A WAVENET FOR SPEECH DENOISING

Dario Rethage, Jordi Pons* and Xavier Serra*

Music Technology Group, Universitat Pompeu Fabra, Barcelona

ABSTRACT

Most speech processing techniques use magnitude spectrograms as front-end and are therefore by default discarding part of the signal: the phase. In order to overcome this limitation, we propose an end-to-end learning method for speech denoising based on Wavenet. The proposed model adaptation retains Wavenet's powerful acoustic modeling capabilities, while significantly reducing its time-complexity by eliminating its autoregressive nature. Specifically, the model makes use of non-causal, dilated convolutions and predicts target fields instead of a single target sample. The discriminative adaptation of the model we propose, learns in a supervised fashion via minimizing a regression loss. These modifications make the model highly parallelizable during both training and inference. Both quantitative and qualitative evaluations indicate that the proposed method is preferred over Wiener filtering, a common method based on processing the magnitude spectrogram.

Index Terms— Speech denoising, convolutional neural networks, end-to-end learning, deep learning, audio

1. INTRODUCTION

Speech recognition is one of the research areas where machine learning has had a very strong impact. However, until today it has been standard practice not to work directly in the time-domain, but rather to explicitly use time-frequency representations as input [1, 2] – for reducing the high-dimensionality of raw waveforms. Similarly, most techniques for speech denoising use magnitude spectrograms as front-end [3, 4, 5]. Nevertheless, this practice comes with its drawbacks of discarding potentially valuable information (phase) and utilizing general-purpose feature extractors (magnitude spectrogram analysis) instead of learning specific feature representations for a given data distribution. Most recently, neural networks have shown to be effective in handling structured temporal dependencies between samples of a discretized audio signal. For example, consider the most local structure of a speech waveform ($\approx$ tens of milliseconds). In this range of context, many sonic characteristics of the speaker (timbre) can be captured and linguistic patterns in the speech become accessible in the form of phonemes. It is important to note that these levels of structure are not discrete, making techniques that explicitly focus on different levels of structure inherently suboptimal. This suggests that deep learning methods, capable of learning multi-scale structure directly from raw audio, may have great potential in learning such structures. To this end, discriminative models have been used in an end-to-end learning fashion for speech classification [6, 7, 8]. But waveforms have also been used for generative tasks [9, 10, 11, 12] and, interestingly, most of these models are autoregressive [9, 10, 11] – except one based on a generative adversarial network [12]. We are not aware of any generative model for raw audio based on variational autoencoders. This discussion motivates our study in adapting Wavenet's model (an autoregressive generative model [11]) for speech denoising. We aim to overcome the inherent limitations of using magnitude spectrogram front-ends by learning multi-scale hierarchical representations from raw audio. Some work in this direction already exists. Back in the 80's, Tamura et al. [13] used a 4-layered feed-forward network operating directly in the raw-audio domain to learn a noise-reduction mapping. And recently: Pascual et al. [12] proposed the use of an end-to-end generative adversarial network for speech denoising, and Qian et al. [14] used a Bayesian Wavenet for speech denoising. In all 3 cases, they provide better results than their counterparts based on processing magnitude spectrograms.

The following lines introduce the original Wavenet architecture. Section 2 describes the modifications we propose, and in Section 3 we discuss our experimental results. Wavenet is the audio domain adaptation of the PixelCNN generative model for images [15, 16] and is capable of synthesizing natural sounding speech [11]. This autoregressive model shapes the (discrete) probability distribution of the next sample given some fragment of previous samples. The next sample is produced by sampling from this distribution. An entire sequence of samples is produced by sequentially feeding previously generated samples back into the model. A high-level visual depiction of the model is presented in Figure 1. Some of Wavenet's key features are presented below:

Gated units. As in LSTMs, sigmoidal gates control the activations' contribution in every layer:

$$z_{t'} = \tanh(W_f * x_t) \odot \sigma(W_g * x_t),$$

where $*$ and $\odot$ denote convolution and element-wise multiplication, respectively. f, t, t' and g stand for filter, input time, output time and gate indices. W_f and W_g are convolutional filters. Figure 2 (*Left*) depicts how sigmoidal gates are utilized.

*Contributed equally.

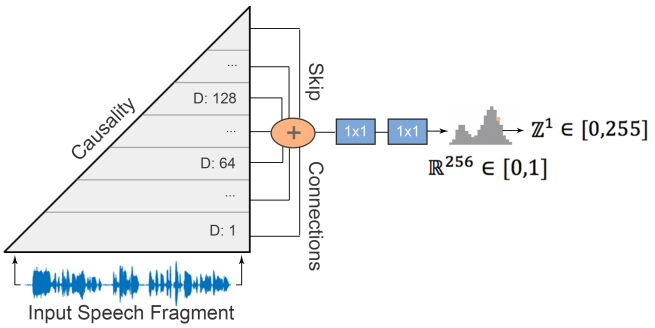

Fig. 1. Overview of Wavenet.

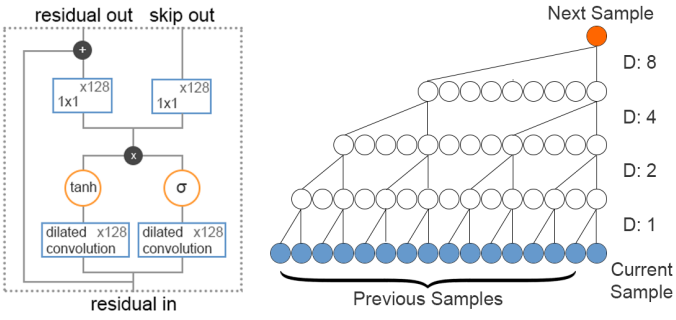

Fig. 2. *Left* – Residual layer. *Right* – Causal, dilated convolutions with increasing dilation factors.

Causal, dilated convolutions. Wavenet makes use of causal, dilated convolutions [11]. It uses a series of small (length = 2) convolutional filters with exponentially increasing dilation factors. This results in an exponential receptive field growth with depth. Causality is enforced by asymmetric padding proportional to the dilation factor, which prevents activations from propagating back in time – see Figure 2 (*Right*). Each dilated convolution is contained in a residual layer, controlled by a sigmoidal gate with an additional 1x1 convolution and a residual connection – see Figure 2 (*Left*).

μ-law quantization. When using a discrete (softmax) output distribution, it is necessary to perform a more coarse 8-bit quantization to make the task computationally tractable. This is accomplished via a μ-law non-linear companding followed by an 8-bit quantization (256 possible values):

$$f(x_t) = sign(x_t)\frac{\ln(1+\mu|x_t|)}{\ln(1+\mu)}$$

Skip connections. These offer two advantages: *(i)* they facilitate training deep models [17], and *(ii)* allow the network to explicitly incorporate features extracted at several hierarchical levels into its final prediction. Figure 1 and 2 (*Left*) provide further details in how skip connections are used.

Context stacks. These deepen the network without increasing the receptive field length as drastically as increasing the dilation factor does. This is achieved by simply stacking a set of layers, dilated to some maximum dilation factor, onto each other – and can be done as many times as desired [11]. For example, Figure 2 (*Right*) is composed of a single stack.

Time-complexity. A significant drawback of Wavenet is its sequential (non-parallelizable) generation of samples. This limitation is strongly considered in our denoising model.

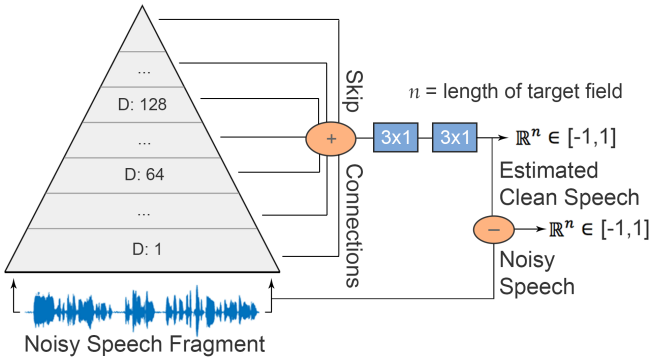

Fig. 3. Overview of the speech-denoising Wavenet.

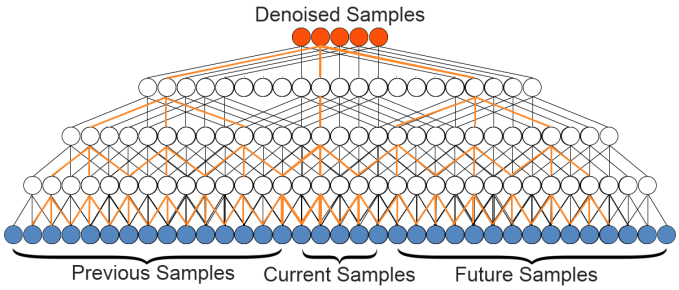

Fig. 4. Predicting on a target field – orange lines: non-causal, dilated convolutions predicting a single sample.

2. WAVENET FOR SPEECH DENOISING

Speech denoising techniques aim to improve the intelligibility and the overall perceptual quality of speech signals with intrusive background-noise. The problem is typically formulated as follows: $m_t = s_t + b_t$, where: $m_t \equiv$ mixed signal, $s_t \equiv$ speech signal, $b_t \equiv$ background-noise signal. The goal is to estimate s_t given m_t. Speech denoising, while sharing many properties with speech synthesis, also has several unique characteristics that motivated the design of this Wavenet adaptation. A high-level visual depiction of the proposed model is presented in Figure 3, and its key features are presented below:

Non-causality. Contrary to audio synthesis, in speech denoising, some future samples are generally available to help make more well informed predictions. Even in real time applications, when a few milliseconds of latency in model response can be afforded, the model has access to valuable information about samples occurring shortly after a particular sample of interest. As a result, and given that Wavenet's time-complexity was a major constraint, the autoregressive causal nature of it was removed in our model. A logical extension to Wavenet's asymmetric dilated convolution pattern, shown in Figure 2, is to increase the filter length to 3 and perform symmetric padding at each dilated layer. If the sample we wish to enhance is now taken to be at the center of the receptive field, this has the effect of doubling the context around a sample of interest and eliminating causality – see Figure 4, in orange.

Real-valued predictions. Wavenet uses a discrete softmax output to avoid making any assumption on the shape of the output's distribution, what is suitable for modeling multi-modal distributions. However, early experiments with discrete softmax outputs proved disadvantageous – the potentially multi-modal output distribution introduced artifacts into the denoised signal. This suggests that real-valued predictions (assuming uni-modal gaussian-shaped output distributions) seem to be more appropriate for our problem. Moreover experiments with discrete softmax outputs resulted in output distributions with high variance, signifying low confidence with the value having highest probability. μ-law quantization was also disadvantageous because it amplified the background-noise. For these reasons, the proposed model learns to directly predict raw audio via minimizing a regression loss – what enables considering alternative costs, as the one used for this study: $L(\hat{s}_t) = |s_t - \hat{s}_t| + |b_t - \hat{b}_t|$ where $\hat{b}_t = m_t - \hat{s}_t$, see Figure 3.

Discriminative model. The proposed model is not autoregressive and its output is not explicitly modeling a probability distribution, but rather the output itself. Furthermore, the model is trained in a supervised fashion – by minimizing a regression loss function. As a result: the proposed model is no longer generative (like Wavenet), but discriminative.

Final 3x1 filters. Since the architecture is not autoregressive, previously generated samples are not fed back into the model to inform future predictions – which enforces time continuity in the resulting signal. Early experiments produced waveforms with sporadic point discontinuities that sounded disruptive. Replacing the kernels of the final layers with 3x1 filters instead of 1x1 filters reimposed this constraint.

Target field prediction. The proposed model does not predict just one, but a set of samples in a single forward propagation – see Figure 4. Parallelizing the inference process from 1 sample to on the order of 1000 samples offers significant memory and time savings. This is because overlapping data is used for predicting neighboring samples, and by predicting target fields these redundant computations are done just once. The receptive field length (rf) of the model is the number of input samples that go into the prediction of a single denoised output sample. In order to maintain that every output sample in the target field (tf) has a full receptive field of context contributing to its prediction, the length of the fragment presented to the model must be equal to: $rf + (tf - 1)$. Finally, note that the cost is computed sample-wise – during training, individual sample costs of a target field are averaged.

Conditioning. The model is conditioned on a binary-encoded scalar corresponding to the identity of the speaker. This condition value is the bias term in every convolution operation. Condition values represent each of the 28 speakers in the training set. We add an auxiliary code (all zeros) denoting any speaker identity so that the trained model can be used for unknown speakers. The same training data is presented to the model either conditioned to its speaker identity or to zeros.

***Noise-only* data augmentation**. A form of augmentation in which 10% of training samples contain only background-noise was also employed after observing that our model had difficulties producing silence.

One-shot denoising. By default, the network is presented with a noisy speech fragment and the condition value is set to zero. The trained model denoises the input in batches, iteratively appending each denoised fragment to the previous. But note that the fully-convolutional nature of the model makes the model flexible in the time-dimension – which permits denoising audio of arbitrary length. As a result of this feature, our model is capable of denoising an entire piece of audio in one-shot – given that sufficient memory is available[1].

3. EXPERIMENTAL RESULTS

The dataset we used [12, 18] was generated from two sources: speech data was supplied by the Voice Bank corpus [19] while environmental sounds were provided by the Diverse Environments Multichannel Acoustic Noise Database (DEMAND) [20]. The subset of the Voice Bank corpus we used features 30 native english speakers from different parts of the world reading out $\approx$ 400 sentences – 28 speakers are used for training and 2 for testing. The subset of DEMAND that we used provides recordings in 13 different environmental conditions such as in a park, in a bus or in a cafe – 8 are mixed with speech during training and 5 are used during testing. During training, 2 artificial noise classes were added – in total 10 different noise classes are available during training. Training samples are synthetically mixed at one of the following four signal-to-noise ratios (SNRs): 0, 5, 10 and 15dB with one of the 10 noise types. This results in 11,572 training samples from 28 speakers under 40 different noise conditions. Test samples are also synthetically mixed at one of the following four different SNRs: 2.5, 7.5, 12.5 and 17.5dB with one of the 5 test-noise types – resulting in 20 noise conditions for 2 speakers. As a result, the test set features 824 samples from unseen speakers and noise conditions. Audios are on average 3 seconds long and are subsampled to 16kHz. No preprocessing is used (*i.e.*: pre-emphasis filtering [12] or μ-law quantization[11]), allowing the pipeline to be end-to-end in the strictest sense.

The proposed model features 30 residual layers as in Figure 2 (*Left*). The dilation factor in each layer increases in the range 1, 2, ..., 256, 512 by powers of 2. This pattern is repeated 3 times (3 stacks). Prior to the first dilated convolution, the 1-channel input is linearly projected to 128 channels by a standard 3x1 convolution to comply with the number of filters in each residual layer. The skip connections are 1x1 convolutions also featuring 128 filters – a RELU is applied after summing all skip connections. The final two 3x1 convolutional layers are not dilated, contain 2048 and 256 filters respectively and are separated by a RELU. The output layer linearly projects the feature map into a single-channel temporal signal by using

[1]When using the model described below on a Titan X Pascal (12GB of VRAM), it is possible to denoise up to 25s of audio with one-shot denoising.

a 1x1 filter. This setup results in a receptive field of 6,139 samples ($\approx$ 384ms) and a target field of 1601 samples ($\approx$ 100ms), optimized to adhere to our memory constraints. The relatively small size of the model (6.3 million parameters) together with its parallel inference on 1601 samples at once, results in a denoising time of ≈ 0.56 seconds per second of noisy audio on GPU. Code and trained models are available online[2].

In our first study we quantify the quality of the denoised speech along three dimensions [21]: signal distortion (with SIG), background-noise interference (with BAK) and overall quality (with OVL). These measures operate in a 1–5 range, aiming to computationally approximate the Mean Opinion Score (MOS) that would be produced from human perceptual trials. We set as baselines: *(i)* the noisy signal, and *(ii)* a signal processing method based on Wiener filtering – widely used for speech-denoising [22, 23] or audio source-separation [24]. The baseline algorithm uses a Wiener filtering method based on a priori SNR estimation [23], as implemented here[3].

Table 1. Quantitative results averaged across all SNRs in the test set. Higher scores are better. Different parameters are studied for Wavenet-based models: *noise-only* data augmentation (0 and 10%) and target field length (1, 101 and 1601 samples).

Wavenet-based	**SIG**	**BAK**	**OVL**
0%, 1 sample*	1.37	1.79	1.28
0%, 101 samples*	1.67	2.07	1.50
0%, 1601 samples	*3.62*	*3.23*	*2.98*
10%, 1601 samples	2.95	3.12	2.49
Wiener filtering	3.52	2.93	2.90
Noisy signal	3.51	2.66	2.79

**Computed on perceptual test set.*

In Table 1 one observes that training with longer target fields is crucial for training models capable of denoising. In addition, we observe that models with a small target field length require impractically long inference times (as a result of many redundant computations). Due to this, the results for smaller target field lengths are computed with the 20-sample perceptual test set (described below). Further, note that the "0%, 1601 samples" model achieves the best results across all metrics. However, informal listening clearly shows that training with 10% *noise-only* augmentation allows the model to produce silence in moments where no speech is present (without degrading the speech signal), which is perceptually pleasant when aurally evaluating denoised samples. When comparing the proposed model with the baseline Wiener filtering method one observes that OVL and SIG results are comparable, showing that Wiener filtering similarly preserves the quality of the speech signal. However, the proposed method removes the background-noise more effectively than Wiener filtering.

In our second study, we conducted perceptual tests with 33 participants to get subjective feedback on the effectiveness of the speech-denoising Wavenet. 20 audio samples were chosen to compose the perceptual test set: 5 samples for each of the four test SNRs, with an equal number of samples coming from each of the 2 speakers in the test set. Aside from these constraints, the samples were chosen randomly. Participants were presented with 4 variants of each sample: *i)* the original mix with speech and background-noise, *ii)* clean speech, *iii)* speech denoised by Wiener filtering, and *iv)* speech denoised with the best performing Wavenet – with 10% noise-only data augmentation and predicting a target field length of 1601 samples. The first two variants were presented as references. Participants were asked to *"give an overall quality score, taking into consideration both: speech quality and background-noise suppression"*[4] for each of the last two variants. Participants were able to give a score between 1–5, with a 1 being described as *"degraded speech with very intrusive background"* and a 5 being *"not degraded speech with unnoticeable background"* [21]. MOS quality measurement is obtained by averaging the scores from all participants. Table 2 presents the results of the perceptual evaluation, showing that participants significantly preferred (t-test: p-value < 0.001) the proposed method over the one based on Wiener filtering.

Audio samples are available online for listening[5].

Table 2. Subjective MOS measures (averaged across all SNRs) on perceptual test set. From 1–5, higher scores are better.

Measurement	**Wiener filtering**	**Proposed Wavenet**
MOS	2.92	3.60

4. CONCLUSION

We have presented a discriminative adaptation of Wavenet's model for speech denoising that features a non-causal and non-autoregressive architecture. The model is able to predict target fields instead of single samples, which significantly reduces the time-complexity of the model – enabling one-shot denoising. Further, we propose using a *noise-only* data augmentation strategy that helps the model to produce silences when only background noise is present. Perceptual tests show that our model's estimates are preferred over the ones based on Wiener filtering. This confirms that it is possible to learn mutli-scale hierarchical representations from raw audio instead of using magnitude spectrograms as front-end for speech denoising.

5. ACKNOWLEDGMENTS

We are grateful for GPUs donated by NVidia. This work is partially supported by the Maria de Maeztu Programme (MDM-2015-0502).

[2] https://github.com/drethage/speech-denoising-wavenet

[3] https://www.crcpress.com/downloads/K14513/K14513_CD_Files.zip

[4] Speech quality and level of noise suppression are jointly rated since preliminary experiments showed that participants were more confident providing overall scores instead of separately rating quality and level of noise.

[5] Speech and background-noise estimates: http://jordipons.me/apps/speech-denoising-wavenet

6. REFERENCES

[1] Dario Amodei, Rishita Anubhai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Jingdong Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, et al., "Deep speech 2: End-to-end speech recognition in english and mandarin," in *ICML*, 2016, pp. 173–182.

[2] W. Xiong, Jasha Droppo, Xuedong Huang, Frank Seide, Mike Seltzer, Andreas Stolcke, Dong Yu, and Geoffrey Zweig, "The microsoft 2016 conversational speech recognition system," *arXiv:1609.03528*, 2016.

[3] Xugang Lu, Yu Tsao, Shigeki Matsuda, and Chiori Hori, "Speech enhancement based on deep denoising autoencoder.," in *Interspeech*, 2013, pp. 436–440.

[4] Shahla Parveen and Phil Green, "Speech enhancement with missing data techniques using recurrent neural networks," in *ICASSP*, 2004, pp. I–733.

[5] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 1, pp. 7–19, 2015.

[6] Ronan Collobert, Christian Puhrsch, and Gabriel Synnaeve, "Wav2letter: an end-to-end convnet-based speech recognition system," *arXiv:1609.03193*, 2016.

[7] Dimitri Palaz, Mathew Magimai Doss, and Ronan Collobert, "Convolutional neural networks-based continuous speech recognition using raw speech signal," in *ICASSP*, 2015, pp. 4295–4299.

[8] Zhenyao Zhu, Jesse H. Engel, and Awni Hannun, "Learning multiscale features directly from waveforms," *arXiv:1603.09509*, 2016.

[9] Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Douglas Eck, Karen Simonyan, and Mohammad Norouzi, "Neural audio synthesis of musical notes with wavenet autoencoders," *arXiv:1704.01279*, 2017.

[10] Soroush Mehri, Kundan Kumar, Ishaan Gulrajani, Rithesh Kumar, Shubham Jain, Jose Sotelo, Aaron Courville, and Yoshua Bengio, "Samplernn: An unconditional end-to-end neural audio generation model," *arXiv:1612.07837*, 2016.

[11] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv:1609.03499*, 2016.

[12] Santiago Pascual, Antonio Bonafonte, and Joan Serra, "Segan: Speech enhancement generative adversarial network," in *Interspeech*, 2017, pp. 3642–3646.

[13] Shinichi Tamura and Alex Waibel, "Noise reduction using connectionist models," in *ICASSP*, 1988, pp. 553–556.

[14] Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, Florencio Dinei, and Mark Hasegawa-Johnson, "Speech enhancement using bayesian wavenet," in *Interspeech*, 2017, pp. 2013–2017.

[15] Aaron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu, "Pixel recurrent neural networks," *arXiv:1601.06759*, 2016.

[16] Aaron van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al., "Conditional image generation with pixelcnn decoders," in *NIPS*, 2016, pp. 4790–4798.

[17] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," in *CVPR*, 2015, pp. 1–9.

[18] Cassia Valentini-Botinhao, Xin Wang, Shinji Takaki, and Junichi Yamagishi, "Investigating rnn-based speech enhancement methods for noise-robust text-to-speech," in *9th ISCA Speech Synthesis Workshop*, pp. 146–152.

[19] Christophe Veaux, Junichi Yamagishi, and Simon King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in *International Conference Oriental COCOSDA held jointly with Conference on Asian Spoken Language Research and Evaluation*. IEEE, 2013, pp. 1–4.

[20] Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent, "The diverse environments multi-channel acoustic noise database: A database of multichannel environmental noise recordings," *JASA*, vol. 133, no. 5, pp. 3591–3591, 2013.

[21] Yi Hu and Philipos C. Loizou, "Evaluation of objective measures for speech enhancement.," in *Interspeech*, 2006.

[22] Philipos C. Loizou, *Speech enhancement: theory and practice*, CRC press, 2013.

[23] Pascal Scalart and Jozue Vieira Filho, "Speech enhancement based on a priori signal to noise estimation," in *ICASSP*, 1996, vol. 2, pp. 629–632.

[24] Jordi Pons, Jordi Janer, Thilo Rode, and Waldo Nogueira, "Remixing music using source separation algorithms to improve the musical experience of cochlear implant users," *JASA*, vol. 140, no. 6, pp. 4338–4349, 2016.