Deep speech denoising by ASR-TTS resynthesis

Iason Hartmann
Chair of Integrated Digital Systems and Circuit Design, RWTH Aachen University
Aachen, Germany
Email: iason.hartmann@rwth-aachen.de

***Abstract*—This written report presents a novel approach, which is proposed as a solution for the Intel Neuromorphic Deep Noise Suppression Challenge (Intel N-DNS Challenge) [1]. We identify the two major approaches: The direct denoising by training a network on noisy and clean waveforms on the one hand [2] and on the other hand an encoder-decoder structure by resynthesis, which extracts the text from the original noisy speech (encoder, ASR) and generates speaker dependent clean speech from the text (decoder, TTS) [3], [4], [5]. Due to the mismatch of synthetic and original audio, a joint network trained on the noisy and synthetic audio mel spectra is developed to align the audios and compensate for wrong generated words due to failed recognition by the encoder as the main challenge of this work. Denoising by parametric resynthetis is able to generate high MOS scores of 2.17, 2.72, 2.97 for OVRL, SIG and BAK. MixNet can decrease the MSE loss compared to synthetic mel spectrogram (mel spec) while decreasing SI-SNR compared to noisy audio. The vocoder is an important factor for speech quality and reduces the alignment significantly.**

***Index Terms*—Speech enhancement, deep audio denoising, parametric resynthesis, text to speech, multi-speaker, Transformer, audio mixing deep neural networks, joint networks, alignment networks**

I. Introduction

DEEP audio denoising is an application of deep learning to signal processing which can exploit the benefits of neuromorphic computing. There exist straight-forward solutions with LSTM networks like the CRNN architecture [2] to extract the features with convolutions and time-dependent information with Long-Short-Term-Memory (LSTM) cells and process the signal. Nevertheless, novel approaches like [7], [3], [4], [6] utilized an encoder-decoder apporach to generate clean audio by complete resynthesis. We pick up this approach in this work and present an encoder-decoder architecture by extracting the text from the noisy speech and resynthesizing the speakers voice, speaking the extracted text from the encoder. The encoder is an automatic speech recognition (ASR) system, the decoder is a speaker dependent text-to-speech (TTS) system, which an generate arbitrary speech from text and a reference audio. The challenges are the high noise sensitivity of ASR systems and therefore high word error rates (WER) in the presence of noise and the temporal alignment of the generated clean speech to the original noisy one for a high signal accuracy according to the SI-SNR metric of the Intel N-DNS challenge. We estimate the quality of such an encoder-decoder architecture by evaluating the state-of-art techniques and compare different approaches to address these challenges. Based on existing solutions, we develop the alignment system *MixNet*, which aims to align the synthetic signal to the noisy one while taking the WER of the first into account and preserving high speech quality.

II. Background

A. Automatic Speech Recognition (ASR)

The task of ASR systems is to convert an input speech to text. Different architectures exist implementing ASR like LSTM [8], Transformer [9] based architectures and recently the conformer [10], [11]. In comparison to the conventional transformer, the conformer can predict the text based on clean speech from the LibriSpeech dataset with a reduced WER of 2.1 % vs 2.69 %. Due to the convolution blocks, the conformer extracts the most important features and further can be scaled down from 19M params of the transformer to 10.3M params in the conformer. However, echo, background noise and competing speech, significantly decreases ASR performance [11]. The cleanformer [12] addresses the problem of noise sensitivity of ASR. Nevertheless, for a SNR of -5dB, the WER of the cleanformer still remains at at least 12.5 %. Another ASR system developed by the Machine Learning and Human Language Technology institute of RWTH Aachen University is "RASR2: The RWTH ASR Toolkit for Generic Sequence-to-sequence Speech Recognition" [13]. It consists of the input feature block, which can perform several transformations of the raw input audio, f.e. mel spectra or mel frequency cepstral coefficients (MFCC) features and a lexicon, which generates based on the input speech the text. The lexicon also utilized the conformer architecture for its encoder. On the LibriSpeech dataset the WER ranges from 11.9% to 4.0% on the test-other set.

B. Text-to-Speech (TTS)

The task of TTS systems is to generate clean speech from a text. For transformer-based [9] training the text and speech data are usually of high quality. To achieve a speaker-dependent output, Multi-Speaker-TTS have been implemented, by training on a speaker embedding [15]. In MultiSpeech TTS [14], the model has in total 1857777 trainable parameters. It was trained on mel spectrogram (mel spec) as output and phonemes as input with WaveNet [16] was used as Vocoder. On the LibriSpeech dataset, Mean Opinion Scores (MOS) of 2.95 ± 0.14 compared to the ground truth signal with 4.04 ± 0.16 are achieved with this method. YourTTS [17] is a multilingual, multi-speaker text-to-speech model trained among others on the LibriTTS partitions train-clean-100 and

train-clean-360. YourTTS takes contrary to MultiSpeech raw text as input, to allow more realistic results for languages without good open-source grapheme-to-phoneme converters available. The text encoder is based on transformer. YourTTS does not train HiFIGan vocoder and transformer separately as usual but jointly to enable high voice similarity of the speaker and the synthesized waveform. Upon evaluation on the LibriSpeech dataset, YourTTS achieves a MOS-score of 4.18 ± 0.05 [17].

C. Speech denoising by resynthesis

Several works in resynthesis of speech have been carried out, some for noise supression [3], [4] as well as for general speech synthesis for arbitrary speakers [6]. In [3] a parametric synthesis system for denoising speech is proposed, which avoids a large and memory expensive speech inventory. The system consist of a prediction model, which is a deep neural network (DNN) supplemented with LSTMs to extract the temporal context of the data. It takes as input several acoustic features. The loss function is the mean squared error (MSE) loss between prediction and ground truth. Upon evaluation, the prediction model achieves a Perceptual Evaluation of Speech Quality (PESQ) with 2.43. In [4], the work is extended by training the WaveNet jointly with the predicition model to generate clean waveform samples.

In [6], the resynthesis-architecture achieves on the LibriSpeech dataset speech natrualness MOS scores of 3.98 ± 0.06 on seen and 4.12 ± 0.05 on unseen speakers compared to 4.49 ± 0.05 and 4.42 ± 0.07 for the ground truth audio. Regarding speaker similarity it achieves similarity MOS scores of 3.03 ± 0.09 3.28 ± 0.08 compared to 4.38 ± 0.08 on the ground truth audio. In [5], so-called "text informed speech enhancement" is utilized to enhance noisy speech with the guide of extracted speech features from text: Mel spectra can be enhanced by first looking for speech units in a lexicon that best match to the underlying clean speech components in the target noisy speech. In [5], matching is guided by the extracted speech information from the text (this can be done by f.e. a TTS system) instead of a lexicon. The text features are time-aligned to the speech features, so that a DNN can learn the correct time dependent mapping to enhanced speech features. Denoising by resynthesis outperforms straight-forward speech denoising with its high PESQ-scores of ¿2.43 (up to 4.18) compared to 2.44 as highest achieved PESQ score by [2].

D. Intel N-DNS evaluation metric

In the following the evaluation metric is presented, based on which the denoising model should be optimized and finally assessed. For more detailed information regarding the challenge we refer to [1].

1) SI-SNR metric: The evaluation of the N-DNS Challenge is performed thorugh audio quality measurement. One metric which specifically refers to noise is the Scale-Invariant Source-to-Noise Ratio (SI-SNR)—SI-SNR. For a single input waveform s_{input}, a real-valued zero-mean vectors, and the corresponding output waveform from the denoising model $s_{predict}$, the SI-SNR is defined as $SI-SNR = log(\frac{||s_{target}||^2}{e_{noise}})$, where $s_{target} = \frac{<s_{noisy}, s_{input}>}{||s_{input}||^2}$ and $e_{noise} = s_{predict} - s_{target}$. This means that if the predicted signal differs highly from the target signal f.e. regarding speed or voice (not volume), the quality is labeled with a high SI-SNR value despite haven clear, high-quality and noise-free speech. Additionally, two measures of audio quality (SI-SNR) improvement are defined

$$SI-SNRi_{data} = SI-SNR_{fs} - SI-SNR_{data} > 3dB \quad (1)$$

and

$$SI-SNRi_{enc+dec} = SI-SNR_{fs} - SI-SNR_{enc+dec} > 3dB \quad (2)$$

with $SI-SNR_{fs}$ being the mean test-set SI-SNR from the full model (encoder, denoiser, decoder) and $SI-SNR_{enc+dec}$ the mean test-set SI-SNR from running only encoder and decoder (encoder, decoder) and $SI-SNR_{data}$ is the mean test-set SI-SNR on the noisy input audio without any transformations.

2) DNSMOS metric: Another important metric in addition to the signal quality measured by SI-SNR is the widely adopted DNSMOS metric to evaluate the perceptual quality of the audio. In DNSMOS from the Microsoft DNS challenge[1], the perceptual quality score is predicted by a deep network that is trained to estimate the human perceptual quality by returning the Mean Opinion Score (MOS) of the input audio. The MOS score ranges from 1 to 5, where 1 corresponds to poor quality, and 5 corresponds to excellent quality.

III. Proposed Approach

Based on the in section II laid foundation, we present our deep denoising model by resynthesis: It takes as input a noisy raw audio waveform and outputs a noise free waveform, which should be similar to the speech without the background noise and consists of three main components: (1) The ASR system, (Encoder), which converts the input waveform with additive noise to a text or phoneme sequence. Due to the noise sensitivity of ASR, it is critical to perform pre-enhancement of the waveform to minimize the WER. Cleanformer [12] could be a suitable approach which can reduce the WER to a maximium of 15 %, (2) a multi-speaker TTS system (Decoder), which can generate based on the noisy audio waveform and the from the ASR model extracted text clean a noise-free mel spec and (3) a joint network, which aligns the artificial synthesized mel spec with respect to the original noisy speech. In addition, a vocoder is necessary to recover the synthesized and aligned waveform from the generated and process mel spec. In section II-C it was already shown by other authors that by resynthesizing speech over a Encoder-Decoder structure is a promising approach for speech enhancement. Nevertheless due to the SNR metric in the Intel N-DNS challenge, which compares the original noise free and recovered waveform, it is cruicial, that the generated waveform and the original clean waveform are aligned regarding speed, duration and voice, which makes

[1]https://github.com/microsoft/DNS-Challenge

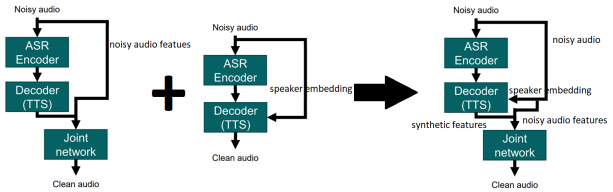

Fig. 1. The two possible models for denoising by speech resynthesis

the addition of a joint alignment network, which mixes the synthetic and noisy reference speech as it is presented in [5]. This paper also shows that one can generate from text and noisy speech enhanced speech, i.e. speech information can be encoded in text (here: ASR) and decoded by text feature extraction (here: TTS) Figure 1 shows three two proposed approaches: One with the ASR Encoder-Decoder structure with a joint network to ensure the alignment of the original and synthetic speech on the left, the simple resynthesis py passing speaker information to the TTS and resynthesis without any alignment in the middel and the complete architecture which unites both approaches on the right.

A. ASR-TTS Encoder-Decoder

For the ASR system, it is necessary to have high noise robustness and a low WER. Based on the literature from [12], [13], we can expect a WER of approximately 10 %. For the Decoder, we use the pretrained YourTTS [17] model from the Coqui-AI TTS library[2]. Now, the reason for our demand for low WER is that by the encoder wrong predicted words propagate through the decoder and generate a signal that differs in the interval with the wrong words resulting in a high SI-SNR. However, the MOS scores should be preserved and may even increase due to the high quality of synthesized speech by multi-speaker TTS systems ([5]). Consequently, we presuppose for the TTS-Decoder a) a high quality of the output speech due to resynthesis under b) a high similarity with the original audio, despite having noise. We feed the TTS parallel with the text from the encoder with the noisy input audio such that the TTS can extract the speaker embedding and resynthesize the speech spoken by the speaker from the reference audio (which is the noisy input audio to be denoised) without noise. However, due to the because state-of-art ASR systems still have WERs about 10 % in the best-case under noisy conditions and thus exhibit a low noise robustness and secondly, other speakers attributes like speed, pitch and pronunciation are still neglected, some alignment of the synthetic and noisy speech may still be necessary as presented in section II-C. This alignment is done by our implementation of the Joint Network *MixNet*, see section IV-D.

B. Joint-Network MixNet

MixNet is a fully connected convolutional deep neural network (FCDNN), whose hyperparameters are inspired by [5]: We stack hidden layers with 2048 neurons each. The synthetic waveform (generated by the TTS-decoder) and the noisy waveform is padded to a fixed number of frames, such that MixNet receives a input with fixed dimensions. This number is determined by the longest audio sample generated by the decoder and from the dataset. Then, from the synthetic and noisy waveform, the mel spec is computed and they are concatenated along an additional dimension. Three convolutions are applied to the resulting mel spec of size $N_{frames} * N_{mels}$ to reduce the number of input features for the FCDNN and reduce the number of parameters. The output of the convolution is flattened and fed into the fully connected network. The three convolutions map from 2 channels (the synthetic and noisy mel specs) to 16, 16 to 32 and 32 to 64 respectively with are kernel size of 3x3 and stride of 1x1. After each convolution layer, batch normalization, ELU-activation and 2x2 max pooling, except for the last one, is performed. Now, instead of training the FCDNN on clean specs we let it compute a mask, which performs a linear transformation on the chunks (the chunk size is set as a hyperparameter) of the input mel spec along the frame domain. The number and size of surrounding chunks which are considered for the mask computation for each chunk can be set and depends on how much the mel specs differ along the different chunks. This mask predicted by the FCDNN is applied on the vector of the chunks of mel spec and the resulting mel spec prediction by MixNet is obtained. The loss of prediction and clean mel spec is then backpropagated through the FCDNN. As loss function we use $l = MSE(mel_{out}, mel_{clean}) + 0.5 * \sum_{chunks} MSE(mel_{out}, mel_n) * MSE(mel_{dec}, mel_n) + 2 * MSE(mel_{out}, mel_{dec})$. MSE is the mean squared loss, mel_{out}, mel_{clean}, mel_n, mel_{dec} are the mels of MixNet, the clean and noisy audio and the decoder, respectively. The second term should penalize the model, when it prefers the noisy input if the encoder audio is similar to the noisy audio for the corresponding chunk. The third term should prefer the clean synthetic over the noisy original audio.[3]

IV. Experiments and Results

In this section, we conduct experiments to estimate the quality of our architecture for speech denoising.

A. Experimental setup

Dataset. We conducted experiments on the LibriSpeech train-clean-100 test-clean dataset. The samples were sampled at at 16 kHz. We trained on a subset of the dataset containing 512 samples each shorter than 5s. We use 10ms hop size and 25ms window size to extract mel spec with 64 mel features from all of our used waveform data (the samples from the datset and those synthesized by the decoder). We simulate the ASR with a WER of 10 % by taking the spoken text given in LibriSpeech by replacing each word with a 10 % chance by a random word from a fixed dictionary of words. Furthermore, we simulate noise by adding noise from background voices with a SNR of 0 dB to the clean audio from the dataset.

[2]https://github.com/coqui-ai/TTS

[3]MixNet is published under https://github.com/iasone99/MixNet and integrated into https://github.com/thebarnable/IntelNeuromorphicDNSChallenge

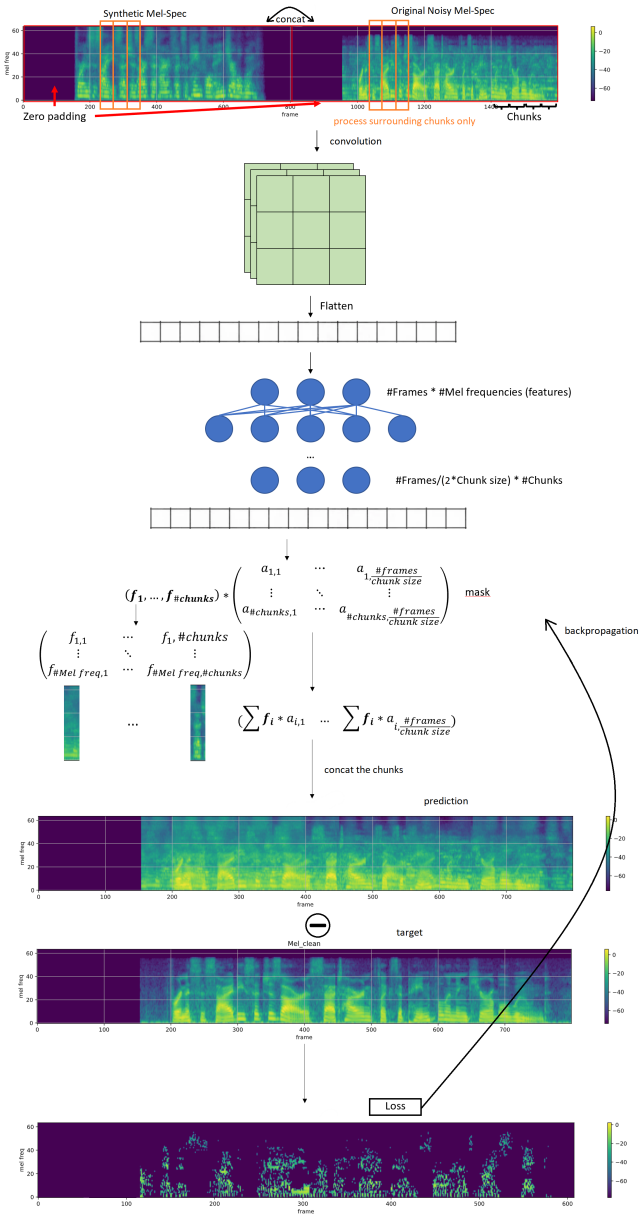

Fig. 2. The structure of the Joint Network MixNet

Model Configuration. We feed the decoder with the text and reference audio and train MixNet on the mel spec of the noisy input waveform according to and the synthetic one from TTS. The latter was stretched and shifted to the length of the noisy audio as a simple alignment. For our FCDNN, we stack 5 fully connected hidden layers with a hidden size of 2048 with leaky ReLU and batch normalization after each layer except for the last for which we use the abs-function to prevent vanishing gradients problem while still generating positive outputs only. We train MixNet with the loss presented in (model I) and process 4 chunk frames with a size of 20 single frames. This model has 5816296 params. We also train the network without the convolutions (model II) with 25214952 params. In the end, we also train the same network as model II to not predict a mask but instead a whole chunk, similar to [5] (model III) with 27823824 params.

Source/ SI-SNR	MixNet+GriffLim to GT	MixNet+GriffLim to GT+GriffLim	TTS+GriffLim to GT	TTS+GriffLim to GT+GriffLim	TTS to GT	Noisy+GriffLim to GT	Noisy+GriffLim to GT+GriffLim	Noisy to GT
model I	-43.2867 dB	-19.7158 dB	-47.3405 dB	-19.0866 dB	-37.9 dB	-40.9666 dB	-24.0065 dB	0 dB
model II	-46.5782 dB	-22.9126 dB						
model III	-49.5593 dB	-29.4156 dB						

Fig. 3. SI-SNRs for for the two MixNet configurations, TTS decoder, noisy original and ground truth audio with and without GriffinLim as Vocoder

Training and Inference. We use the A10 GPU with a batch size of 32 and Adam optimizer with a learning rate of 0.01. We use linear warmup of the learning rate over the period of one epoch. GriffinLim from the Torchaudio library is used as vocoder to recover the waveform from the mel specs.

Evaluation. We evaluate the SI-SNR from section II-D1 of the synthetic speech generated by the standalone TTS-decoder from the text of the simulated ASR, the noisy waveform and the audio recovered by GriffinLim from the mel prediciton of MixNet related to each other on 10 samples from the LibriSpeech test-clean set. Further, we compare the average SI-SNRs of each waveforms after transformation to mel spec and recovery by GriffinLim to study the impact of the Vocoder. For all of these waveforms, we evaluate the average DNSMOS score from section II-D2 by passing them to the DNSMOS network. We also compute the average MSE loss on the test samples.

B. SI-SNR evaluation

Figure 3 lists the SI-SNR values for the test set. Model I surpasses model II. However, the SI-SNR of MixNet is comparable to the SI-SNR of the decoder. Another observation is the effect of the GriffinLim vocoder, which worsens the performance significantly. While the noisy audio has a SI-SNR of 0 dB as set in the training configuration, after conversion of both, the noisy audio and the ground truth (GT) clean one to mel spec and recovery with GriffinLim, the SI-SNR is reduced by 24 dB. To solve this issue, we propose joint training of MixNet and a vocoder as done in [17], which resulted in high quality speech. Further, we observe that the network tends towards the noisy mel spec, if the synthetic and noisy mel spec are highly different, despite the second term in the loss function. The reason for this is the missing alignment of the synthetic mel: The synthetic mel differs too strongly from the clean one such that the model decides for the noisy input. To resolve this issue, we propose an alignment in the style of Dynamic Time Warping (DTW) such that the network only needs to compare the two mels chunk for chunk and decide for one or another or a superposition of both to generate the to the GT mel most similar output. Model III exhibits the lowest SI-SNR. For the improvement metrics, only model I matches $SI-SNRi_{enc+dec} = 4.3412dB > 3dB$ but only w.r.t. the GT signal after mel transformation and GrifinLim recovery. $SI-SNRi_{data} > 3dB$ is not matched due to the high change by the vocoder.

Source	MixNet+GriffLim I	MixNet+GriffLim II	MixNet+GriffLim III	TTS+GriffLim	TTS	Noisy+GriffLim	Noisy	GT+GriffLim	GT
OVRL	1.1505	1.1519	1.2128	1.1447	2.1662	1.1570	1.0945	1.4945	3.3313
SIG	1.2370	1.2408	1.4102	1.3876	2.7197	1.2534	1.2006	1.8394	3.5969
BAK	1.2478	1.2577	1.473	1.7752	2.9715	1.2822	1.1449	3.0977	4.1288

Fig. 4. DNSMOS for the two MixNet configurations, TTS decoder, noisy original and ground truth audio

Source/MSE Loss	MixNet	TTS	Noisy
model I	0.0318	0.2205	0.0295
model II	0.0295		
model III	1.0056		

Fig. 5. The MSE loss for two configurations of MixNet

C. DNSMOS evaluation

Figure 4 depicts the DNSMOS scores for the two models, the TTS decoder, noisy original and ground truth audio with and without vocoder. The clean waveform generated by the decoder exhibits high DNSMOS scores of 2.1662, 2.7197 and 2.9715 for OVRL, SIG and BAK respectively. The two MixNet models perform similar and lie at around 1.15, 1.24 and 1.25. This shows together with section IV-B that there is a SI-SNR-MOS trade-off because recombination for better alignment harms the speech naturalness due to echos and discontinuities in the mel specs. The MOS scores may increase with joint training of MixNet and the vocoder as done in [17], since after mel conversion and recovery with GriffinLim, the DNSMOS scores of the TTS worsen to 1.1447, 1.3876 and 1.7752. For model III, the DNSMOS scores increase sightly because the discontinuities can be avoided by predicting the mel spec directly.

D. Loss evaluation

Figure 5 lists the MSE loss of the two models and the TTS and noisy mels. As expected the loss of MixNet lies between the one from the noisy and from the TTS, as it takes the best of both worlds to reduce the overall loss function from . Interestingly, despite having a lower MSE, the noisy mel exhibits a lower SI-SNR than the one from MixNet after conversion to waveform. The synthetic audio SI-SNR after alignment is comparable to the one from the model but has a much higher loss. This indicates, that the MSE loss alone is not sufficient as loss function to minimize SI-SNR. We instead propose training directly on SI-SNR jointly with a vocoder. Model III yields the highest loss of all models, mainly because it fails to remove the noise component of the mel spec while maintaining mel similarity: Compared to model III, in model II we do not need to care about the vertical axis of the mel specs and only need to train our model on the time domain. Together with a larger chunk size, we can therefore reduce the problem of generating a complete new audio to the problem of simply permuting respectively superposing two audios by a linear transformation.

V. Conclusion

In this paper, we evaluated the suitability of a deep speech denoising system for the Intel N-DNS challenge via ASR-TTS resynthesis. Resynthesis via multi-speaker TTS is able to generate high quality clean speech while maintaining the speakers voice. On the other hand, due to high WER of ASR in the presence of noise and different speed and pronounciation, the alignment of the synthetic to the noisy/clean audio is the main challenge of this architecture. We presented MixNet which can align the different frames from both mel specs and exposed the vocoder as the limiting factor for alignment in the waveform domian (SI-SNR) as well as speech quality (MOS).

For future works, we propose training the vocoder jointly with the TTS-MixNet architecture and implement training not on mel spectra but on SI-SNR directly for better alignment. Further we suggest that MixNet may learn the alignment more efficiently when the target is no the mel spec itself but the optimal mapping of the chunks. This presupposed, that an optimal mapping is known from the mel specs beforehand. Lastly, we recommend to reduce the task of MixNet further: While it now learns discrimination of correct and wrongs words as well as the alignment of the chunks, we recommend to perform preprocessing of the mel spec via stretching and shifting in the style of DTW, let it process only the respective chunks from both audios itself instead of additionally the surrounding ones and train on discrimination between the two chunk candidates.

Acknowledgments

This work was supported by Institute of Integrated Digital Systems, RWTH Aachen University. The authors thank Tobias Gemmeke, Malte Wabnitz and Johnson Loh for their guidance and support.

References

[1] J. Timchek et al., “The Intel Neuromorphic DNS Challenge,” Mar. 2023, doi: https://doi.org/10.48550/arxiv.2303.09503.

[2] K. Tan and D. Wang, “A Convolutional Recurrent Neural Network for Real-Time Speech Enhancement,” Conference of the International Speech Communication Association, Sep. 2018, doi: https://doi.org/10.21437/interspeech.2018-1405.

[3] Soumi Maiti and M. I. Mandel, “Speech Denoising by Parametric Resynthesis,” May 2019, doi: https://doi.org/10.1109/icassp.2019.8683130.

[4] Soumi Maiti and M. I. Mandel, “Parametric Resynthesis With Neural Vocoders,” Oct. 2019, doi: https://doi.org/10.1109/waspaa.2019.8937165.

[5] K. Kinoshita, M. Delcroix, A. Ogawa, and T. Nakatani, “Text-informed speech enhancement with deep neural networks,” Sep. 2015, doi: https://doi.org/10.21437/interspeech.2015-409.

[6] Ye Jia, Yu Zhang, Ron J. Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, Yonghui Wu, ”Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis”, Jan. 2019, doi: https://doi.org/10.48550/arXiv.1806.04558

[7] H. Abouzid, O. Chakkor, O. G. Reyes, and S. Ventura, “Signal speech reconstruction and noise removal using convolutional denoising audioencoders with neural deep learning,” Analog Integrated Circuits and Signal Processing, vol. 100, no. 3, pp. 501–512, Mar. 2019, doi: https://doi.org/10.1007/s10470-019-01446-6.

[8] J. T. Geiger, F. Weninger, J. F. Gemmeke, M. Wollmer, B. Schuller, and G. Rigoll, “Memory-Enhanced Neural Networks and NMF for Robust ASR,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 6, pp. 1037–1046, Jun. 2014, doi: https://doi.org/10.1109/taslp.2014.2318514.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need” in Advances in neural information processing systems, Dec. 2017, pp. 5998–6008, doi: https://doi.org/10.48550/arXiv.1706.03762

[10] A. Gulati et al., “Conformer: Convolution-augmented Transformer for Speech Recognition,” Conference of the International Speech Communication Association, May 2020, doi: https://doi.org/10.21437/interspeech.2020-3015.

[11] T. O’Malley, A. Narayanan, Q. Wang, A. Park, J. J. Walker, and N. Howard, “A Conformer-Based ASR Frontend for Joint Acoustic Echo Cancellation, Speech Enhancement and Speech Separation,” Dec. 2021, doi: https://doi.org/10.1109/asru51503.2021.9687942.

[12] J. Caroselli, A. Naranayan, and T. O’Malley, “Cleanformer: A multichannel array configuration-invariant neural enhancement frontend for ASR in smart speakers,” Apr. 2022, doi: https://doi.org/10.48550/arxiv.2204.11933.

[13] RASR2: Wei Zhou, Eugen Beck, Simon Berger, Ralf Schlüter, Hermann Ney” The RWTH ASR Toolkit for Generic Sequence-to-sequence Speech Recognition”, 2023

[14] M. Chen et al., “MultiSpeech: Multi-Speaker Text to Speech with Transformer,” Oct. 2020, doi: https://doi.org/10.21437/interspeech.2020-3139.

[15] L. Wan, Q. Wang, A. Papir, and I. Moreno, “Generalized End-to-End Loss for Speaker Verification,” International Conference on Acoustics, Speech, and Signal Processing, Apr. 2018, doi: https://doi.org/10.1109/icassp.2018.8462665.

[16] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, Koray Kavukcuoglu, ”WaveNet: A Generative Model for Raw Audio” Sept. 2016, doi: https://doi.org/10.48550/arXiv.1609.03499.

[17] E. Casanova, J. Weber, C. Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir Antonelli Ponti, “YourTTS: Towards Zero-Shot Multi-Speaker TTS and Zero-Shot Voice Conversion for everyone,” Dec. 2021, doi: https://doi.org/10.48550/arxiv.2112.02418.

[18] J. Kong, J. Kim, and J. Bae, “HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis,” vol. 33, pp. 17022–17033, Oct. 2020, doi: https://doi.org/10.48550/arXiv.2010.05646.

[19] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, Yonghui Wu, ”Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions”, Dec. 2017, doi: https://doi.org/10.48550/arXiv.1712.05884.