IMPROVING THE QUALITY OF SYNTHESIZED SPEECH OF A VIENNESE DIALECT SPEAKER THROUGH SPEAKER ADAPTATION

Lorenz Gutscher, Michael Pucher
Acoustics Research Institute
lorenz.gutscher@oeaw.ac.at

Abstract: Text-to-speech systems recently experienced a push towards Deep Neural Network-based approaches that achieve high quality for standard languages. Such systems especially benefit from big datasets that often are not available for dialects. By the example of a Standard Austrian German speaker and a Viennese Dialect speaker, it is examined if a combined training can improve the quality of the dialect speaker. We use the term "dialect" here because it is well known for describing regional language variation, although the Viennese dialect is a sociolect in the strict sense, since the variation is on a social dimension. It is shown that the similarities between those varieties are sufficient to benefit from the additional data. Using an open source neural network speech synthesis system [1], an average voice model is built and afterwards used to fine-tune it to the Viennese Dialect speaker. In a subjective listening test, participants are asked to rate stimuli in relation to perceived naturalness. Synthetic audio samples of the proposed model are judged as more natural compared to a baseline model where training is only based on the Viennese Dialect speaker. An objective evaluation indicates that – in reference to the natural recording – mel-cepstral distortion is roughly the same for both systems. When fine-tuning the average voice to the Standard Austrian German speaker, no remarkable benefits can be found. Additionally, it is tested whether a method called multi-task learning can further improve the synthesis quality by using additional pole-zero features for modeling nasal and lateral phones. Looking at objective and subjective measures, we conclude that in a multi-task learning setting the use of additional pole-zero features does not increase speech quality.

1 Introduction

Deep Neural Networks (DNNs) and the supply of big datasets for training play a key part in the progression of speech synthesis towards more natural speech [2]. With the flexibility of statistical parametric speech synthesis, it is possible to adapt models [3] to unseen speakers with less available data, capturing speech characteristics like timbre and fundamental frequency. Even though small corpora can benefit from pre-built models, limitations arise when a different language is chosen or speaking style differs greatly. Many dialects in Austria partly overlap with Standard Austrian German, but also differ significantly from the Standard in a phonetic context. Austrian German and Austria's dialect landscape offer interesting data for research on speaker adaptation in speech synthesis. Using a speaker adaptation method, it is tested whether Standard Austrian German can have a beneficial effect on the synthesis quality of a Viennese Dialect and vice versa. In a separate attempt to improve synthesis quality, an additional feature representation is tested using multi-task learning. It is shown by Enzinger et al. [4] that synthesis systems based on low-dimensional mel-frequency cepstral coefficients often underestimate formants and anti-formants for nasals and laterals. By using pole-zero features as a secondary task in the output layer of a DNN, it is expected to achieve increased accuracy of the predicted parameters of the acoustic model [5].

2 Method

2.1 Speaker adaptation

We propose a method (which will be referenced to as "speaker adaptation") to improve speech synthesis quality of a Viennese Dialect speaker by training an average voice model and fine-tuning it to the dialect speaker. In detail, the method is a two-step approach: Firstly, a combined training of both speakers is carried out using Merlin as a framework [1]. The model created resembles an average voice, as it is a mixture of the two speakers used for training. Secondly, the average voice model is adapted to a target speaker (Viennese Dialect speaker) by adjusting the weights of the DNN using back propagation. In addition to the proposed method, a "baseline" synthesis model is created for comparison using the same toolkit, but with the corpus of the dialect speaker only and without further fine-tuning. Training (90%), validation (5%), and test data (5%) are split up the same for both systems. We work with phone level representation. The phonetic labels used for testing are obtained from the transcription of the natural recordings. This makes comparison easier because we do not have to rely on a grapheme-to-phoneme mapping, which introduces additional possible errors that are not related to the acoustic model. Duration modeling for the test files is fully calculated by the model and not used from the natural recordings.

For the baseline model and speaker adaptation model, we use a frameshift of 80 samples (5 ms) and a window size of 1024 samples (64 ms). The order of mel-cepstral coefficients n and the frequency warping factor α are set to their default values ($n = 60$ and $\alpha = 0,58$). Additional features include Logarithmic Fundamental Frequency (LF0), Band Aperiodicity (BAP) and Voiced/Unvoiced (VUV) information. Derivatives of all features (except VUV) are calculated, which adds up to 187 parameter values per sample. A feed-forward network with six hidden layers of size 1024 and a tanh activation function is used. For the training data, we use phone alignment and contextual features. Contextual features can be derived from question files, containing linguistic information at a segmental, syllable, and word level as well as prosodic features on positioning within utterances. In summary, we end up with binary/numeric input vectors of size 2140, six hidden layers and a 187-dimensional output vector. Batch size for computation is set to 256.

2.2 Multi-task learning

In a separate process, the synthesis quality is attempted to be improved by multi-task learning. Multi-task learning is a method where a DNN is trained jointly for a main task and a secondary task that is related to the main task [6] (see Figure 1).

In our case, the main task uses the same features as the baseline model and the secondary task uses pole-zero features. Nasals and laterals have strongly developed formants and anti-formants due to the involvement of nasal cavities during the process of human speech production. Anti-formants in particular are often underdeveloped and poorly modeled by all-pole models [7]. Enzinger et al. [4] supposes that pole-zero models can improve the feature representation of such sounds. To obtain poles and zeros, an estimation of the spectrum is calculated. After that, a pole-zero model is fitted to it using estimations of the numerator and denominator (ND) coefficients [8] as shown in Figure 2.

Feature extraction for poles and zeros is done with an order of 11 in the z-plane. The numerator and denominator are thereafter converted to the frequency domain and finally only the lowest three center frequencies are used for further processing, similar to Enzinger et al. [4]. The lowest three formants and anti-formants are then converted to a logarithmic representation

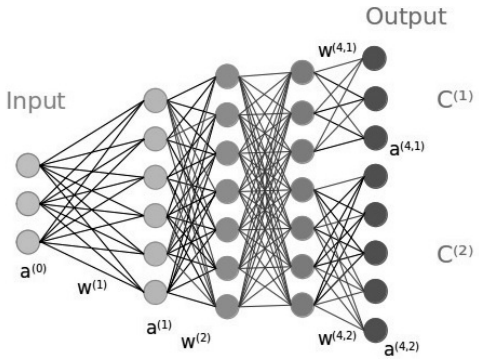


Figure 1 – Schematic structure of a multi-task learning network.

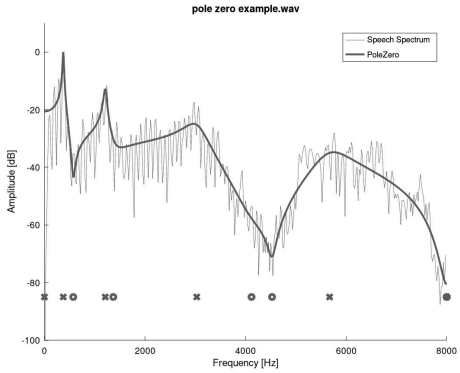


Figure 2 – Example of pole-zero estimation for one sample. Poles are marked as crosses and zeros as circles.

and used as parameters for the secondary task, together with their respective derivatives, adding 18 features to the output vector. While pole-zero features perform well for sounds with strong harmonics, their accuracy decreases for unvoiced sounds. Therefore, the DNN is adapted such that the secondary task is only used for nasals and laterals, while the main task is used for every other sound. This is solved by using a custom loss function, dependent on the current phone in each sample. If the current sample k is a nasal or lateral, b_k is set to one. For all other phones, its value is set to zero. The differentiation between nasals, laterals and all other sounds is implemented in Keras [9] and the loss function is set up as a Mean Square Error (MSE) for both tasks $C^{(1)}$ and $C^{(2)}$:

$$C^{(1)}_{nasals\,and\,laterals} = \beta \frac{1}{n} \sum_{k=1}^{n} (y_k b_k - \hat{y}_k b_k)^2$$

$$C^{(1)}_{non-nasals\,and\,non-laterals} = \frac{1}{n} \sum_{k=1}^{n} (y_k \bar{b_k} - \hat{y}_k \bar{b_k})^2$$

$$C^{(2)}_{nasals\,and\,laterals} = (1-\beta) \frac{1}{n} \sum_{k=1}^{n} (y_k^{pz} b_k - \hat{y}_k^{pz} b_k)^2$$

where y is the true output and $\hat{y}$ is the predicted output. β is a weight-factor that makes it possible to control the amount of influence that the main and secondary task have. After trying

different parameters for β in a preliminary test setting, it was set to $0,9$ as this yielded the best perceived results.

3 Results

The available audio recordings for the Standard Austrian German speaker amount to 3 hours and 24 minutes. The recordings for the Viennese Dialect speaker amount to 2 hours and 20 minutes. All files are down-sampled from $44,1$ kHz (16 bit) to 16 kHz (16 bit). Subjective evaluation is done with a MUSHRA (MUltiple Stimuli with Hidden Reference and Anchor) test, whereas the lower anchor is selected as a $3,5$ kHz band-passed version of the baseline system to ensure it is always judged the worst. The listeners are requested to use headphones and to rate four presented stimuli by moving a slider that ranges from 0 to 100. For each utterance, four stimuli and the reference are presented on the same page. Listeners can listen to each of the stimuli as many times as they want. One of the four stimuli always consists of the natural recording (reference) and one of the lower anchor (band-passed version of baseline system). The other two stimuli consist of the baseline synthesis and the method under investigation (i.e. speaker adaptation method). As there are many features in speech that participants might focus on, we tried to get an overall rating of the samples by asking the participants to rate each of the stimuli in comparison to the reference in terms of naturalness of speech. The presented order of the samples per page is chosen randomly for each test call to avoid ordering effects based on the position on the screen. The files for listening tests are chosen beforehand from the test pool, so that the selection comprises many categories of the corpus with different context. The listening tests are carried out using a web-based interface called webMUSHRA [10]. All participants are proficient listeners that are used to do listening tests.

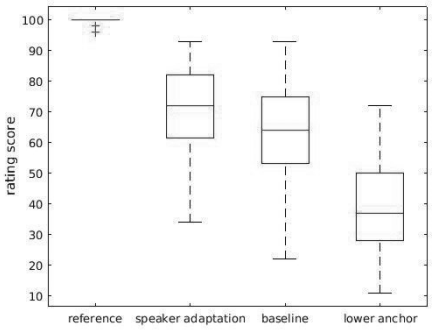
For an objective evaluation the synthesized utterances are each compared to the respective natural recordings using mel-cepstral distortion. Since the duration modeling of the synthesis most likely differs from the natural recording, Dynamic Time Warping (DTW) is used to align the synthesized version with the natural recording in time. Using the DTW function within the Speech Signal Processing Toolkit (SPTK) (`http://sp-tk.sourceforge.net/`), a distance score for each utterance is obtained.

3.1 Speaker adaptation – AVM fine-tuned to Viennese Dialect

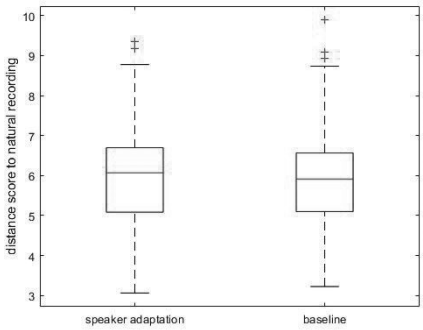
Figure 3a shows the result of the MUSHRA listening test. It contains ratings from nine participants who were presented with 15 (out of 142) test utterances.[1] Because the resulting rating scores of the subjective listening tests are not normally distributed for all stimuli, a Wilcoxon rank sum test is performed. As expected, the lower anchor point is rated the worst (mean score: $39,33$) and the reference stimulus is rated the best with scores close to 100 (mean score: $99,96$). Utterances created with the speaker adaptation method are judged better than the baseline system with differences being statistically significant ($p < 0,001$).

Figure 3b shows a boxplot of the distance score of all test utterances (142) to the reference, whereas smaller values indicate higher similarity. Looking at the mel-cepstral distortion of both synthesized stimuli, no significant difference is found in the objective evaluation ($p > 0,5$). Therefore, we conclude that the speaker similarity of the adaptation matches the speaker similarity of the baseline method. By listening carefully to both versions, we can confirm that the fine-tuned AVM sounds very similar to the baseline speaker in terms of timbre and pronunciation. Connecting those findings, one can conclude that even though the perceived

[1] Data available at: https://speech.kfs.oeaw.ac.at/essv2022/



(a) Subjective rating score (MUSHRA listening test.)



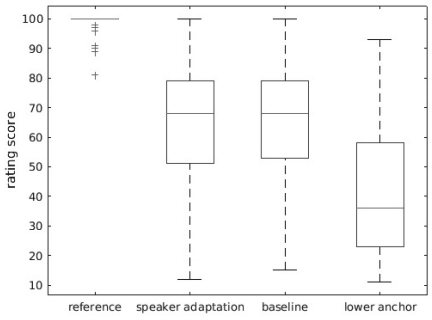
(b) Mel-cepstral distortion over all test utterances.

Figure 3 – Evaluation of synthesis quality (speaker adaptation = AVM fine-tuned to Viennese Dialect).

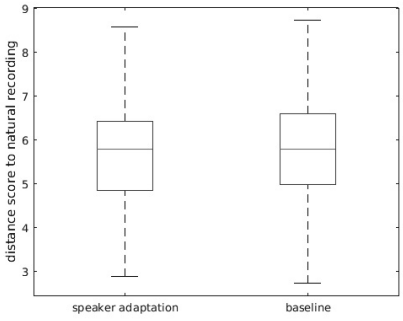
quality of the samples increased, speaker similarity is not impaired nor improved by the speaker adaptation method.

3.2 Speaker adaptation – AVM fine-tuned to Standard Austrian German

Out of 219 test files, 14 are selected for the MUSHRA listening test with 11 participants. No significant difference is found between the speaker adaptation method and the baseline model ($p - value = 0,097$) (see Figure 4a). This shows that fine-tuning the AVM to the Standard Austrian German speaker does not improve the perceived synthesis quality compared to the baseline system. It is assumed that this is due to the higher amount of data in the corpus of the Standard Austrian German speaker and because the model cannot find much new relevant information within the Viennese Dialect. Mel-cepstral distortion of all 219 test utterances shows similar results for the two methods (see Figure 4b), which correlates with the findings of the listening test.



(a) Subjective rating score (MUSHRA listening test.)

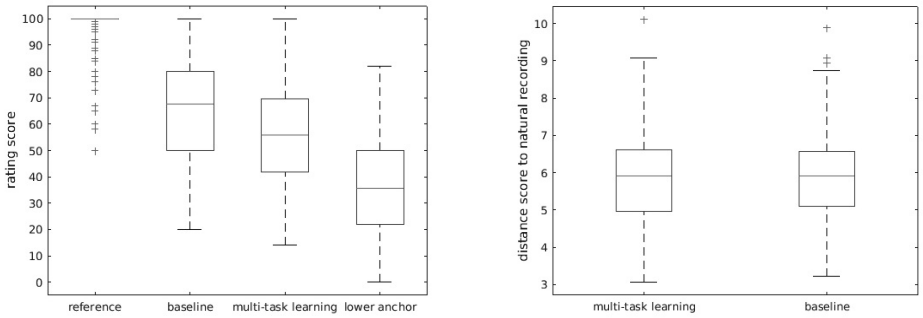


(b) Mel-cepstral distortion over all test utterances.

Figure 4 – Evaluation of synthesis quality (speaker adaptation = AVM fine-tuned to Standard Austrian German).

3.3 Multi-task learning

A subjective listening test with 12 participants shows that the baseline system is rated as superior compared to the multi-task learning model (see Figure 5). The listening test contains 7 test utterances from the Viennese Dialect speaker and 8 test utterances from the Standard Austrian German. Both speakers are trained separately. It can be observed that the rating of the reference stimuli has some outliers, which might come from accidental mix-ups by the participants or audio dropouts because of browser incompatibility. Nevertheless, the results confirm our impression that additional pole-zero features do not increase the quality in this test case. We conclude that the high order of mel-cepstral coefficients makes the use of additional parameters unnecessary. Therefore, we experimentally decreased the order of mel-cepstral coefficients ($n = 12$) to see if multi-task learning would benefit in this test case. Still, we received similar results showing no improvement in comparison to the baseline system (with equally reduced order).



(a) Subjective rating score (MUSHRA listening test.)

(b) Mel-cepstral distortion over all test utterances.

Figure 5 – Evaluation of synthesis quality (multi-task learning).

4 Conclusions

Combined training with Viennese Dialect and Standard Austrian German speakers has a beneficial effect on perceived speech quality of synthesized samples in the case of an Austrian Dialect speaker. While the perceived naturalness of speech samples created with the speaker adaptation method is judged higher than with the baseline system, speaker-specific features like timbre and pronunciation are preserved, which is reflected in the evaluation of mel-cepstral distortion. This indicates that Viennese Dialect and Standard Austrian German have sufficient phonetic similarities that can be exploited in the context of a speaker adaptation method. Since it is difficult and time-consuming to collect good recordings of dialect speakers to build a corpus usable for speech synthesis, this method offers an interesting approach to improve the audio quality of synthesized samples for small corpora. Fine-tuning the average voice model to the Standard Austrian German speaker does not increase the quality noticeably.

In a separate attempt to increase synthesis quality, multi-task learning is applied in a speech synthesis setting using pole-zero features as a secondary learning task. Listening tests show no improvement in perceived naturalness for this method compared to the baseline system.

References

[1] WU, Z., O. WATTS, and S. KING: *Merlin: An open source neural network speech synthesis system*. In *SSW*. 2016.

[2] SHEN, J., R. PANG, R. J. WEISS, M. SCHUSTER, N. JAITLY, Z. YANG, Z. CHEN, Y. ZHANG, Y. WANG, R. SKERRV-RYAN, R. A. SAUROUS, Y. AGIOMVRGIANNAKIS, and Y. WU: *Natural tts synthesis by conditioning wavenet on mel spectrogram predictions*. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4779–4783. 2018.

[3] WU, Z., P. SWIETOJANSKI, C. VEAUX, S. RENALS, and S. KING: *A study of speaker adaptation for dnn-based speech synthesis*. In *INTERSPEECH*. 2015.

[4] ENZINGER, E., P. BALAZS, D. MARELLI, and T. BECKER: *A logarithmic based pole-zero vocal tract model estimation for speaker verification*. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4820–4823. 2011.

[5] SELTZER, M. L. and J. DROPPO: *Multi-task learning in deep neural networks for improved phoneme recognition*. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6965–6969. 2013.

[6] WU, Z., C. VALENTINI-BOTINHAO, O. WATTS, and S. KING: *Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis*. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4460–4464. 2015.

[7] SHI, L., J. K. NIELSEN, J. R. JENSEN, and M. G. CHRISTENSEN: *A variational em method for pole-zero modeling of speech with mixed block sparse and gaussian excitation*. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pp. 1784–1788. 2017.

[8] MARELLI, D. and P. BALAZS: *On pole-zero model estimation methods minimizing a logarithmic criterion for speech analysis*. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(2), pp. 237–248, 2010.

[9] CHOLLET, F.: *Keras*. `https://keras.io`, 2015.

[10] SCHOEFFLER, M., S. BARTOSCHEK, F.-R. STÖTER, M. ROESS, S. WESTPHAL, B. EDLER, and J. HERRE: *webmushra — a comprehensive framework for web-based listening tests*. *Journal of Open Research Software*, 6, 2018.