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]. Different deep learning architectures for speech recognition and synthesis are compared and based on that, an encoder-decoder structure is presented for the task of deep audio denoising via resynthesis, which extracts the text from the original noisy speech (encoder, ASR) and generates speaker dependent clean speech from the text (decoder, TTS). 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 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 LSTMs and process the signal. Nevertheless, novel approaches like [3], [15], [16], [17] 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 Long-Short-Term-Memory (LSTM) cells [6], Transformer [4] based architectures and recently the conformer [7], [5]. 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 [5]. In the presence of background noise competing with the speech signal with a SNR of -5dB, the WER can increase up to 30-70%, even with the conformer based ASR system. Another proposed ASR system is the cleanformer [8], which addresses the problem of noise sensitivity of ASR. Cleanformer takes as inputs a single channel each of raw (noisy) and enhanced signals, and derives a time-frequency mask self-attention to derive a time-frequency mask. The enhanced input is generated by Speech Cleaner, a multichannel adaptive noise cancellation algorithm. The time-frequency mask is applied to the noisy input to produce enhanced features for ASR. In the presence of noise at an SNR of -6dB this ASR system can reduce the WER by about 80% compared to conventional state-of-art ASR systems. 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” [9]. 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. Conventional single speaker state-of-art TTS system are based on the transformer [4] architecture, which was already mentioned in section II A. For transformer-based training the text and speech data are usually of high quality and the text-to-speech adaptations are easy to learn and afterwards high quality speech can be generated from text. However further extensions of transformer based TTS towards Multi-Speaker-TTS have been implemented, allowing to generate clean speech spoken by arbitrary voices, by training on either a fixed set of Speaker IDs as in MultiSpeech [10], which is converted to a speaker embedding space [11]. Based on the text encoder and the speaker encoder, the transformer layers are trained. One challenge of this approach is the increased difficulty of the transformer to learn the speech alignments to generalize speech synthesis on multiple speakers with various speed, pitch and acoustic conditions while maintaining speech quality: Monotonic and diagonal alignments in the attention weights between text and speech are critical to ensure the quality of synthesized speech. In [10], adding diagonal constraint on the attention weights is implemented to force the model to learn correct alignments. With 4 transformer layers each ($N = 4$), the model has in total 1857777 trainable parameters. MultiSpeech TTS was trained on mel spec as output and phonemes as input and WaveNet [13] 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 [12] is a multilingual, multi-speaker text-to-speech model, which was not only trained on VCTK, TTS-Portuguese and MAILABS french datasets (trilingual) but also on 1151 additional English speakers from both 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. 4-dimensional trainable language embeddings are concatenated into the embeddings of each input character for multi-language training. The number of transformer blocks is 10 and the number of hidden channels is 196. As vocoder for final waveform generation, a modified version of HiFi-GAN is chosen. YourTTS does not train vocoder and transformer separately as usual but jointly: The transformer and the vocoder are connected by a posterior encoder, which consists of 16 non-causal WaveNet residual blocks and recieves a linear spec as input. Its output is directly fed into the vocoder and thus avoids the need for conversion from mel spec to waveform. The Speaker Consistency Loss (SCL) measures the speaker similarity and enables 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 [12].

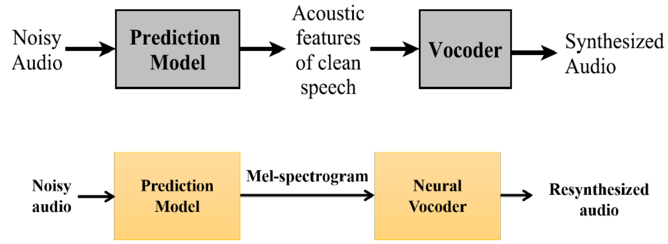

Fig. 1. Top: The parametric resynthesis model [15]. Bottom: The neural vocoder [16]

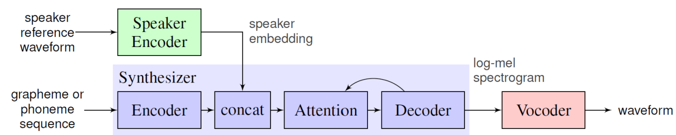

Fig. 2. The transfer learning model for Multispeaker Text-To-Speech Synthesis [17]

C. Speech denoising by resynthesis

Several works in resynthesis of speech have been carried out, some for noise supression [15], [16] as well as for general speech synthesis for arbitrary speakers [17]. In [15] 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 such as the log mel spectra of the noisy audio and predicts clean speech acoustic features at a fixed frame rate. The WORLD vocoder transforms the clean features to an audio waveform. 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 [16], the work is extended by training the WaveNet jointly with the predicition model to generate clean waveform samples. However, it is observed that the PR resynthesis files are not perfectly aligned with the clean signal itself, which affects the objective scores significantly and that PR-neural-joint performance decreases for the joint case. Figure 1 shows the structure of both models.

In [17], a neural network-based system for text-to-speech (TTS) synthesis is presented that is able to generate speech audio in the voice of different during training seen as well as unseen speakers. The core structure of three main components as follows: The speaker encoder network, which is the speaker embedding and maps the reference waveform to the speaker embedding space, a TTS synthesizer network based on Tacotron 2, that generates a mel spec from text and the speaker embedding and an auto-regressive WaveNet-based vocoder network that converts the mel spec waveform samples. The resulting architecture is depicted in Figure 2.

On the LibriSpeech dataset, the model achieves 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.

D. Text-informed speech enhancement with DNNs

In [19], so-called ”text informed speech enhancement” is utilized to enhance noisy speech with the guide of extracted speech features from text. This approach, often referred to as corpus-/inventory-based approach, is greatly inspired by unit-selection-based text-to-speech synthesis (TTS) (as introduced in section II B) technologies. [19] refers to literature which reports that ”very high audible-quality enhancement is possible by first looking for speech units in the training data that best matches to the underlying clean speech components in the target noisy speech, and then generating enhanced spectra by concatenating the units” [19]. The work in [19] modifies this approach by not looking for speech units in the training data (which would require a large memory) but instead for speech units generated by a TTS system from the raw input text. Guided by the extracted speech information from the text, which is performed by TTS, a deep neural network (DNN) enhances the noisy input speech by processing its extracted features through its nonlinearities. The resulting architecture is depicted in Figure 4: From the raw input text, text features are extracted and from the noisy speech the speech features. Afterwards, the text features are time-aligned to the speech features, as opposed to a simple TTS case, so that the DNN can learn the correct time dependent mapping to enhanced speech features. The aligned featres are the fed into the DNN and trained on the clean speech features (f.e. mel spec). The extracted text can be fed into the TTS system to generate a candidate for the final speech synthesis, which in turn is performed by the joint network.

E. 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. SI-SNR measures ”how clear the human speech is above the noise in the output of the N-DNS system”, similar to the conventional Source-to-Noise Ratio (SNR) The difference of SI-SNR compared to SNR is the scale-invariance of SI-SNR, e.g. the output volume independence such that solutions over others that simply increase the output volume are not favored over others. 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

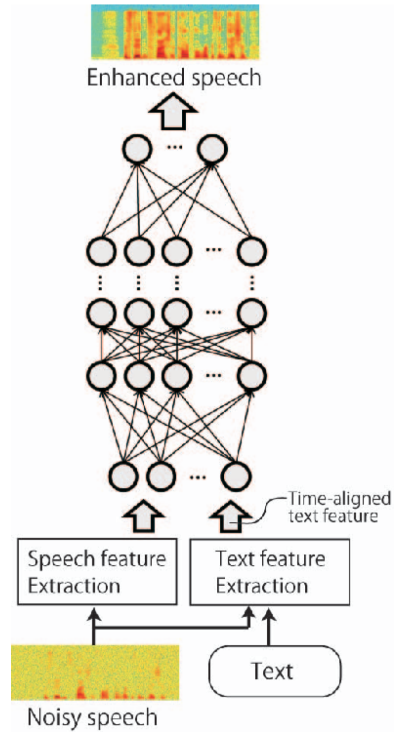

Fig. 3. Text-informed speech enhancement based on a DNN [19].

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

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

and

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

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. DNSMOS has proven to be able to generate scores that are highly correlated with human perceptual assessment compared to other similar

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

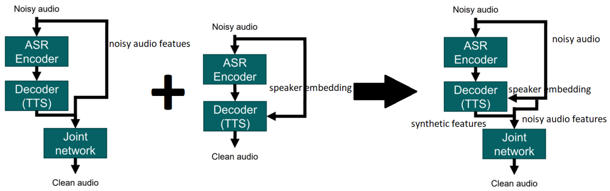

Fig. 4. The two possible models for denoising by speech resynthesis

methods like Perceptual Evaluation of Speech Quality and therefore is a reliable prediction of MOS scores for audio without the need for human surveys.

III. Proposed Approach

On the in II laid foundation, we propose our deep denoising model. 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, which contains additive noise, to a text or phoneme sequence. Due to the noise sensitivity of ASR it is critical to perform preprocessing of the waveform to minimize the WER. Cleanformer [8] cound 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 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 the addition of a joint alignment network, which mixes the synthetic and noisy reference speech an attractive approach as it is presented in II-D. 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 4 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 and the complete architecture which unites both approaches.

A. ASR-TTS Encoder-Decoder

For the ASR system, it is necessary to have high noise robustness and a low WER, even under noisy environments. Suitable candidates are the RWTH Speech Recognition Framework[2] or Cleanformer, which can be utilized to predict text from noisy input speech [9], [8]. Based on the literature from [8], [9], we can expect a WER of approximately 10%. For the Decoder, we use the pretrained YourTTS [12] model from the Coqui-AI TTS library[3]. 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 (II-B). Consequently, we presuppose for the TTS-Decoder a) a high quality of the output speech under b) a high similarity with the original audio, despite having noise. It should extract the speech features without synthesizing the noise. More precisely, this high similarity should be achieved with the multi-speaker property of TTS. 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 II-C. This alignment is done by our implementation of the Joint Network *MixNet* IV-D.

B. Joint-Network MixNet

MixNet is a fully connected convolutional deep neural network (FCDNN), whose hyperparameters are inspired by [19], shortly summarized in II-D: 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

[2]https://github.com/rwth-i6/rasr

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

size is set as a hyperparameter) of the input mel spec along the frame domain. It should cherry pick the best chunks to generate the most similar mel spec possible via recombination and/or superposition by taking the best of both (clean but faulty and orignal but noisy) worlds. 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. There are multiple reasons which lead to this different architecture: i) If the model would simply be trained on clean features it would perform a much more difficult task, the task of denoising the noisy audio. Not only would a simple FCDNN be overwhelmed, but also would this, if it would work, make the Encoder-Decoder structure in III unnecessary. ii) We already got a good candidate for our denoised audio, the synthetic one from the decoder. Only wrong words or very different sounding sections of the synthetic audio should simply replaced by the corresponding chunks from the noisy audio. Thus, we would increase the SI-SNR at the cost of more background noise, because the signals are better aligned (II-E1. iii) Together with i) and 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. 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.

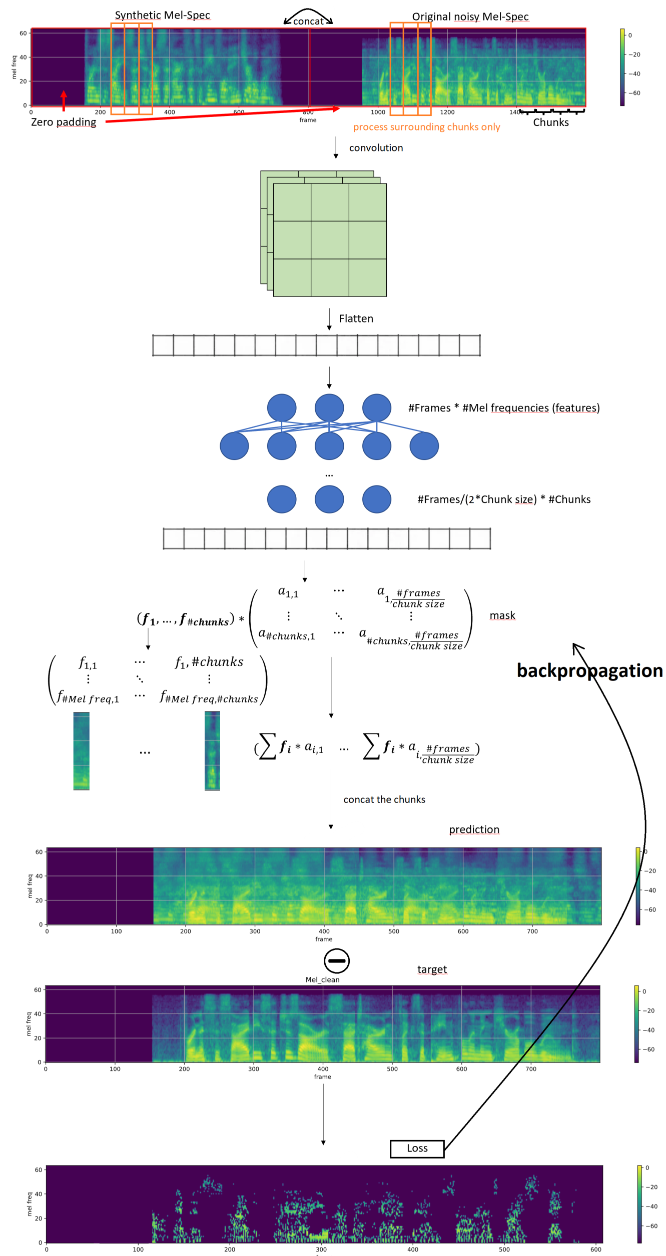

Fig. 5. The structure of the Joint Network MixNet

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 dataset which contains 836 speakers and LibriSpeech test-clean dataset. The samples were sampled at at 16kHz. We trained on a subset of the dataset containing 500 samples each shorter than 5s. We use 10ms hop size and 25ms window size to extract the mel spec 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 0dB to the clean audio from the dataset.

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

Source	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
SI-SNR model I	-43.2867 dB	-19.7158 dB	-47.3405 dB	-19.0866 dB	-37.9634 dB	-40.9666 dB	-24.0065 dB	0 dB
SI-SNR model II	-46.5782 dB	-22.9126 dB						

Fig. 6. SI-SNRs for for the two MixNet configurations, TTS decoder, noisy original and ground truth audio with and without GriffinLim as Vocoder

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 II-D (model III) with 27823824 params.

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 II-E1 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 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 DNSMOS score from II-E2 by passing them to the DNSMOS network.

B. SI-SNR evaluation

Figure 6 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 [12], 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.

C. DNSMOS evaluation

Figure 7 depicts the DNSMOS scores for the two models, the TTS decoder, noisy original and ground truth audio with

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

Fig. 7. DNSMOS for the two MixNet configurations, TTS decoder, noisy original and ground truth audio

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

Fig. 8. The MSE loss for two configurations of MixNet

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 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 [12], since after mel conversion and recovery with GriffinLim, the DNSMOS scores of the TTS worsen to 1.1447, 1.3876 and 1.7752.

D. Loss evaluation

Figure 8 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 . Interesingly, 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.

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.

References

[1] Jonathan Timcheck, Sumit Bam Shrestha, Daniel Ben Dayan Rubin, Adam Kupryjanowy, Garrick Orchard, Lukasz Pindory, Timothy Shea, and Mike Davies, "The Intel Neuromorphic DNS Challenge", 2023

[2] Ke Tan, DeLiang Wang, "A Convolutional Recurrent Neural Network for Real Time Speech Enhancement", 2018

[3] Houda Abouzid, Otman Chakkor, Oscar Gabriel Reyes, Sebastian Ventura, "Signal speech reconstruction and noise removal using convolutional denoising audioencoders with neural deep learning", 2019

[4] 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", 2017, pp. 5998–6008.

[5] Tom O'Malley, Arun Narayanan, Quan Wang, Alex Park, James Walker, Nathan Howard, "A conformer based ASR frontend for joint acoustic echo cancellation", 2021

[6] Jürgen T. Geiger, Felix Weninger, Jort F.Gemmeke,, Martin Wöllmer, Björn Schuller, and Gerhard Rigoll, "Memory-Enhanced Neural Networks and NMF for Robust ASR", IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 22, NO. 6, JUNE 2014

[7] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, Ruoming Pang, "Conformer: Convolution-augmented Transformer for Speech Recognition", 2020

[8] Joseph Caroselli, Arun Narayanan, Nathan Howard, Tom O'Malley, "Cleanformer: A multichannel array configuration-invariant neural Enhancement frontend for ASR in smart speakers", ICASSP 2023

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

[10] Mingjian Chen, Xu Tan, Yi Ren, Jin Xu, Hao Sun, Sheng Zhao, Tao Qin, Tie-Yan Liu, "MultiSpeech: Multi-Speaker Text to Speech with Transformer ", 2020

[11] Li Wan, Quan Wang, Alan Papir, Ignacio Lopez Moreno, "Generalized End-to-End Loss for Speaker Verification", 2017

[12] Edresson Casanova, Julian Weber, Christopher Shulby, Arnaldo Candido Junior, Eren Gölge and Moacir Antonelli Ponti, "YourTTS: Towards Zero-Shot Multi-Speaker TTS and Zero-Shot Voice Conversion for everyone", 2023

[13] 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"2016

[14] Jungil Kong, Jaehyeon Kim, Jaekyoung Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis", 2020

[15] Soumi Maiti and Michael I. Mandel, "Speech denoising by parametric resynthesis", 2019

[16] Soumi Maiti and Michael I. Mandel, "parametric resynthesis with neural vocoders", 2019

[17] 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", 2019

[18] 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", 2017

[19] Keisuke Kinoshita, Marc Delcroix, Atsunori Ogawa, Tomohiro Nakatani, "Text-informed speech enhancement with deep neural networks", 2015