ATTENTION WAVE-U-NET FOR SPEECH ENHANCEMENT

Ritwik Giri, Umut Isik, and Arvindh Krishnaswamy

Amazon Inc.

{ritwikg, umutisik, arvindhk}@amazon.com

ABSTRACT

We propose a novel application of an attention mechanism in neural speech enhancement, by presenting a U-Net architecture with attention mechanism, which processes the raw waveform directly, and is trained end-to-end. We find that the inclusion of the attention mechanism significantly improves the performance of the model in terms of the objective speech quality metrics, and outperforms all other published speech enhancement approaches on the Voice Bank Corpus (VCTK) dataset. We observe that the final layer attention mask has an interpretation as a soft Voice Activity Detector (VAD). We also present some initial results to show the efficacy of the proposed system as a pre-processing step to speech recognition systems.

Index Terms— Speech denoising, speech enhancement, deep learning, attention, U-Net.

1. INTRODUCTION

Speech enhancement is the task of improving the general audio quality and intelligibility of speech-containing audio. A major part of speech enhancement is the task of speech denoising, i.e. removing unwanted noise from speech recordings. Speech denoising is a special case of audio source separation, which is the task of separating out individual audio sources from an input mixture. This paper proposes a new neural network architecture for the tasks of speech denoising and source separation from background.

Deep learning approaches with U-Net architectures have been used on various machine learning tasks in medical diagnostics [1], semantic segmentation [2], monocular depth estimation [3], and singing voice separation [4], and others. They have found success in speech enhancement tasks as well, with some architectures working with the raw waveform [5, 6, 7, 8], and others working in the time-frequency (STFT output) domain [9].

Recently, some works have explored the use of attention mechanisms for U-Net [10] architectures in medical diagnostics. *Attention*, in a neural network architecture, refers to the existence of layers that recall, or combine, features from farther parts of the input and/or different/earlier layers of the network. This is done via multiplying with attention masks/coefficients, which are often interpretable. The masks can often be interpreted, for example, as word correspondences in sequence-to-sequence neural machine translation [11], as word relationships in transformer networks for language models [12], or as simple saliency maps in medical diagnostics [10, 13].

Deep learning approaches have also greatly improved performance on the source separation [14], singing voice separation [5], and speech enhancement [8] tasks. Since deep learning techniques that apply to one of these tasks usually apply to the others, we consider works on these tasks together. There are two fundamental approaches to these tasks, processing the audio in the time-frequency domain, or in the raw waveform domain. Recent time-frequency domain approaches [7, 9] rely on real or complex ratio masking which has been shown to be more effective than direct prediction of the spectrum for time-frequency domain based source separation [15].

Raw-waveform approaches [16, 17], operate directly on audio samples. Recent works in the raw audio domain that use the U-Net architecture include the following. SEGAN [8], which proposes a U-Net-based Generative Adversarial Network for denoising; Wave-U-Net [5], which employs strided convolutions in lieu of max-pools and other architectural improvements for avoiding audible artifacts in singing voice separation; and [6] which employs Wave-U-Net for speech enhancement. Wave-U-Net is also the starting point for our proposed architecture.

In this paper, we approach the speech enhancement task in the raw waveform domain, and propose a U-Net architecture with a *local self-attention* (or local intra-attention) mechanism. The local self-attention mechanism, similar to [10], recalls features from previous layers centered at the same location. Specifically, we apply the attention mechanism to the skip connections in the U-Net architecture. Instead of directly concatenating the earlier layer features from the same scale as is usual in the U-Net architecture, we first multiply the skip-connected layers by an attention-mask.

We show that, based on speech quality metrics on the VCTK dataset, the added attention mechanism significantly improves denoising quality, achieving state-of-the-art results outperforming all published speech enhancement methods on this dataset. We also observe that the final-layer attention mask can be interpreted as a soft voice activity detector. To best of our knowledge, this is the first work where attention-techniques have been used in speech enhancement.

The rest of the article is organized as follows: In Section 2, the proposed method is presented in detail, in Section 3 we present evaluation results in terms of widely used speech quality metrics of our proposed method and other competing methods over a benchmark dataset, and finally Section 4 concludes the paper and talks about some future research directions.

2. ARCHITECTURE: WAVE-U-NET WITH ATTENTION

In this section, we describe the Wave-U-Net architecture and the attention mechanism we employ.

2.1. Wave-U-Net

Our baseline architecture follows the Wave-U-Net architecture proposed in [5], which is essentially a one-dimensional variant of the popular U-Net, and which operates directly on the time domain signal.

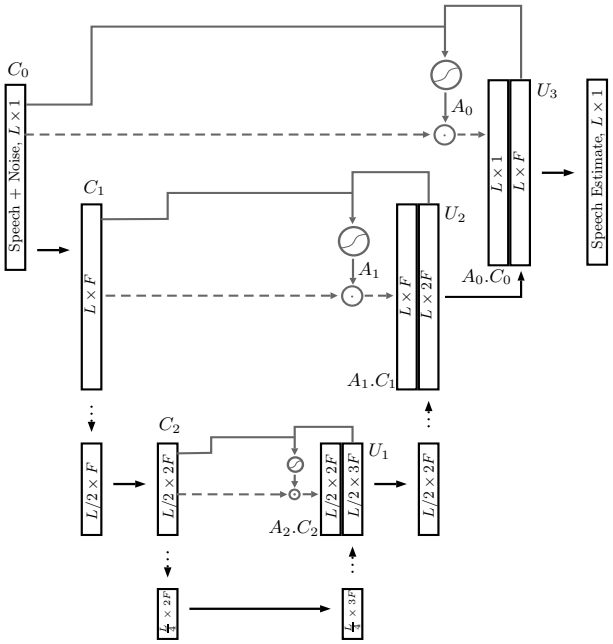

Figure 1: Wave-U-Net with attention architecture, with 2 down-blocks. Network used in experiments has 12 down-blocks. Solid arrows represent convolutional layers, dotted vertical arrows represent down-sampling and up-sampling layers, and the dashed central arrows represent the skip connections for concatenation.

Wave-U-Net has been described in detail in [5], for the music-vocal separation task. It consists of a series of down-blocks, followed by one 'bottom' convolutional layer, followed by a series of up-blocks with skip connections from the down-blocks to the up-blocks. Because of the downsampling blocks, the model can compute a number of higher level features on coarser time scales, which are concatenated with the local, high resolution features computed from the same level upsampling block. This concatenation results into multi-scale features for predictions. Figure 1 shows the entire architecture with only two downblocks for simplicity.

Each down-block is a 1D-convolution followed by 'decimation' i.e., subsampling by 2 (implemented together as a strided convolution with stride 2). The first convolutional layer has F feature channels, and each such layer, up to and including the bottom layer adds another extra F number of features. Each up-block is a $2\times$ upsampling, which is achieved using a linear interpolation layer, followed by concatenation of features from the same-scale down-block, followed by a convolution layer. After each convolutionl layer in the network, a leaky ReLU non-linearity with $\alpha = 0.2$.

Referring to Figure 1, let C_i, $i = 1, \ldots, d$ be the output of the convolution in the ith down block. Let U_{d-i} be the output of the linear interpolation (for up-sampling) in the ith up-block. The skip connections in Wave-U-Net consists of concatenating C_i with U_{d-i}.

2.2. Attention Mechanism

In our proposed architecture, instead of directly concatenating features computed during the contracting path with the same hierarchical level among the upblocks, we use attention gates to identify relevant features from downblock by multiplying it with an attention mask as shown in Figure 1.

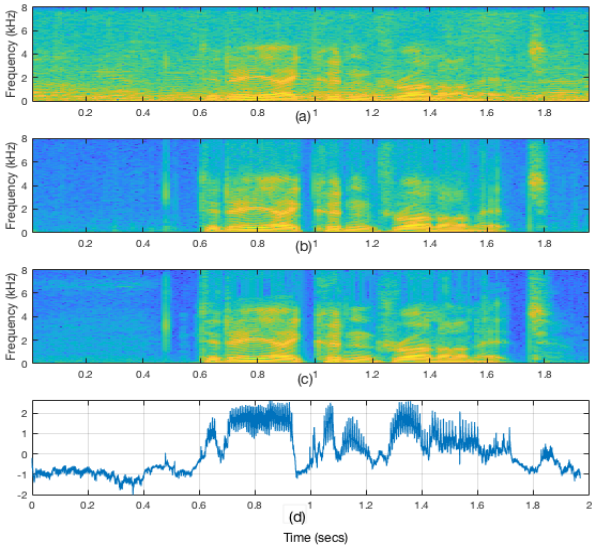

Figure 2: (a) Noisy, (b) Processed, (c) Clean Spectrograms, (d) Normalized final layer attention mask

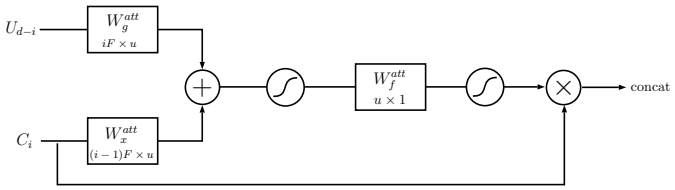

Figure 3: The Attention Mechanism.

To model the attention mechanism, there are additional 1-d convolutions, W_g^{att} and W_x^{att} (u convolutions each), of kernel size 1, which are used to compute an intermediate u-feature layer,

$$B_i = \sigma(W_x^{\text{att}} C_i + W_g^{\text{att}} U_{d-i} + b_{i,1}), \quad (1)$$

which is fed to a single convolution W_f^{att} with kernel size 1 to give the attention mask: $A_i = \sigma(W_f^{att} B_i + b_{i,2})$. The term-wise product $A_i.C_i$ is then concatenated with U_{d-i}. This concludes the attention mechanism for one block. See Figure 3 for a visual description of the attention mechanism.

At the final layer before the output layer, there is an attention mechanism computed the same way as done previously, but with with different inputs: U_{d+1} is the output of the convolution in the last up-block, rather than upsampling layer, and C_0 is the noisy neural net input rather than an intermediate layer (c.f. top of Figure 1).

3. EXPERIMENTAL RESULTS

3.1. Datasets

To evaluate our proposed method, we used the CSTR VCTK Corpus [21] which is made available under the ODC Attribution License. This corpus consists of 56 speakers, 28 male and 28 female - of different accent regions (Scotland and United States).

Table 1: Objective Evaluation of different Algorithms over VCTK Test Set

Methods	PESQ	CSIG	CBAK	COVL	SSNR
Noisy	1.97	3.35	2.44	2.63	1.68
Wave-U-Net (16 kHz, no aug)	2.53	3.77	3.12	3.14	7.42
Wave-U-Net (16 kHz, no aug) + Attention	2.58	3.79	3.17	3.18	7.87
Wave-U-Net (16 kHz, with aug)	2.52	3.73	3.25	3.11	9.56
Wave-U-Net (16 kHz, with aug) + Attention	2.56	3.83	3.28	3.19	**9.57**
Wave-U-Net (48 kHz, no aug)	2.55	3.88	3.27	3.21	9.46
Wave-U-Net (48 kHz, no aug) + Attention	2.62	3.93	**3.30**	3.27	9.48
Wave-U-Net (48 kHz, with aug)	2.58	3.88	3.28	3.22	9.44
Wave-U-Net (48 kHz, with aug) + Attention	**2.63**	**3.95**	**3.30**	**3.29**	9.35

Table 2: Comparison with Competing Methods on VCTK Test Set with 28 speakers training set

Methods	PESQ	CSIG	CBAK	COVL	SSNR
Noisy	1.97	3.35	2.44	2.63	1.68
Wiener [18]	2.22	3.23	2.68	2.67	5.07
SEGAN [8]	2.16	3.48	2.94	2.80	7.73
Wave-U-Net [6]	2.40	3.52	3.24	2.96	9.97
Wavenet [19]	-	3.62	3.23	2.98	-
Deep Feature Loss [20]	-	3.86	3.33	3.22	-
Wave-U-Net (16 kHz, with aug) + Attention (Ours)	2.57	3.79	3.32	3.18	**10.05**
Wave-U-Net - ℓ_1 (16 kHz, with aug) + Attention (Ours)	**2.62**	**3.91**	**3.35**	**3.27**	**10.05**

There are around 400 sentences available from each speaker. All data is sampled at 48 kHz and orthographic transcription is also available. For noise recordings, the Diverse Environments Multichannel Acoustic Noise Database (DEMAND) [22] was used. The signal-to-noise (SNR) values used for training were: 15 dB, 10 dB, 5 dB and 0 dB for 10 different noise types. Therefore total of 40 different noisy conditions (10 noises $\times$ 4 SNRs) were simulated. Hence per speaker, there were around ten different sentences in each condition (total 400 sentences per talker). Testing conditions are mismatched from those of the training. The test set inputs were made with four SNR settings different from the training set (17.5, 12.5, 7.5, and 2.5 dB), using the 5 noise types (unseen during training) from DEMAND and the 2 hold-out speakers from the Voice Bank corpus.

Finally, for a fair comparison with competing methods, i.e., end-to-end speech enhancement methods that operate directly on the raw waveform, we also trained our proposed method using a smaller training set (28 speakers) at 16 kHz, following [8, 19], which were only evaluating the respective models on the 28-speaker dataset.

3.2. Experimental Setup

Unlike in [6], where speech and noise estimates were separate outputs of the model, in this work, the target output is only the foreground/clean speech. Being employed in a model that is only trying to predict the foreground, the attention mechanism can attenuate features from background-only regions relative to voice-activity regions, rather than having mixed responsibilities. For our 16 kHz models, we use $L_{in} = L_{out} = 8192$ samples, whereas for 48 kHz models we use, $L_{in} = L_{out} = 16384$ input and output samples. Following [5], our network size is also 12 down-blocks. All the models are trained on randomly-sampled audio excerpts using Adam optimizer, with learning rate = 10^{-4} and batch size of 16. We use $F = 24$ filters for convolution in first layer of network, downsampling block filters of size 15 and upsampling block filters of size 5 as in [5]. We train all our models using either ℓ_2 or ℓ_1 cost function over foreground output samples in a batch and apply early stopping if there is no improvement on the validation set for 20 epochs, with each epoch consisting of 5000 iterations. After that, we perform finetuning of the final model with the batch size doubled and learning rate lowered to 10^{-5}. Finetuning is performed until the validation loss stops improving for 20 epochs. We randomly selected 1% of the training utterances as our validation set.

3.3. Competing Methods

To evaluate the efficacy of our proposed approach we provide comparisons with several recently proposed speech enhancement algorithms, specifically recent deep-learning (DL) based models that use deep networks to perform end-to-end enhancement directly on the raw waveform. As a baseline, Wiener filtering [18] with a priori noise SNR estimation was used. Other notable DL based methods are,

- **SEGAN** [8]: Time-domain U-Net model optimized with generative adversarial networks.
- **Wavenet** [19]: Time-domain non-causal dilated wavenet-based network.
- **Wave-U-Net** [6]: Time domain U-Net based network, motivated from source separation work of [5].

Table 3: WER Comparison (%) on VCTK and Simulated Datasets

Methods	VCTK	Babble - 0 dB	Babble - 5 dB	Babble - 10 dB
Clean	7.62			
Noisy	11.55	64.54	25.57	13.17
Wave-U-Net (16 kHz, no aug)	11.48	48.13	24.55	13.64
Wave-U-Net (16 kHz, with aug) + Attention	**10.69**	**44.62**	**21.59**	**12.80**

- **Deep Feature Loss** [20]: Time-domain dilated convolution network trained with feature loss from a classifier network.

3.4. Results

In this subsection, to demonstrate the efficacy of our proposed method we present both speech enhancement results and some initial ASR results.

3.4.1. Speech Enhancement Results

We present the results of our proposed approach on the VCTK test set using the widely used quality metrics [23] for speech enhancement. Each measurement compares the enhanced signal with the clean reference of each of the test stimuli provided in the dataset. The first metric is the well known Perceptual Evaluation of Speech Quality (PESQ) - more specifically the wide-band version recommended in ITU-T P.862.2 (from –0.5 to 4.5). We also use the composite scores proposed in [23], that were found to be correlated with human listener ratings. Those are, **CSIG:** MOS predictor of speech distortion, **CBAK:** MOS predictor of intrusiveness of background noise, **COVL:** MOS predictor of overall processed speech quality. Finally, we also report the segmental SNR improvement (**SSNR**).

In Table 1, we show the usefulness of a simple data augmentation method (random attenuation), and also the attention mechanism by systematically comparing the baseline (Wave-U-net) with different variants of our proposed approach for both 16 kHz and 48 kHz models. We found that augmentation helps the ability of our model in denoising specifically for 16 kHz model as indicated by the metric **CBAK** and **SSNR** improvement. Whereas attention mechanism improves all of the performance metrics for both 16 kHz and 48 kHz models.

Finally, in Table 2 we demonstrate that our proposed approach produces state of the art results in terms of speech quality metrics as discussed above by comparing against four recently proposed methods that use deep neural networks to perform end-to-end denoising directly on the raw waveform. Our proposed model, *Wave-U-Net (16 kHz, with aug) + Attention* was trained using both ℓ_2 and ℓ_1 loss. With ℓ_1 loss, our results improved significantly and produced state of the art result among all methods that perform denoising in raw waveform domain. Our proposed model also outperforms all other published approaches that operate on time frequency representation (specifically the STFT domain), including recent work on MMSE-GAN [9]. We would also like to point out that for a fair comparison among all the competing methods for reported results in Table 2, we used a 16 kHZ model and trained using the smaller training set (28 speakers), as reported in [8, 19, 20].

For interpretation purposes, during testing for a noisy utterance at SNR 2.5 dB, we also visualize the final layer attention mask (normalized to mean 0 and variance 1 for visualization purposes) in Figure 2. As shown in Figure 2 (d), we observe that the attention mask for the final layer is essentially learning a soft Voice Activity Detector (VAD), which enables the network to only keep lower-level features from regions with voice activity.

3.4.2. ASR Results

We demonstrate that along with improving the perceptual speech quality metrics, our proposed speech denoising/enhancement approach also shows potential usefulness as a part of an acoustic front-end for an Automatic Speech Recognition (ASR) system. We used a publicly available ASR system, trained for conversational speech recognition. The VCTK test set comes with the ground truth transcriptions, which has been used as ground truth to compute Word Error Rate (WER). Since the VCTK test set does not include harsher conditions (SNR $\approx$ 0 dB), we simulate more challenging conditions by mixing babble noise with the clean recordings for 0, 5 and 10 dB SNR, with the clean recordings of VCTK test set. Babble noise recording was taken from NOISEX-92 database [24], which is a different database than DEMAND noise database (used to create VCTK training and test set), hence this condition has not been seen by the model during training. We present WER for 4 different conditions: Clean, Noisy, processed by baseline model (Wave-U-Net) and our proposed approach with attention and augmentation. As shown in Table 3, while baseline model shows improvement over no processing in terms of WER (except Babble - 10 dB case), our proposed approach shows further improvement. As expected the performance gap is more significant for harsher conditions (Babble - 0 dB).

4. CONCLUSIONS

In this article, we proposed the Attention Wave-U-Net structure for speech enhancement, which allows our model to attend to the salient portions of the raw-waveform. We also provided visualization of the learned attention mask for final layer, which shows that it is learning a soft VAD, hence keeping features focused more on the voice activity region. Our extensive experimental evaluation showed the efficacy of the attention mechanism for the speech enhancement task, and produced state-of-the-art numbers among all published speech enhancement approaches. Finally, some preliminary ASR results also indicated that this model has the potential to be a very useful part of acoustic front end for an ASR system. Future directions of this work will involve employing a similar attention mechanism for spectral U-Net and possibly for a complex U-Net architecture that operates on the complex-valued spectrogram.

5. REFERENCES

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation,"

in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[2] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[3] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1851–1858.

[4] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep u-net convolutional networks," 2017.

[5] D. Stoller, S. Ewert, and S. Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," *arXiv preprint arXiv:1806.03185*, 2018.

[6] C. Macartney and T. Weyde, "Improved speech enhancement with the wave-u-net," *arXiv preprint arXiv:1811.11307*, 2018.

[7] O. Ernst, S. E. Chazan, S. Gannot, and J. Goldberger, "Speech dereverberation using fully convolutional networks," in *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2018, pp. 390–394.

[8] S. Pascual, A. Bonafonte, and J. Serrà, "Segan: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

[9] M. H. Soni, N. Shah, and H. A. Patil, "Time-frequency masking-based speech enhancement using generative adversarial network," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5039–5043.

[10] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, *et al.*, "Attention u-net: learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[11] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[13] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, and D. Rueckert, "Attention gated networks: Learning to leverage salient regions in medical images," *Medical image analysis*, vol. 53, pp. 197–207, 2019.

[14] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 31–35.

[15] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[16] S.-W. Fu, T.-W. Wang, Y. Tsao, X. Lu, and H. Kawai, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 26, no. 9, pp. 1570–1584, 2018.

[17] A. Pandey and D. Wang, "A new framework for cnn-based speech enhancement in the time domain," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 27, no. 7, pp. 1179–1188, 2019.

[18] P. Scalart *et al.*, "Speech enhancement based on a priori signal to noise estimation," in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, vol. 2. IEEE, 1996, pp. 629–632.

[19] D. Rethage, J. Pons, and X. Serra, "A wavenet for speech denoising," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5069–5073.

[20] F. G. Germain, Q. Chen, and V. Koltun, "Speech denoising with deep feature losses," *arXiv preprint arXiv:1806.10522*, 2018.

[21] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating rnn-based speech enhancement methods for noise-robust text-to-speech." in *SSW*, 2016, pp. 146–152.

[22] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database: A database of multichannel environmental noise recordings," *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 3591–3591, 2013.

[23] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2007.

[24] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication*, vol. 12, no. 3, pp. 247–251, 1993.