Evaluation of state mapping based foreign accent conversion

Markus Toman, Michael Pucher

FTW Telecommunications Research Center Vienna, Austria

toman@ftw.at, pucher@ftw.at

Abstract

We present an evaluation of the perception of foreign-accented natural and synthetic speech in comparison to accent-reduced synthetic speech. Our method for foreign accent conversion is based on mapping of Hidden Semi-Markov Model states between accented and non-accented voice models and does not need an average voice model of accented speech. We employ the method on recorded data of speakers with first language (L1) from different European countries and second language (L2) being Austrian German. Results from a subjective evaluation show that the proposed method is able to significantly reduce the perceived accent. It also retains speaker similarity when an average voice model of the same gender is used. Accentedness of synthetic speech was rated significantly lower than natural speech by the participants and listeners were unable to identify accents correctly for 81% of the natural and 85% of the synthesized samples. Our evaluation shows the feasibility of accent conversion with a limited amount of speech resources.

Index Terms: speech synthesis, accent perception, foreign accent conversion

1. Introduction

Accented speech is ubiquitous in everyday communication. It has been shown that adaptive speech synthesis [1] can be applied to create voices that retain the speaker accent [2]. Foreign accent conversion aims to convert from one speaker variety to another one. The main direction of conversion is from accented speech to some form of standard where foreign accent conversion becomes accent reduction.

Foreign accent reduction has possible applications in language learning where the non-accented synthetic voice of an originally accented speaker can be used as a target for language learning [7]. Through the application of interpolation methods [3] it is also possible to gradually transform an accented speakers synthetic voice into a non-accented voice. These methods also have a possible application in adaptive spoken dialog systems where the accent or grade of accent is changed based on contextual factors.

In this paper we are investigating how well a speaker's accent can be reduced in a Hidden Semi-Markov Model (HSMM) based speech synthesizer, and if the speaker identity can be retained. We apply a transformation method that does not use an average accent-specific voice. We adapt an HSMM state mapping method [4] that we also used for dialect transformation in [5]. While in [5] the method was applied to mapping a dialect average voice with a standard language average voice, here we investigate the case when no average voice model of a specific accent is available. We recorded accented data from 10 speakers with first languages (L1) from different European countries and their second language (L2) being Standard Austrian German (SAG) [6]. In a subjective listening test we evaluate the perception of natural and synthetic accented speech and accent-reduced synthetic speech.

This paper is organized as follows: in Section 2 we describe related work and in Section 3 we present the corpus and models used. Section 4 describes the accent conversion method and Section 5 analyzes the state mappings produced by the method. In Section 6 we present our evaluation, which is discussed and concluded in Section 7 and 8.

2. Related work

In [7] it was shown that it is beneficial for language learning when students are able to hear to their own voice producing native-accented utterances. In their work they use contours of F0, local speech rate, and intensity from a native reference speaker and copy them to the learners' speech signals. They found these resynthesized less accented stimuli to be more effective than natural stimuli of the non-accented reference speaker in a language learning experiment. This copy synthesis method only allows for the modification of accented recordings.

In [8] accented portions of speech are replaced with alternative less accented segments from the same speaker corpus by finding the segment that is closest to the respective segment from a native reference speaker. They report a 20% reduction in perceived accent compared to the natural accented speech and also observe a strong coupling between accent and speaker identity. This concatenative method relies on a method for detecting the most accented speech segments within an utterance.

[9] and [2] have shown that speaker adaptation with a set of 105 sentences is able to overrule the influence of the accent of the average voice and produce accented voice models. In our method we use voice models of accent speakers adapted from a non-accented average voice model as baseline. [10] showed that synthesized foreign accent received lower accent ratings by listeners than naturally produced accent. With our adaptive approach we assume to have similar findings concerning accent ratings.

In [9] is was also shown that listeners perform worse in speaker discrimination tasks when facing mixed conditions with natural and synthetic voices. Therefor we conclude that the baseline to evaluate accent conversion should not be natural but synthetic speech. [11] also found that listeners perform worse in cross-lingual speaker discrimination tasks. This is very likely to also influence speaker discrimination tasks between accented and non-accented speech of the same speaker, as also suggested in [8]. We therefore also expect that accent conversion has an effect on the speaker discrimination performance.

3. Corpus and models

We recorded 5 female and 5 male speakers with (Austrian) German as second language (L2) and the following first languages (L1) (with ISO 639-1 codes): Bulgarian (BG), UK-English (EN), Estonian (ET), French (FR), Greek (EL), Slovakian (SK), Spanish (ES), Hungarian (HU), Serbian (SR). 320 utterances were recorded for each speaker from which 23 were held out as test set. The remaining 297 utterances were used for adaptation from an average voice of 9 non-accented, male Austrian German speakers, trained from 1790 utterances. The CSTR/EMIME HTS system [12] was used for adaptation. Sound samples were recorded and used for training in 44100 Hz. Cutting and selection was performed manually. Noise cancellation was applied to the recordings before training and volume normalization to the synthesized samples used in the evaluation. A 5 ms frame shift was used for the extraction of 40-dimensional mel-cepstral features, fundamental frequency and 25-dimensional band-limited aperiodicity [13] measures.

The basis for the accent conversion method was the system presented previously in [5]. In contrast to the dialectal models used in [5], we used the same Standard Austrian German (SAG) phone set for the accented and non-accented models.

4. Accent conversion

The method for foreign accent conversion used here is based on our previous work on language variety transformation in [5] which employs Kullback-Leibler-Divergence (KLD) to map similar HSMM state models between two average voice models [4].

As average models of specific accents of a certain language are often not available, we apply HSMM state mapping between an accented, speaker specific voice model and a non-accented average voice model. We assume here that phone identity has a stronger influence on the KLD metric than speaker identity and that therefore this approach is feasible.

In [5] we also introduced constraints on the mappings for the overlapping phone sets of two language varieties to improve the quality of the transformed voice. As in this work we use the same Standard Austrian German (SAG) phone set for accented and non-accented models, these constraints are still likely to improve quality but also constrain the degrees of freedom for transforming the voice models (i.e. removing the accent). Our hypothesis was that the constrained mapping actually leads to a less effective accent reduction. To verify this, we included both the unconstrained and the constrained mapping methods in our experiments. As can be seen in the evaluation results in Section 6 we could not confirm this hypothesis.

The procedure to generate accent-reduced voice models is firstly to find the best mappings between HSMM states from the accented speaker specific model and the non-accented average model. For the constrained mapping method, the center phone that occurred most often in the data used to train an HSMM state model of the accented voice has to occur at least once in the training data of the mapped HSMM state model from the average voice model. From the mappings that satisfy this constraint, the mapping with the highest similarity rating is selected. According to this set of selected mappings, the accented adaptation data is then placed at the nodes of the decision tree of the non-accented voice model. In a next step, the decision tree is pruned until every node has at least one (accented) adaptation datum associated with it. Finally, regular speaker adaptation is performed.

5. Mappings in KLD-based methods

We wanted to see how many KLD mappings actually differ for our models when constraining the set of possible mappings. Figure 1 shows the number of matching mappings where the unconstrained and the constrained method selected the same mapping (i.e. the KLD-wise best mapping is not affected by the constraints), and the number of not matching mappings where both methods selected different mappings (i.e. due to the constraints, the KLD-wise best mapping was not used). In total 69% of the mappings were affected by the constraints, meaning that for the constrained method 69% of the selected mappings were not the KLD-wise best matches. This shows us that both methods are indeed very different and will produce different adapted acoustic models.

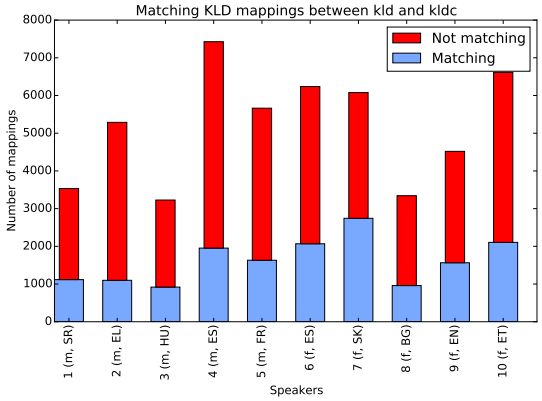

Figure 1: Number of mappings that differ when using constrained KLD for each speaker model.

6. Evaluation

For our evaluation we wanted to assess the accent degree, accent identity and speaker identity as perceived by the listeners for different methods. The methods used for the evaluation are:

- accented natural speech ("rec")
- accented synthetic speech ("adapt")
- accent-reduced synthetic speech using KLD-based state mapping ("kld")
- accent-reduced synthetic speech using constrained KLD-based state mapping ("kldc")

20 listeners participated in the evaluation, 10 female and 10 male from age 25 to age 65. They had to perform a listening test consisting of two parts with three different tasks. In the first part, the listeners had to identify and rate accents of 48 sound samples, in the second part they had to discriminate speakers in 120 sound sample pairs (i.e. 240 samples)[1].

6.1. Accent identity and rating

In this part of the evaluation each listener was presented 48 samples to rate. The first step for each sample was to rate the degree of the accent using a slider that ranged from "0 - no accent" to "100 - very strong accent". The second step was to try to identify the heard accent. The listeners could select an accent from

[1] Examples at http://userver.ftw.at/~mtoman/interspeech2015/

a list showing the accents used in the evaluation. There was also the option to select none. If the listener moved the accent rating slider to "no accent", the accent selection disappeared. The listeners were asked to ignore the quality of the samples and focus on the accent only. The evaluation data consisted of 12 utterances uttered by 10 speakers using 4 methods, resulting in 480 unique samples. These were distributed equally to all listeners twice (i.e. 960 samples), resulting in 48 accented samples per listener. As a baseline we also included recordings and adapted synthetic speech from a non-accented speaker. These were 24 additional samples that were also distributed to all listeners twice, resulting in 2 to 3 non-accented samples for each listener.

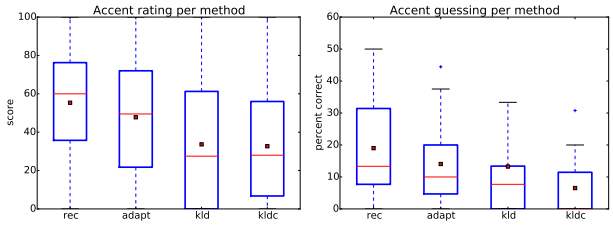

Figure 2: Accent ratings (left) and guesses (right) per method.

Figure 2 shows the accent ratings and identity guesses for all accented speakers. It can be seen that the median accent rating for the recorded samples across all speakers was 60. The adapted voices were rated with a median rating of 49.5. While the authors found the synthetic voices to retain the characteristics of the accent quite well, the difference between recorded and adapted speech was still significant ($p < 0.007$, double-sided Mann-Whitney U test). This result agrees with the findings of [10]. "kld" and "kldc" achieved median ratings of 27.5 and 28 respectively with the differences to "adapt" and to "rec" being highly significant ($p \ll 0.001$, double-sided Mann-Whitney U test). The difference between "kld" and "kldc" was not significant.

The results of the accent identity guessing task in Figure 2 are based on the samples for which the listeners selected an accent rating > 0 and therefore had the option to select an accent. Also, the data from the non-accented listener was excluded. As expected, accents in "rec" samples were recognized correctly most often but still only with a median correctness of 13.3%. The speaker identified most often was the French male speaker, who was identified correctly with median 16% and mean 22%.

Figure 3 presents histograms of the accent ratings for all methods. It can be seen that the histograms for the accent reduced samples have high counts of ratings close to zero, which was the "no accent" setting for the rating slider in the evaluation interface.

Figure 4 shows the accent ratings for all speakers and methods. It can be seen that the ratings between speakers varied a lot, with speaker 2 (Greek, male) having the highest ratings (median 81.5 for "rec", 86.5 for "adapt"). The non-accented speaker is not shown as he was correctly identified as non-accented and received a median accent rating of 0.

After the evaluation the listeners reported that they felt able to detect the presence of an accent but were unable to identify it most of the time. Many stated that they rated accents not too high as long as "they were able to understand the speaker." Conversely they tended to give higher accent ratings if the quality of the sample was degraded, despite being told to try to ignore

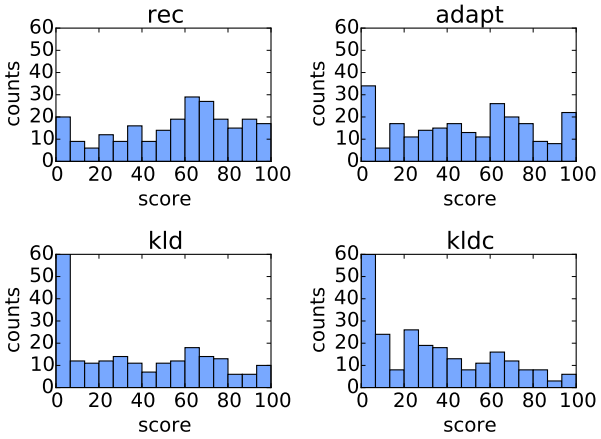

Figure 3: Accent rating histograms per method.

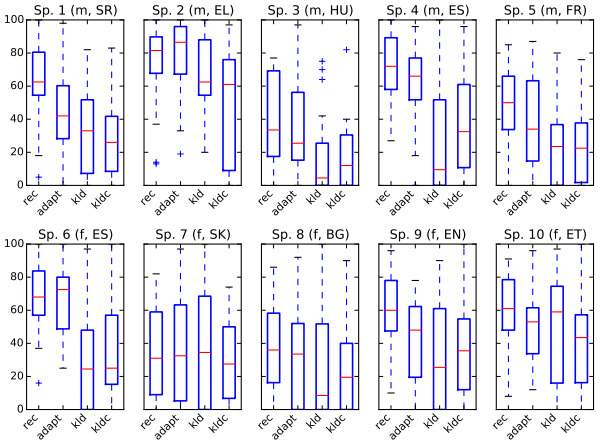

Figure 4: Accent ratings per speaker and method. For each speaker the gender (m/f) and L1 language is given in the title.

quality issues. This was especially true for samples created by the KLD-based methods which often introduced errors in the synthetic speech.

6.2. Speaker identity

We assessed speaker identity by a speaker discrimination task. Our experiment setup was similar to the setup proposed in [9]. The listeners were presented with two samples at a time and had to vote if the same speaker could be heard in both samples. The listeners were told to try to ignore quality and accent of the samples and focus on the voice itself. The evaluation data consisted of 12 utterances uttered by 10 speakers using 4 methods. Each of these samples was paired with a sample of the same speaker and with a sample of another random speaker of the same gender for all other methods. Also the two utterances within each pair differed. The sample pairs were then equally distributed to all listeners, resulting in 120 pairs (i.e. 240 samples) for each listener.

The overall results for the speaker discrimination task for "rec" and "adapt" samples can be seen in Figure 5. This shows that the KLD-based methods actually degrade speaker similarity. When listening to the synthesized samples, we noticed a

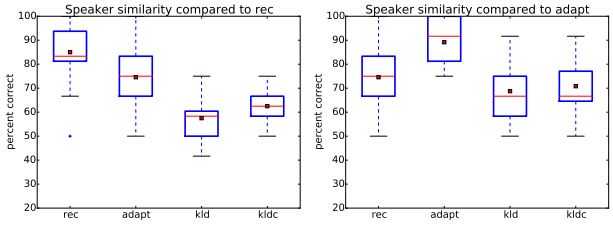

Figure 5: Speaker discrimination task results with reference method “rec” (left) and “adapt” (right).

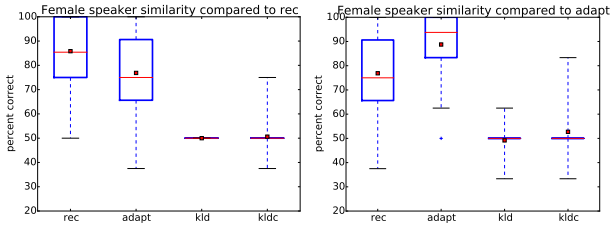

Figure 6: Speaker discrimination task results for female speakers with reference method “rec” (left) and “adapt” (right).

bias in speaker similarity between male and female speakers. We hypothesized that this is due to the fact that the average voice model was built from male speakers only.

In Figures 6 and 7 it can be seen that the speaker similarity between recorded and adapted synthetic speech was not significantly different for male and female speakers. This means that the adaptation of a female voice from a male average voice retained the speaker identity at a similar level as the adaptation of a male voice.

In Figure 7 it can also be seen that the median percentage of correct discriminations of adapted male speakers from KLD-based speakers was at about 90% compared to 50% for the female speakers shown in Figure 6. This suggests that the KLD-based methods are able to retain speaker similarity when a similar enough (in this case gender-matched) average voice model is used.

7. Discussion

[9] showed that listeners perform significantly worse in speaker discrimination tasks when natural and synthetic voices are mixed. We could also observe this in our experiments as well as the fact that natural voices received higher accentedness ratings from our listeners than the synthetic voices. [10] attributes this effect to the over-smoothing nature of HSMM-based speech

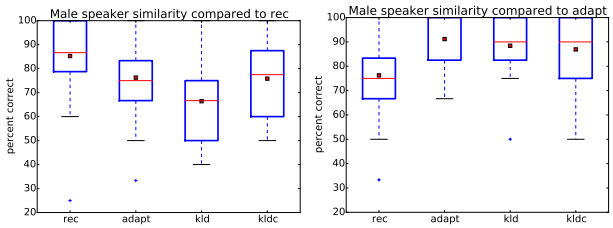

Figure 7: Speaker discrimination task results for male speakers with reference method “rec” (left) and “adapt” (right).

synthesis. In [11] is was also shown that listeners were significantly less accurate in cross-lingual speaker discrimination and [8] also revealed a strong coupling between accent and speaker identity. We hypothesize that the listeners of our experiment were also influenced by these effects, although we are not able to show the influence of accent on speaker discrimination with our experiments. To measure this effect we would need a different experiment design.

While the advantage of the method presented here is that it does not need an average accent voice we also noticed a higher quality degradation in the accent-reduced samples when compared with our previous experiments on dialect transformation [14] where we have used a dialect average voice. Our hypothesis that the constrained mapping method is less effective for accent reduction could not be confirmed. On the contrary, this method is able to alleviate the quality problem to a certain extent. Future work would include a subjective listening test on the quality difference between the constrained and unconstrained KLD methods. It would also be interesting to quantify the difference in quality when using an accented average voice model instead of the accented speaker dependent voice model.

8. Conclusion

In this contribution we analyzed the perception of accented natural and synthetic speech as well as accent-reduced synthetic speech at the example of accented Austrian German speech.

We presented a method for reducing accents for HSMM-based speech synthesis employing KLD-based state mapping between an accented HSMM voice model and a non-accented HSMM average voice model. We also presented an extension to this method which includes phonetic constraints on the space of possible state mappings. A listening test showed that a significant reduction of accent was possible with both KLD-based methods. We know from previous experiments that the method incorporating constraints introduces fewer errors into the synthesis, suggesting that including the constraints should be preferred.

We have also shown that with these methods, speaker similarity of the synthetic voices was degraded when using an average model of different gender than the accented speaker. When using an average voice model of the same gender, speaker similarity was on a similar level as that of the adapted voices. This means that a more careful selection of the average voice model used is necessary than with regular adaptation.

We also found that listeners rated the accentedness of synthetic speech lower than that of recorded speech. Listeners were able to detect the presence and rate the degree of an accent but were barely able to identify the (European) accents of Austrian German correctly.

9. Acknowledgements

This work was supported by the Austrian Science Fund (FWF): P22890-N23. The Competence Center FTW Forschungszentrum Telekommunikation Wien GmbH is funded within the program COMET - Competence Centers for Excellent Technologies by BMVIT, BMWA, and the City of Vienna. The COMET program is managed by the FFG.

10. References

[1] J. Yamagishi and T. Kobayashi, "Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training," *IEICE Transactions on Information and Systems*, vol. E90-D, no. 2, pp. 533–543, Feb. 2007.

[2] R. Karhila and M. Wester, "Rapid adaptation of foreign-accented HMM-based speech synthesis," in *Proceedings of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Florence, Italy, 2011.

[3] M. Pucher, D. Schabus, J. Yamagishi, F. Neubarth, and V. Strom, "Modeling and interpolation of Austrian German and Viennese dialect in HMM-based speech synthesis," *Speech Communication*, vol. 52, no. 2, pp. 164–179, feb 2010.

[4] Y.-J. Wu, Y. Nankaku, and K. Tokuda, "State mapping based method for cross-lingual speaker adaptation in HMM-based speech synthesis," in *Proceedings of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Brighton, United Kingdom, 2009, pp. 528–531.

[5] M. Toman, M. Pucher, and D. Schabus, "Cross-variety speaker transformation in HSMM-based speech synthesis," in *Proceedings of the 8th ISCA Workshop on Speech Synthesis (SSW)*, Barcelona, Spain, Aug. 2013, pp. 77–81.

[6] S. Moosmüller, C. Schmid, and J. Brandstätter, "Standard austrian german," *International Journal of the International Phonetic Association*, In press.

[7] M. P. Bissiri and H. R. Pfitzinger, "Italian speakers learn lexical stress of german morphologically complex words," *Speech Commun.*, vol. 51, no. 10, pp. 933–947, Oct. 2009.

[8] D. Felps, C. Geng, and R. Gutierrez-Osuna, "Foreign accent conversion through concatenative synthesis in the articulatory domain," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 8, pp. 2301–2312, Oct 2012.

[9] M. Wester and R. Karhila, "Speaker similarity evaluation of foreign-accented speech synthesis using hmm-based speaker adaptation," in *Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011, pp. 5372–5375.

[10] M. L. G. Lecumberri, R. Barra-Chicote, R. P. Ramón, J. Yamagishi, and M. Cooke, "Generating segmental foreign accent," in *Proceedings of the 15th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2014, pp. 1302–1306.

[11] M. Wester, "Cross-lingual talker discrimination," in *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Makuhari, Japan, September 2010.

[12] J. Yamagishi and O. Watts, "The CSTR/EMIME HTS system for Blizzard challenge 2010," in *Proceedings of the Blizzard Challenge Workshop*, Kansai Science City, Japan, Sep. 2010, pp. 1–6.

[13] H. Kawahara, I. Masuda-Katsuse, and A. de 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*, vol. 27, no. 3–4, pp. 187–207, 1999.

[14] M. Toman and M. Pucher, "Structural KLD for cross-variety speaker adaptation in HMM-based speech synthesis," in *Proceedings of the 10th IASTED International Conference on Signal Processing, Pattern Recognition and Applications (SPPRA)*, Innsbruck, Austria, 2013.