Improved Speech Enhancement with the Wave-U-Net

Anonymous Author(s)
Affiliation
Address
email

Abstract

We study the use of the Wave-U-Net architecture for speech enhancement, a model introduced by Stoller et al for the separation of music vocals and accompaniment. This end-to-end learning method for audio source separation operates directly in the time domain, permitting the integrated modelling of phase information and being able to take large temporal contexts into account. Our experiments show that the proposed method improves several metrics, namely PESQ, CSIG, CBAK, COVL and SSNR, over the state-of-the-art with respect to the speech enhancement task on the Voice Bank corpus (VCTK) dataset. We find that a reduced number of hidden layers is sufficient for speech enhancement in comparison to the original system designed for singing voice separation in music. We see this initial result as an encouraging signal to further explore speech enhancement in the time-domain, both as an end in itself and as a pre-processing step to speech recognition systems.

1 Introduction

Audio source separation refers to the problem of extracting one or more target sources while suppressing interfering sources and noise[18]. Two related tasks are those of speech enhancement and singing voice separation, both of which involve extracting the human voice as a target source. The former involves attempting to improve speech intelligibility and quality when obscured by additive noise [7, 9, 18]; whilst the latter's focus is on separating music vocals from accompaniment [12].

Most audio source separation methods operate not directly in the time-domain, but with time-frequency representations as input and output (front-end). Since 2017, the U-Net architecture on magnitude spectrograms has achieved new state of the art results in audio source separation for music [6] and speech dereverbration [2]. Also, neural network architectures operating in the time domain have recently been proposed for speech enhancement [9, 11]. These approaches have been combined in the Wave-U-Net [12] and applied to singing voice separation. In this paper we apply the Wave-U-Net to speech enhancement and show that it produces results that are better than the current state of the art.

The remainder of this paper is structured as follows. In section 2, we briefly review related work from the literature. In section 3, we introduce briefly the Wave-U-Net architecture and its application to speech. Section 4 presents the experiments we conducted and their results including comparison to other methods. Section 5 concludes this article with a final summary and perspectives for future work.

2 Related work

Source separation of audio has seen great improvement in recent years through deep learning models [5, 8]. These methods, as well as more traditional ones, mostly operate in the time-frequency domain, from deep recurrent architectures predicting soft masks, such as [4], to convolutional encoder-decoder

Submitted to 32nd Conference on Neural Information Processing Systems (NIPS 2018). Do not distribute.

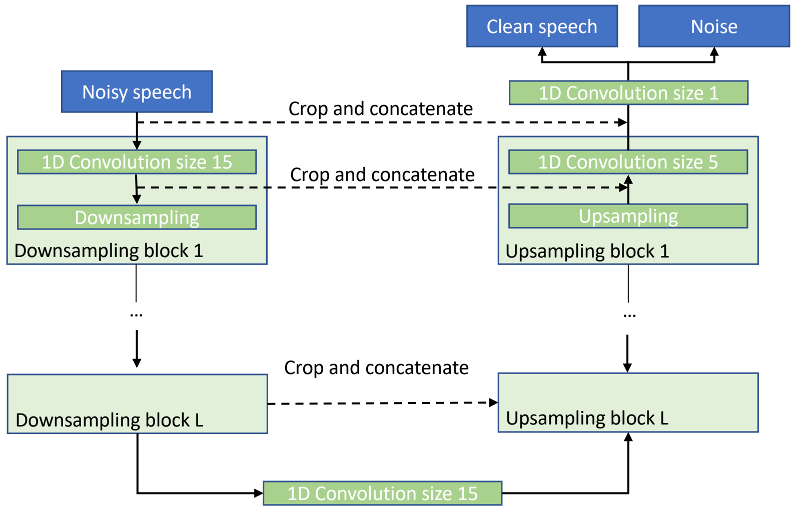

Figure 1: The Wave-U-Net architecture following [12].

architectures like that of [1]. Recently, the U-Net architecture on magnitude spectrograms has achieved new state of the art results in audio source separation for music [6] and speech dereverbration [2].

Also recently, models operating in the time domain have been developed. The development of Wavenet [17] inspired other developments, including [9, 11]. The SEGAN [9] architecture was developed for the purpose of speech enhancement and denoising. It employs a neural network in the time-domain with an encoder and decoder pathway that successively halves and doubles the resolution of feature maps in each layer, respectively, and features skip connections between encoder and decoder layers. It offers state-of-the-art results on the Voice Bank (VCTK) dataset ([14]).

The Wavenet for Speech Denoising [11], another architecture to operate directly in the time domain, takes its inspiration from [17]. It has a non-causal conditional input and a parallel output of samples for each prediction and is based on the repeated application of dilated convolutions with exponentially increasing dilation factors to factor in context information.

3 Wave-U-Net for Speech Enhancement

The Wave-U-Net architecture of [12] combines elements of both of the abovementioned architectures with the U-Net. The overall architecture is a one-dimensional U-Net with down and upsampling blocks.

As per the spectrogram-based U-Net architectures (e.g. [6]), the Wave-U-Net uses a series of downsampling and upsampling blocks to make its predictions, whilst at each level of the network, the time resolution is halved. At the final level, as described by [12], in estimating the two target sources, a 1D convolution prepares the features at each audio sample for source prediction of each sample. To yield an estimate of the target sources, a $tanh$ nonlinearity follows, succeeded by a final LeakyReLU.

In applying the Wave-U-Net architecture to the application of speech enhancement, our objective is to separate a mixture waveform $m \in [-1, 1]^{L_m \times C}$ into K source waveforms $S^1, ..., S^K$ with $S_k \in [1, 1]^{L_S \times C}$ for all $k \in 1, ..., K, C$ as the number of audio channels and L_m and L_S as the respective numbers of audio samples. In our case of monaural speech enhancement we have $K = 2$ and $C = 1$.

Table 1: Objective evaluation - comparing the mean results of the untreated noisy signal, the Wiener-, SEGAN- and Wave-U-Net-enhanced signals. Higher scores are better for all metrics.

Metric	Noisy	Wiener	SEGAN	Wave-U-Net
PESQ	1.97	2.22	2.16	**2.40**
CSIG	3.35	3.23	3.48	**3.51**
CBAK	2.44	2.68	2.94	**3.32**
COVL	2.63	2.67	2.80	**2.95**
SSNR	1.68	5.07	7.73	**9.77**

Table 2: Objective evaluation - mean results, comparing variations of the Wave-U-Net model with different numbers of layers, without fine-tuning applied.

Metric	12-layer	11-layer	9-layer
PESQ	2.40	2.38	**2.41**
CSIG	3.49	3.47	**3.54**
CBAK	**3.23**	3.22	**3.23**
COVL	2.95	2.92	**2.97**
SSNR	9.79	**9.95**	9.87

4 Experiments

4.1 Datasets

We use the same VCTK dataset [14] as the SEGAN [9], which is available publicly, encouraging comparisons with future speech enhancement methods.

The dataset includes clean and noisy audio data at 48kHz sampling frequency. However, like the SEGAN, we downsample to 16kHz for training and testing. The clean data are recordings of sentences, sourced from various text passages, uttered by 30 English-speakers, male and female, with various accents – 28 intended for training and 2 reserved for testing [16]. The noisy data were generated by mixing the clean data with various noise datasets, as per the instructions provided in [9, 14, 15].

With respect to the training set, 40 different noise conditions are considered [9, 16]. These are composed of 10 types of noise (2 of which are artificially-generated[1] and 8 sourced from the DEMAND database [13], each mixed with clean speech at one of 4 signal-to-noise ratios (SNR) (15, 10, 5, and 0 dB). In total, this yields $11,572$ training samples, with approximately 10 different sentences in each condition per training speaker.

The separate test set with 2 speakers consists of a total of 20 different noise conditions: 5 types of noise sourced from the DEMAND database at one of 4 SNRs each (17.5, 12.5, 7.5, and 2.5 dB) [14, 15]. This yields 824 test items, with approximately 20 different sentences in each condition per test speaker [14, 15].

4.2 Experimental setup

As per [12], our baseline model trains on randomly-sampled audio excerpts, using the ADAM optimization algorithm, a learning rate of 0.0001, decay rates $\beta 1 = 0.9$ and $\beta 2 = 0.999$ and a batch size of 16. We specify a network layer size of 12 with 16 extra filters per layer, downsampling block filters of size 15 and upsampling block filters of size 5 like in [12]. We train for 2,000 iterations with mean squared error (MSE) over all source output samples in a batch as loss and apply early stopping if there is no improvement on the validation set for 20 epochs. We use a fixed validation set of 10 randomly selected tracks. Then, the best model is fine-tuned with the batch size doubled and the learning rate lowered to 0.00001, again until 20 epochs have passed without improved validation loss.

4.3 Results

To evaluate and compare the quality of the enhanced speech yielded by the Wave-U-Net, we mirror the objective measures provided in [9]. Each measurement compares the enhanced signal with the clean reference of each of the 824 test set items. They have been calculated using the implementation provided in [7] [2]. The first metric is that of the 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) [7, 9]. Secondly, composite measures of metrics that aim to computationally approximate the Mean Opinion Score (MOS) that would be produced from human perceptual trials are computed [11]. These are: CSIG, a prediction of the signal distortion attending only to the speech signal [3] (from 1 to 5); CBAK, a prediction of the intrusiveness of background noise [3] (from 1 to 5); and COVL, a prediction of the overall effect [3] (from 1 to 5). Last is the Segmental Signal-to-Noise Ratio (SSNR) [10] (from 0 to ∞).

Table 1 shows the results of these metrics for comparison across different speech enhancement architectures. As a comparative reference, it also shows the results of these metrics when applied: directly to the noisy signals; to signals filtered using the Wiener method, based on a priori SNR estimation; and to the SEGAN-enhanced signal, as provided in [9]. The results indicate that the Wave-U-Net is the most effective model for speech enhancement.

Table 2 shows the performance differences between different variations of the Wave-U-Net, with different numbers of layers.[3] In this experiment no fine-tuning was performed, which explains the difference between the 12-layer Wave-U-Nets in Table 1 and in Table 2. The results suggest that fine-tuning does not make a meaningful difference, except on the CBAK measurement, and that smaller models perform better. This is likely due to the size of the receptive field, where for speech the optimal size is probably smaller than for music.

5 Conclusions

5.1 Summary

The Wave-U-Net combines the advantages of several of the most recent successful architectures for music and speech source separation and our results show that it is particularly effective at speech enhancement. The results improve over the state of the art by a good margin even without significant adaptation or parameter tuning. This indicates that there is great potential for this approach in speech enhancement.

5.2 Future work

In comparison to the SEGAN architecture, it is possible that the advantage stems from the upsampling that avoids aliasing, which should be further investigated. The results indicate that there is room for increasing effectiveness and efficiency by further adapting the model size and other parameters, e.g. filter sizes, to the task and expanding to multi-channel audio and multi-source-separation.

References

[1] Pritish Chandna, M. Miron, Jordi Janer, and Emilia Gómez. Monoaural Audio Source Separation Using Deep Convolutional Neural Networks. *13th International Conference on Latent Variable Analysis and Signal Separation (LVA ICA2017)*, (2017), 2017. URL http://mtg.upf.edu/node/3680.

[2] Ori Ernst, Shlomo E. Chazan, Sharon Gannot, and Jacob Goldberger. Speech dereverberation using fully convolutional networks. *CoRR*, abs/1803.08243, 2018. URL http://arxiv.org/abs/1803.08243.

[3] Yi Hu, Philipos C Loizou, and Senior Member. Evaluation of Objective Quality Measures for Speech Enhancement. *IEEE transactions on audio, speech and language processing*, 16(1):229, 2008. doi: 10.1109/TASL.2007.911054. URL http://www.utdallas.edu/loizou/speech/noizeus/.

[1] available here: http://data.cstr.ed.ac.uk/cvbotinh/SE/data/ [14]

[2] available here: http://www.crcpress.com/downloads/K14513

[3] The 10-layer version is missing due to technical problems before the submission and will be completed shortly.

[4] Po-Sen Huang, Minje Kim, Mark Hasegawa-Johnson, and Paris Smaragdis. Singing-Voice Separation from Monaural Recordings Using Deep Recurrent Neural Networks. Technical report. URL http://www.isle.illinois.edu/sst/pubs/2014/huang-ismir2014.pdf.

[5] Po-Sen Huang, Minje Kim, Mark Hasegawa-Johnson, and Paris Smaragdis. Joint optimization of masks and deep recurrent neural networks for monaural source separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(12):2136–2147, 2015.

[6] Andreas Jansson, Eric J. Humphrey, Nicola Montecchio, Rachel M. Bittner, Aparna Kumar, and Tillman Weyde. Singing voice separation with deep u-net convolutional networks. In *Proceedings of the 18th International Society for Music Information Retrieval Conference, ISMIR 2017, Suzhou, China, October 23-27, 2017*, pages 745–751, 2017. URL https://ismir2017.smcnus.org/wp-content/uploads/2017/10/171_Paper.pdf.

[7] Philipos C Loizou. *Speech Enhancement: Theory and Practice*. CRC Press, Inc., Boca Raton, FL, USA, 2nd edition, 2013. ISBN 1466504218, 9781466504219.

[8] Aditya Arie Nugraha, Antoine Liutkus, and Emmanuel Vincent. Multichannel audio source separation with deep neural networks. *IEEE/ACM Trans. Audio, Speech & Language Processing*, 24(9):1652–1664, 2016.

[9] Santiago Pascual, Antonio Bonafonte, and Joan Serrà. SEGAN: Speech Enhancement Generative Adversarial Network. doi: 10.7488/ds/1356. URL http://dx.doi.org/10.7488/ds/1356.

[10] Schuyler R. Quackenbush, T. P. (Thomas Pinkney) Barnwell, and Mark A. Clements. *Objective measures of speech quality*. Prentice Hall, 1988. ISBN 0136290566.

[11] Dario Rethage, Jordi Pons, and Xavier Serra. A Wavenet for Speech Denoising. Technical report. URL https://arxiv.org/pdf/1706.07162.pdf.

[12] Daniel Stoller, Sebastian Ewert, and Simon Dixon. Wave-U-Net: A Multi-Scale Neural Network for End-to-End Audio Source Separation. 6 2018. URL https://arxiv.org/abs/1806.03185.

[13] Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent. The Diverse Environments Multi-channel Acoustic Noise Database (DEMAND): A database of multichannel environmental noise recordings The Diverse Environments Multi-channel Acoustic Noise Database (DEMAND): A database of multichannel environmental noise record-ings. page 10, 2013. doi: 10.5281/zen. URL https://hal.inria.fr/hal-00796707.

[14] Cassia Valentini-Botinhao. Noisy speech database for training speech enhancement algorithms and TTS models, 2016 [sound]. *University of Edinburgh. School of Informatics. Centre for Speech Technology Research (CSTR)*, 2017. URL http://dx.doi.org/10.7488/ds/2117.

[15] Cassia Valentini-Botinhao, Xin Wang, Shinji Takaki, and Junichi Yamagishi. Investigating RNN-based speech enhancement methods for noise-robust Text-to-Speech. Technical report. URL https://www.research.ed.ac.uk/portal/files/26581510/SSW9_Cassia_1.pdf.

[16] Cassia Valentini-Botinhao, Xin Wang, Shinji Takaki, and Junichi Yamagishi. Speech Enhancement for a Noise-Robust Text-to-Speech Synthesis System using Deep Recurrent Neural Networks. 2016. doi: 10.21437/Interspeech.2016-159. URL http://dx.doi.org/10.21437/Interspeech.2016-159.

[17] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. In *The 9th ISCA Speech Synthesis Workshop, Sunnyvale, CA, USA, 13-15 September 2016*, page 125, 2016. URL http://www.isca-speech.org/archive/SSW_2016/abstracts/ssw9_DS-4_van_den_Oord.html.

[18] Emmanuel Vincent, Tuomas Virtanen, and Sharon Gannot, editors. *Audio Source Separation and Speech Enhancement*. John Wiley & Sons Ltd, Chichester, UK, 9 2018. ISBN 9781119279860. doi: 10.1002/9781119279860. URL http://doi.wiley.com/10.1002/9781119279860.