance to the recorded samples.

Table 4. *Dialect classification error (DCE) for the different methods in %.*

Error / Method	Tosk rec.	Tosk spk. dep.	Tosk formant	Gheg rec.	Gheg spk. dep.	Gheg adapt.
DCE	4.6	4.4	16.1	14.7	19.4	10.7

For Gheg the Dialect Classification Error (DCE) was in general higher than for Tosk, which might be related to our listeners' competence. Here we can see that the adapted HMM-based voice has the lowest error followed by the recordings and the speaker dependent voice.

Interestingly the speaker-adapted voice is more often found to produce Gheg samples than the original Gheg samples. This result shows that there is a difficulty with classifying the Gheg dialect, which either comes from the language competence of our Gheg speakers or from the Gheg classification competence of our listeners.

Pair-wise comparison: Figure 3 shows the results for the pair-wise comparison task. For Tosk the HMM-based speaker dependent voice is significantly better than the formant voice ($p < 0.05$ according to a paired T-test). The recordings are not surprisingly significantly better than both synthetic voices. Thereby we can show that the HMM-based method outperforms the formant-based method also for this task.

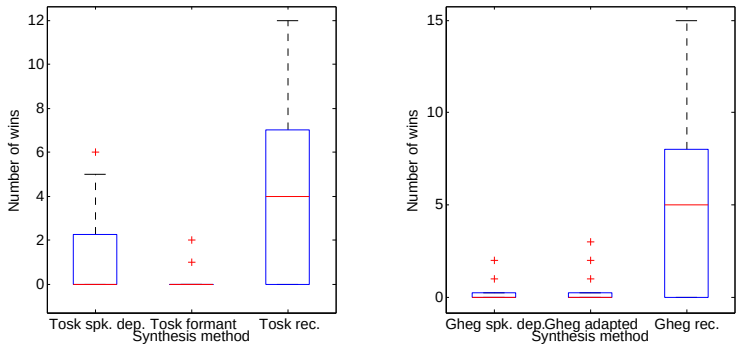

Fig. 3. Results of pair-wise comparison between different methods for Tosk (left) and Gheg (right).

For the Gheg voices we can see no significant differences between the synthetic voices. Only the recorded voice is significantly better in terms of overall quality.

8 Conclusion

We have developed a state-of-the-art HMM-based synthesizer for the Albanian Tosk and Gheg dialects. We showed that our speaker dependent Tosk model outperforms an existing formant-based synthesizer on an intelligibility, dialect classification and pair-wise comparison task. We also saw that the speaker-adapted Gheg model outperformed the Tosk model on a dialect classification and pair-wise comparison task, while there was no performance difference for the pair-wise comparison task. We also showed that the synthesizers are as intelligible as natural speech, which makes them usable in many application scenarios such as spoken dialog systems, web readers, and speech output for blind users.

9 Acknowledgements

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

References

1. J. Yamagishi, M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi. A training method of average voice model for HMM-based speech synthesis. *IEICE Trans. Fundamentals*, E86-A(8):1956–1963, August 2003.
2. J. Yamagishi and T. Kobayashi. Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training. *IEICE Trans. Inf. & Syst.*, E90-D(2):533–543, February 2007.
3. M. Pucher, D. Schabus, Y. Yamagishi, F. Neubarth, and V. Strom. Modeling and interpolation of Austrian German and Viennese dialect in HMM-based speech synthesis. *Speech Communication*, 52:164–179, 2010.
4. eSpeak. eSpeak Text-to-Speech. `http://espeak.sourceforge.net/`, 2007.
5. HTS. HMM-based speech synthesis system (hts). `http://hts.sp.nitech.ac.jp/`, 2014.
6. K. Tyshchenko. *Metatheory of Linguistics*. 1999.
7. Shaban Demiraj. *Gjuha Shqipe dhe historia e saj*. Onufri, 2013.
8. Eqrem Çabej. *Studime gjuh'esore*. Rilindja, 1976.
9. Sylvia Moosmüller and Theodor Granser. The spread of standard albanian: An illustration based on an analysis of vowels. *Language Variation and Change*, 18:121–140, 2006.
10. Christos Papadimitriou. *Computational Complexity*. Addison Wesley, 1994.
11. Bálint Tóth and Géza Németh. Improvements of Hungarian Hidden Markov Model-based text-to-speech synthesis. *Acta Cybern.*, 19(4):715–731, January 2010.