SPEAKER INTERPOLATION BASED DATA AUGMENTATION FOR AUTOMATIC SPEECH RECOGNITION

Lisa Kerle[1], Michael Pucher[2], Barbara Schuppler[3]

[1,2,3]Signal Processing and Speech Communication Laboratory, Graz University of Technology

[2]Austrian Research Institute for Artificial Intelligence

[1]lisa-kerle@gmx.at, [2]michael.pucher@tugraz.at, [3]b.schuppler@tugraz.at

ABSTRACT

In recent years, the development of automatic speech recognition systems has ensured their widespread use in a broad range of areas. Most of these systems, however, require large amounts of training data, making them less suitable for low-resourced languages and for smaller varieties of (well-resourced) languages. This paper focuses on improving automatic speech recognition for Austrian German by means of training data augmentation through neural network-based text-to-speech synthesis. For this purpose, speaker embedding vectors are extracted from an existing corpus and subsequent interpolation between these vectors is used for the generation of new voices. Synthesised speech is then used to train an automatic speech recognition system, while comparing differently large portions of synthesised speech in the training data. Overall, we find that performance improves when the ratio of real and synthesised speech is in the same order of magnitude.

Keywords: speech synthesis, automatic speech recognition, data augmentation, Austrian German

1. INTRODUCTION

Given the increasing performance of automatic speech recognition (ASR) systems, they are also used more broadly in everyday life. Modern ASR systems are based on neural networks (e.g., [1, 2]), requiring thousands of hours of speech, ideally covering a high degree of acoustic diversity [3]. When operating in a low-resource scenario, including ASR for smaller varieties of (well-resourced) languages, or for less resourced speaking styles, large performance drops have been reported [4].

In order to improve ASR performance, different approaches on data augmentation have been investigated. One method is text augmentation by means of the mix-up method, which describes the swapping of words between two or more sentences [5]. Another method is to acoustically augment the training data, for which Wang et al. [6] modified speaker characteristics using a non-autoregressive, non-parallel voice conversion model. To improve robustness in ASR training, Lam et al. [7] applied aligned data augmentation, that uses the replacement of certain tokens in an original audio-text pair and the corresponding adjustment of the audio representations.

Parallel to the development of ASR systems, also speech synthesis has advanced by integrating neural networks, leading to the generation of more naturally-sounding, human-like speech. State-of-the-art speech synthesis systems at an end-to-end level (e.g., Tacotron2) generate speech with high naturalness based on a sequence-to-sequence prediction network with attention mechanism [8, 9]. However, more conventional systems using deep neural networks for acoustic modeling (e.g., feed-forward deep neural network (FFDNN)-based synthesis) also produce intelligible speech at high quality [10].

Previous research has shown the successful combination of speech synthesis and ASR: In [3], speech is synthesised with randomly selected voice profiles and used in addition to real speech data for training an ASR system resulting in an improved ASR performance of 12.5% WER compared to speech recognition using real speech only. [11] generated speaker information by sampling from observed speaker representations to increase the speaker diversity of a training set of an ASR system and showed the effectiveness of augmenting real training data with synthesised training data (4% WER improvement).

This paper aims at investigating whether the augmentation of training data for an ASR system by means of synthesised voices improves its performance for Austrian German, a low-resourced variety of German and contains a variation of the work described in [12]. For this purpose, we generate new speaker characteristics by applying linear interpolation to speakers from an already existing corpus. The generated speaker information is passed to a speech synthesis system in order to generate speech

with new voice characteristics. Synthesised speech is subsequently used to train the ASR system, while comparing differently large portions of synthesised speech in the training data. In all experiments, we use the text from the original training, thus keeping the lexical variety equal across all experiments.

2. MATERIALS

Experiments were based on two corpora of read Austrian German. The first one, a corpus created for speech synthesis experiments, was the Wiener Corpus of Austrian Varieties for Speech Synthesis (WASS) with read speech from a total of 19 speakers of standard Austrian German (6f, 13m) [13, 14, 15]. The reading material contains, among others, sentences from the Berlin-Marburg corpus and the Kiel corpus, resulting in a total of 8293 utterances. Approx. half of the material was read by the same professional male. Second, we used the Graz corpus of Read And Spontaneous Speech (GRASS) [16, 17], which contains a total of approx. 30h of speech from 38 speakers of similar social but different regional backgrounds. In this work, we only used the read speech component of GRASS. Table 1 provides an overview of the data used in the experiments of this paper. One main difference between the WASS and the GRASS corpus is that whereas in WASS a small number of speakers produced a large number of sentences, in GRASS a larger number of speakers read a smaller number of sentences.

set	speakers		utterances	duration (s)
	m	f		
WASS				
train	10	3	6955	20483
dev	1	1	446	1433
test	1	1	446	1395
GRASS				
train	17	16	3820	19914
dev	1	1	249	1329
test	1	1	254	1157

Table 1: Training, development and test set of the WASS and GRASS corpus, with the corresponding number of utterances and duration in seconds.

3. GENERATING SYNTHESISED VOICES

3.1. Method

We used the feed-forward deep neural network speech synthesis system Merlin [10, 18]. To generate speech with specific speaker characteristics, we passed speaker information in form of speaker embedding vectors to the speech synthesis system in addition to linguistic features (i.e., quinphone, syllable, word, phrase).

We generated speaker embedding vectors from the WASS corpus using a Pytorch-Kaldi based speaker recognition system [19, 20]. The system employs a learnable dictionary encoding layer as pooling layer, after which the speaker information is extracted in form of speaker embedding vectors [21]. The quality of the speaker embedding vectors depends on the *minimum required frame length*, the *chunk size* and the *speaker embedding dimension*. First, we chose these parameters in accordance with previous work [20]. Next, we tuned these parameters by considering the *equal error rate* (EER) and the *minimum detection cost function* (minDCF) of the speaker recognition system, and additionally the visual representation of the extracted speaker embedding vectors using the *t-distributed stochastic neighbour embedding* (t-SNE). The parameter combination leading to the lowest EER (6.05%) and minDCF (0.27) and a good separation in terms of the t-SNE was used to extract the speaker embedding vectors from the WASS corpus.

For generating new speaker characteristics, we linearly interpolated between the extracted speaker embedding vectors according to:

$$v_{12} = \alpha \cdot v_1 + (1-\alpha) \cdot v_2,$$

where v_1 and v_2 describe the embedding of the original speakers and by choosing an interpolation factor $\alpha \in \{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$, a vector v_{12} containing new speaker characteristics is obtained. We then passed the generated speaker embedding vectors to a previously trained average voice model (AVM), in addition to linguistic input features to generate speech with new speaker characteristics. We synthesised two different data sets: *Wsyn* and *Gsyn*. For *Wsyn*, text and interpolated speaker embeddings were based on the WASS corpus. For *Gsyn*, we used text from the GRASS corpus and speaker information was obtained from the interpolated speaker embedding vectors from the WASS corpus. We evaluated the quality of the interpolated speaker embedding vectors by passing utterances of *Gsyn* to the speaker recognition system and considering the EER and the minDCF. Additionally, the extracted speaker embedding vectors were analysed by means of the t-SNE. For example, speech synthesised by interpolation between speakers *kep* and *lsc* led to an EER of 16.05% and a minDCF of 0.53.

Figure 1 shows the t-SNE of the interpolated

speaker embedding vectors. First, speaker embeddings were interpolated, then the interpolated embeddings were used to generate synthetic test data with the AVM, and finally the dimension of the utterance embeddings was reduced using t-SNE. Figure 1 shows that the linear interpolation of speaker embeddings leads to a non-linear but continuous behaviour in the space of utterance embeddings, allowing us to create a continuum of speakers using this DNN architecture. We also performed experiments with attention-based recurrent DNN architectures that did not show such a continuous behavior in the interpolated utterance space. A further qualitative evaluation of the interpolated embedding vectors revealed that the transition from speaker *kep* to *lsc* was clearly perceptible.

Figure 1: t-SNE for utterances of *Gsyn* for speaker combinations of *kep* and *lsc* using different interpolation factors α.

4. ASR RESULTS AND DISCUSSION

4.1. Methods

For ASR, we used Kaldi, a widely used state-of-the-art speech recognition toolkit [22]. In this paper, we applied a recipe described in [23], which follows the conventional GMM/HMM approach and consists of an acoustic model, a language model and a lexicon. The acoustic model includes monophone training as well as triphone training and is based on 13-dimensional MFCCs, which we normalised using cepstral mean and variance normalisation (CMVN). The lexicon was created using a G2P online tool [24] for German German. In order to adapt this lexicon for Austrian German, some adjustments were made in the used recipe (e.g., devoicing all alveolar and postalveolar fricatives and affricates), resulting in a set of 38 phones. The language model based on the SRILM toolkit [25] used a N-gram model with order $N = 3$.

4.2. Baseline Experiments

As baseline, we used the original WASS and GRASS corpus for training, developing and testing. With both training sets, we achieved good results, i.e., a WER below 1% (cf. Table 2), whereby the ASR performed better using the WASS corpus. Although the GRASS corpus had a larger number of speakers, the hours of training data was almost equal for both corpora (cf. Table 1). The better performance of the ASR system on WASS than on GRASS may be explained by the fact that one professional speaker contributed a large proportion of the speech in WASS, i.e., about half of all utterances.

4.3. Corpus Augmentation using Synthesised Speech

We used the synthesised *Wsyn* and *Gsyn* data to augment the WASS and GRASS corpus for training the ASR system. The development and test sets remained the same as in the baseline experiment. We experimented with different amounts of synthesised utterances to find the "optimal" ratio of original and synthesised data. We successively augmented the original GRASS corpus by means of utterances from the synthesised GRASS corpus, for which the synthesised utterances were selected randomly. The amount of synthesised utterances ranged from 0% (original utterances only) to 100% (synthesised utterances only), with tighter steps in the area around 50%. To compare our results with findings from earlier studies [11], we calculated the percentage of synthesized utterances and experimented also with exclusively synthesized speech. Figure 2 shows that in the area of around 50% (i.e., when the amount of original utterances equals the amount of synthesised utterances) best results were obtained and that the performance outperformed that of the baseline experiment (WER of 0.88% is reduced to 0.64%). We made the same observations for the experiments with the WASS corpus, where at a 50% mix of real and augmented data the baseline WER of 0.55% was reduced to 0.33% (cf. Table 2). The augmentation for the WASS corpus thus leads to larger improvements than for the GRASS corpus (i.e., the augmentation is more successful if the interpolated speaker information originates from the same corpus).

To investigate which amount of synthesised data needs to be added to a training corpus to improve ASR, Rosenberg et al. [11] used a fixed amount of synthesised speech and reduced the amount of utterances from the original corpus. They report the

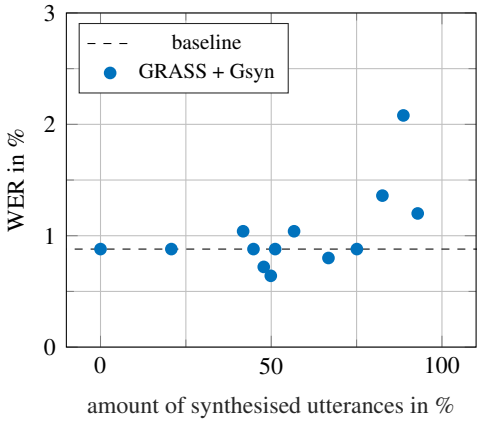

Figure 2: WER in % for different amounts of synthesised utterances of the GRASS corpus. The dashed line shows the baseline experiment.

highest performance when the amount of synthesised and original utterances were equal (at 50%), which is consistent with our observations.

4.4. Using the Synthesised Corpus for Training

We analysed the performance of the ASR system when training it on synthesised speech only (i.e., *Wsyn* and *Gsyn*). The development and test sets were the same as for the baseline experiment. Our experiments showed that ASR performance degrades drastically for *Wsyn* and for *Gsyn*, ASR was almost impossible (cf. Table 2). The poor performance for *Gsyn* might reflect the mismatch between the speakers of the training set (*Gsyn*) and the speakers of the development and test set (GRASS), as the speakers of *Gsyn* resulted from the interpolation between speakers from the WASS corpus and development and test set from speakers of the GRASS corpus. When using *Wsyn* for training, where both text and speaker information come from the same corpus (WASS) results in better ASR performance than *Gsyn*. These results indicate that synthesised speech performs better as ASR training set when generated with text and speaker information from the same corpus. The drastic degradation in ASR performance when using only synthesised material for training was observed in previous work from [11] (32.44% WER).

data		results	
train	**dev/test**	**WER**	**SER**
WASS	WASS	0.55	2.69
GRASS	GRASS	0.88	3.54
WASS + *Wsyn*	WASS	0.33	1.57
GRASS + *Gsyn*	GRASS	0.64	2.36
Wsyn	WASS	13.85	36.55
Gsyn	GRASS	72.92	95.03

Table 2: ASR results in % WER and SER, for the original training set (WASS, GRASS), the augmented training corpus with 50% synthesised speech (WASS + *Wsyn*, GRASS + *Gsyn*) and the synthesised training corpus (*Wsyn*, *Gsyn*).

5. CONCLUSION

The aim of this paper was to investigate whether augmentation of training data by means of synthesised voices improves ASR performance for read Austrian German, a variety of German lacking large amounts of annotated speech resources. We used two corpora of read Austrian German, the WASS and the GRASS corpus, which were created having different applications in mind. WASS, created for speech synthesis applications, contains a large amount of sentences read by few (partly professionally trained) speakers. GRASS, created for ASR, contains fewer different sentences, read by a larger number of speakers. For both corpora, the baseline ASR experiments (i.e., trained on the original GRASS or WASS data and tested on GRASS or WASS respectively) showed good results (WER below 1%). Next, we trained the ASR system with synthesised utterances only. For both WASS and GRASS, the WER degraded drastically, for the WASS corpus to 13.85% WER and for the GRASS corpus to 72.92%, meaning that ASR was basically impossible. Finally, we augmented the original training data with synthesised utterances, which improved the performance of the ASR system for both corpora, even more so when the amount of original utterances was in the order of magnitude of the synthesised utterances, i.e., 0.33% WER for WASS and 0.64% for GRASS. We observed a higher improvement through data augmentation for the WASS corpus, which is the data constellation where the linguistic information and the speaker information originated from the same corpus.

The experiments of this paper dealt with read speech only. As a next step, we will explore synthesis-based data augmentation for conversational speech, where ASR is even more challenging and data sparsity is even more of an issue (e.g., [26]).

6. ACKNOWLEDGEMENTS

This work was supported by grant P-32700-N and I2539-G23 from the Austrian Science Fund (FWF). We thank Julian Linke for his help with the KALDI recipes for GRASS.

7. REFERENCES

[1] Baevski, A., Zhou, Y., Mohamed, A., Auli, M. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems* volume 33, 12449–12460.

[2] Luo, H., Zhang, S., Lei, M., Xie, L. 2021. Simplified self-attention for transformer-based end-to-end speech recognition. *2021 IEEE Spoken Language Technology Workshop (SLT)*, 75–81.

[3] Fazel, A., Yang, W., Liu, Y., Barra-Chicote, R., Meng, Y., Maas, R., Droppo, J. 2021. SynthASR: Unlocking synthetic data for speech recognition. *Proceedings of Interspeech*, 896–900.

[4] Linke, J., Garner, P. N., Kubin, G., Schuppler, B. 2022. Conversational speech recognition needs data? Experiments with Austrian German. *Proceedings of LREC*, 4684–4691.

[5] Kwon, S., Lee, Y. Apr 2022. Explainability-Based Mix-up Approach for Text Data Augmentation. *ACM Transactions on Knowledge Discovery from Data*. Just Accepted.

[6] Wang, G., Rosenberg, A., Ramabhadran, B., Biadsy, F., Emond, J., Huang, Y., Moreno, P. J. 2022. Non-Parallel Voice Conversion for ASR Augmentation. *Proceedings of Interspeech*, 3408–3412.

[7] Lam, T. K., Ohta, M., Schamoni, S., Riezler, S. 2021. On-the-Fly Aligned Data Augmentation for Sequence-to-Sequence ASR. *Proceedings of Interspeech*, 1299–1303.

[8] Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R. J., Saurous, R. A., Agiomyrgiannakis, Y., Wu, Y. 2018. Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions. *Proceedings of ICASSP*, 4779–4783.

[9] Lim, D., Jung, S., Kim, E. 2022. JETS: Jointly Training FastSpeech2 and HiFi-GAN for End to End Text to Speech. Ko, H., Hansen, J. H. L. (eds), *Proceedings of Interspeech*, 21–25.

[10] Wu, Z., Watts, O., King, S. 2016. Merlin: An Open Source Neural Network Speech Synthesis System. *Proceedings of 9th ISCA Workshop on Speech Synthesis Workshop*, 202–207.

[11] Rosenberg, A., Zhang, Y., Ramabhadran, B., Jia, Y., Moreno, P., Wu, Y., Wu, Z. 2019. Speech Recognition with Augmented Synthesized Speech. *Proceedings of ASRU*, 996–1002.

[12] Kerle, L. K. 2022. Speaker Interpolation based Data Augmentation for Automatic Speech Recognition. Master's thesis Graz University of Technology.

[13] Pucher, M., Toman, M., Schabus, D., Valentini-Botinhao, C., Yamagishi, J., Zillinger, B., Schmid, E. 2015. Influence of speaker familiarity on blind and visually impaired children's perception of synthetic voices in audio games. *Proceedings of Interspeech*, 1625–1629.

[14] Toman, M., Pucher, M. 2015. An Open Source Speech Synthesis Frontend for HTS. *Proceedings of the 18th International Conference on Text, Speech, and Dialogue - Volume 9302*, 291–298.

[15] Pucher, M., Rausch-Supola, M., Moosmueller, S., Toman, M., Schabus, D., Neubarth, F. 2016. Open data for speech synthesis of Austrian German language varieties. *12. Tagung Phonetik und Phonologie im deutschsprachigen Raum*, 147–150.

[16] Schuppler, B., Hagmüller, M., Morales-Cordovilla, J. A., Pessentheiner, H. 2014. GRASS: the Graz corpus of Read And Spontaneous Speech. *Proceedings of LREC*, 1465–1470.

[17] Schuppler, B., Hagmüller, M., Zahrer, A. 2017. A corpus of read and conversational Austrian German. *Speech Communication* 94C, 62–74.

[18] Ronanki, S., Wu, Z., Watts, O., King, S. 2016. A Demonstration of the Merlin Open Source Neural Network Speech Synthesis System. *Proceedings of 9th ISCA Workshop on Speech Synthesis Workshop*, 124–124.

[19] Villalba, J., Chen, N., Snyder, D., Garcia-Romero, D., McCree, A., Sell, G., Borgstrom, J., García-Perera, L. P., Richardson, F., Dehak, R., others, 2019. State-of-the-art speaker recognition with neural network embeddings in NIST SRE18 and speakers in the wild evaluations. *Computer Speech & Language* 101026.

[20] Cooper, E., Lai, C.-I., Yasuda, Y., Fang, F., Wang, X., Chen, N., Yamagishi, J. 2020. Zero-Shot Multi-Speaker Text-To-Speech with State-of-the-art Neural Speaker Embeddings. *Proceedings of ICASSP*, 6184–6188.

[21] Cai, W., Chen, J., Li, M. 2018. Exploring the Encoding Layer and Loss Function in End-to-End Speaker and Language Recognition System. *The Speaker and Language Recognition Workshop*, 74–81.

[22] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Vesely, K. Dec. 2011. The Kaldi Speech Recognition Toolkit. *Proceedings of ASRU*.

[23] Linke, J., Wepner, S., Kubin, G., Schuppler, B. 2023. Using Kaldi for automatic speech recognition of conversational Austrian German. *url: http://arxiv.org/abs/2301.06475v1*.

[24] Reichel, U. D. 2012. PermA and Balloon: Tools for string alignment and text processing. *Proceedings of Interspeech*, 1874–1877.

[25] Stolcke, A. 2002. SRILM - an extensible language modeling toolkit. *Proceedings of Interspeech*, 901–904.

[26] Wepner, S., Schuppler, B., Kubin, G. 2022. How prosody affects ASR performance in conversational Austrian German. *Speech Prosody 2022*, 195–199.