SPEECH DEREVERBERATION USING RECURRENT NEURAL NETWORKS

Shahan Nercessian and Alexey Lukin

iZotope, Inc.
Cambridge, MA, USA
shahan@izotope.com, alex@izotope.com

ABSTRACT

Advances in deep learning have led to novel, state-of-the-art techniques for blind source separation, particularly for the application of non-stationary noise removal from speech. In this paper, we show how a simple reformulation allows us to adapt blind source separation techniques to the problem of speech dereverberation and, accordingly, train a bidirectional recurrent neural network (BRNN) for this task. We compare the performance of the proposed neural network approach with that of a baseline dereverberation algorithm based on spectral subtraction. We find that our trained neural network quantitatively and qualitatively outperforms the baseline approach.

1. INTRODUCTION

Reverberation is an effect that can be created naturally when a source sound is reflected off various surfaces before reaching an observer (e.g. a microphone). The characteristics of this reverberation are defined by its acoustic environment (the dimensions and objects in a space, their material properties, etc.), as well as the positions of the source and observer in this environment. Though desirable in some creative and musical contexts, reverberation has the overall effect of reducing intelligibility, which may be particularly undesirable for speech applications, such as telecommunications, automated voice systems, and dialogue editing in post-production. *Dereverberation* is the process of automatically removing reverberation from audio signals. It is an extremely difficult problem, as neither the source signal nor the characteristics of the acoustic environment are known a priori. Moreover, moving sources or microphones can induce time-varying reverberation effects and further complicate its removal.

Standard algorithms for speech dereverberation either exploit properties of speech, or attempt to blindly estimate the reverberant channel [1]. In the former case, algorithms explicitly track harmonic content or leverage linear predictive coding (LPC) of speech to estimate components of the underlying direct-path speech signal [2]. In contrast, blind channel estimation methods generally involve an explicitly parameterized model of the reverberation process, techniques for estimating its parameters, and finally, a reverb removal step based on this parameter estimation through some form of inverse filtering. Moreover, when multiple microphones are available, these algorithms leverage beamforming techniques to further improve reverb cancellation [3], though this is not expected to be the case for most practical applications. The performance of existing dereverberation algorithms is limited by the flexibility of their model assumptions and the ability to accurately estimate parameters over a wide range of applicable scenarios. They are also laborious to develop and may require a fair amount of intensive hand-tuning.

Source separation techniques using deep learning have become increasingly popular, particularly for the application of separating speech from non-stationary noise [4]. Arguably the most common source separation approaches involve time-frequency masking, wherein models are trained to estimate the amount of speech and noise present in each spectrogram bin, and accordingly create a time-varying masking filter to separate speech from noise through Wiener filtering [5]. To this end, recurrent neural network (RNN) architectures have been shown to be extremely effective, as they are a natural choice for modeling sequential data [6]. Bidirectional RNNs (BRNNs) can further improve separation quality by performing a forward and backward pass over the data, thus incorporating future temporal context at the cost of offline (non-realtime) operation [7].

In this paper, we propose the adaptation of blind source separation techniques using deep learning for single-channel dereverberation. One of the main advantages of this method, in addition to its performance, is that we no longer require learning an explicit reverberation profile, but instead simply learn to distinguish dry and reverberant signal components through several synthesized examples observed during network training. With a targeted interest in dialogue editing for post-production, we deliberately limit ourselves to speech signals.

The rest of this paper is structured as follows: We review a baseline dereverberation algorithm based on spectral subtraction [8], and outline our proposed neural network approach in Section 2. We compare the performance of our neural network solution against the baseline algorithm in terms of reverberation reduction and speech intelligibility in Section 3. Finally, we draw conclusions and allude to future work in Section 4.

2. DEREVERBERATION ALGORITHMS

We consider monaural speech signals and model their reverberation by the convolution $y(t) = h(t) * s(t)$, where $y(t)$ is the observed reverberated signal, $h(t)$ is the impulse response of the acoustic environment, factoring the position of the source and observer in said environment, and $s(t)$ is the direct-path speech signal. The dereverberation process is a blind deconvolution in which we attempt to find an estimate $\hat{s}(t)$ from $y(t)$. Rather than operating on the time-domain waveform, both the baseline and proposed neural network solutions operate primarily on the short-time Fourier transform (STFT) magnitude spectrogram of $y(t)$, denoted as $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_T] \in \mathbb{R}^{d \times T}$, where d is the number of frequency bins and T is the number of STFT time frames.

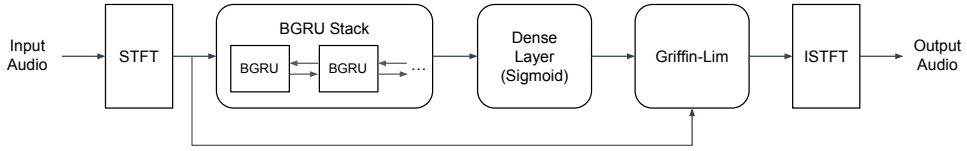


Figure 1: *Speech dereverberation neural network architecture.*

2.1. Baseline Algorithm

The baseline dereverberation algorithm uses short-time spectral attenuation with directly calculated spectral masks. In this case, the multi-path signal is modeled as a first-order recursive filter that "smears" the spectrogram in time, defined as

$$\mathbf{r}_t = \alpha \mathbf{s}_t + (1-\alpha)\mathbf{r}_{t-1}, \tag{1}$$

where t is the STFT frame index, $\mathbf{s}_t$ is the magnitude spectrum of the direct path signal, $\mathbf{r}_t$ is the magnitude spectrum of the multi-path signal, and α is a coefficient related to the reverberation time (RT_{60}). This implies an exponential model for the reverberation process, which is considered to be an adequate assumption for a large number of reverberation types. The observed reverberant signal $\mathbf{y}_t$ at time t is a mixture of the direct and multi-path signals defined as

$$\mathbf{y}_t = \mathbf{s}_t + \beta \mathbf{r}_t, \tag{2}$$

where β is the wet-to-dry ratio coefficient controlling the relative amount of reverberation. Assuming that the reverberation parameters α and β are known, we can easily invert (1) and (2) to compute an estimate of the dry signal spectral magnitude $\hat{\mathbf{s}}_t$ from $\mathbf{y}_t$. The obtained signal-to-noise ratio $\hat{\mathbf{s}}_t/\mathbf{y}_t$ can be used in a spectral attenuation algorithm with extra time-frequency smoothing for reduction of "musical noise" artifacts [8]. In our implementation, α is frequency-independent, while β is independently estimated in 4 frequency bands. Parameter estimation is an offline process which analyzes an entire audio waveform. The parameter α is estimated from a histogram of spectral decay rates, while β is estimated in each frequency band such that the resulting spectral subtraction maximally reduces the energy of $\mathbf{y}_t$ while remaining non-negative.

2.2. Neural Network Algorithm

For any practical application, we can consider impulse responses $h(t)$ normalized such that $h(0) = 1$. This allows us to reformulate the reverberation process as

$$y(t) = s(t) + h_0(t) * s(t) = s(t) + n(t), \tag{3}$$

where

$$h_0(t) = \begin{cases} 0, & t = 0, \\ h(t), & \text{otherwise} \end{cases} \tag{4}$$

and $n(t)$ is the multi-path signal. As such, we have converted the problem of dereverberation into one of blind source separation, and can leverage advances in deep learning which we have successfully used to this end [7, 9]. Note that, in contrast to (2), the frequency-dependent wet-to-dry ratio β is simply embedded in $n(t)$ and its respective magnitude spectrum $\mathbf{n}_t$ (e.g. $\mathbf{n}_t = \beta \mathbf{r}_t$). In general, one of the attractive features of a deep learning-based approach is that we can forego the need to define an explicit parameterization of the reverberation process, easing design and giving flexibility to perform well over a wider range of possible reverberation types.

Given a training set with examples of isolated direct-path and multi-path speech signals, we create mixtures with known ground truth to learn a mapping that estimates the direct-path signal $\hat{s}(t)$ from the reverberated mixture $y(t)$. We use the magnitude ratio mask as a time-varying filter for separating dry and reverberated speech, which is defined as

$$\mathbf{m}_t = \frac{\mathbf{s}_t}{\mathbf{s}_t + \mathbf{n}_t}, \tag{5}$$

where the division operation in (5) is performed element-wise. Because magnitude spectra $\mathbf{s}_t$ and $\mathbf{n}_t$ are nonnegative, the mask elements $\mathbf{m}_t$ are in the interval $[0, 1]$. The output of our neural network is $\hat{\mathbf{m}}_t$, which we use to obtain estimated magnitude spectra for separated direct-path and multi-path speech, i.e.,

$$\hat{\mathbf{s}}_t = \hat{\mathbf{m}}_t \odot \mathbf{y}_t, \tag{6}$$

$$\hat{\mathbf{n}}_t = (\mathbf{1} - \hat{\mathbf{m}}_t) \odot \mathbf{y}_t, \tag{7}$$

where $\odot$ represents an element-wise product. For inference, we use (6) to obtain the estimated time-domain waveform $\hat{s}(t)$ through the inverse STFT. This usually involves using phase information taken from the noisy mixture $y(t)$, which can introduce some noticeable artifacts. To improve upon this, we use the mixture phase as our initial estimate of the direct-path phase, and apply a few iterations of the Griffin-Lim algorithm [10]. Though not strictly necessary for inference (unless we, for some reason, want to retain a portion of the multi-path signal in the processed output), we still make use of (7) for network training.

We estimate $\hat{\mathbf{m}}_t$ using a bidirectional recurrent neural network architecture as depicted in Figure 1. At 48 kHz, spectrograms are computed with Hann-windowed FFTs of size 2048 and a stride of 512. A stack of bidirectional gated recurrent units (BGRU) [11] take the reverberated mixture spectrogram as input, and produce outputs which incorporate temporal context from both past and future spectrogram frames. We opt for a non-causal architecture because their lookahead capabilties allow them to perform better than their causal counterparts, and because the baseline approach already required an initial offline learning pass. The output of the BGRU stack is projected to the appropriate number of frequency bins by means of a dense layer, whose sigmoid activation ensures that spectral masks are in the desired [0, 1] range. Several studies on neural network-based speech separation have shown the utility of using the error in the estimated spectrum $\hat{\mathbf{s}}_t$ (as opposed to the error in the estimated mask $\hat{\mathbf{m}}_t$) as the network training objective [4, 5, 12]. To this end, a common objective function for source separation compares $\hat{\mathbf{s}}_t$ and $\mathbf{s}_t$ in a mean-squared sense. As

Table 1: *(*SDR/SI-SNR*) metrics as a function of SNR.*

Method	−5dB	0dB	5dB	10dB	15dB
Mixture	−4.2/ − 5.3	0.4/ − 0.2	5.4/4.9	10.3/ 9.9	15.3/15.0
Baseline	−2.3/ − 3.6	2.6/ 1.7	7.0/6.2	11.1/10.4	15.7/15.1
Proposed	**1.3/ − 0.4**	**5.6/ 4.5**	**9.6/8.7**	**13.7/13.1**	**17.6/17.1**
Oracle	4.8/ 3.1	8.1/6.9	11.6/10.7	15.3/14.5	19.1/18.5

Table 2: ΔSTOI *metric as a function of SNR.*

Method	−5dB	0dB	5dB	10dB	15dB
Baseline	4.5	2.9	0.6	0.4	0.3
Proposed	**16.2**	**9.3**	**3.8**	**1.5**	**0.5**
Oracle	38.6	17.9	7.3	2.8	1.1

in [4, 7], we use the modified mean-squared error function

$$J = \frac{1}{T}\sum_{t=1}^{T}\Big(||\hat{\mathbf{s}}_t - \mathbf{s}_t||_2^2 + ||\hat{\mathbf{n}}_t - \mathbf{n}_t||_2^2 - \gamma||\mathbf{s}_t - \hat{\mathbf{n}}_t||_2^2 - \gamma||\mathbf{n}_t - \hat{\mathbf{s}}_t||_2^2\Big), \quad (8)$$

where the parameter γ provides a trade-off between interference and artifacts caused by the source separation process. We found that the addition of cross-term penalties helped to improve dereverberation performance at lower wet-to-dry ratios.

3. EXPERIMENTAL RESULTS

3.1. Dataset description

While there are several speech datasets available for machine learning research, most of them are band-limited (usually sampled at 16 kHz), and are insufficient for the full audio rate processing needs of post-production. With a target sampling rate of 48 kHz in mind, we have opted to use speech from the pitch tracking corpus in [13], the processed speech from the DAPS experiments [14], and the TSP speech dataset [15]. We have also supplemented our clean speech training with several hours of audio from iZotope tutorial videos. While these are not truly anechoic speech recordings, they were found to be representative enough to serve as "ideal" dereverberated outputs of our system. In addition to the speech dataset, we have gathered reverb impulse responses (RIRs) from many open sources. We have also developed a RIR generator factoring different reverb characteristics (RT_{60}, wet-to-dry ratio, etc.), and added hundreds of simulated RIRs to our dataset. We considered reverberation types and parameter ranges that resemble the naturally occurring environments that the system was targeted for.

3.2. Performance assessment

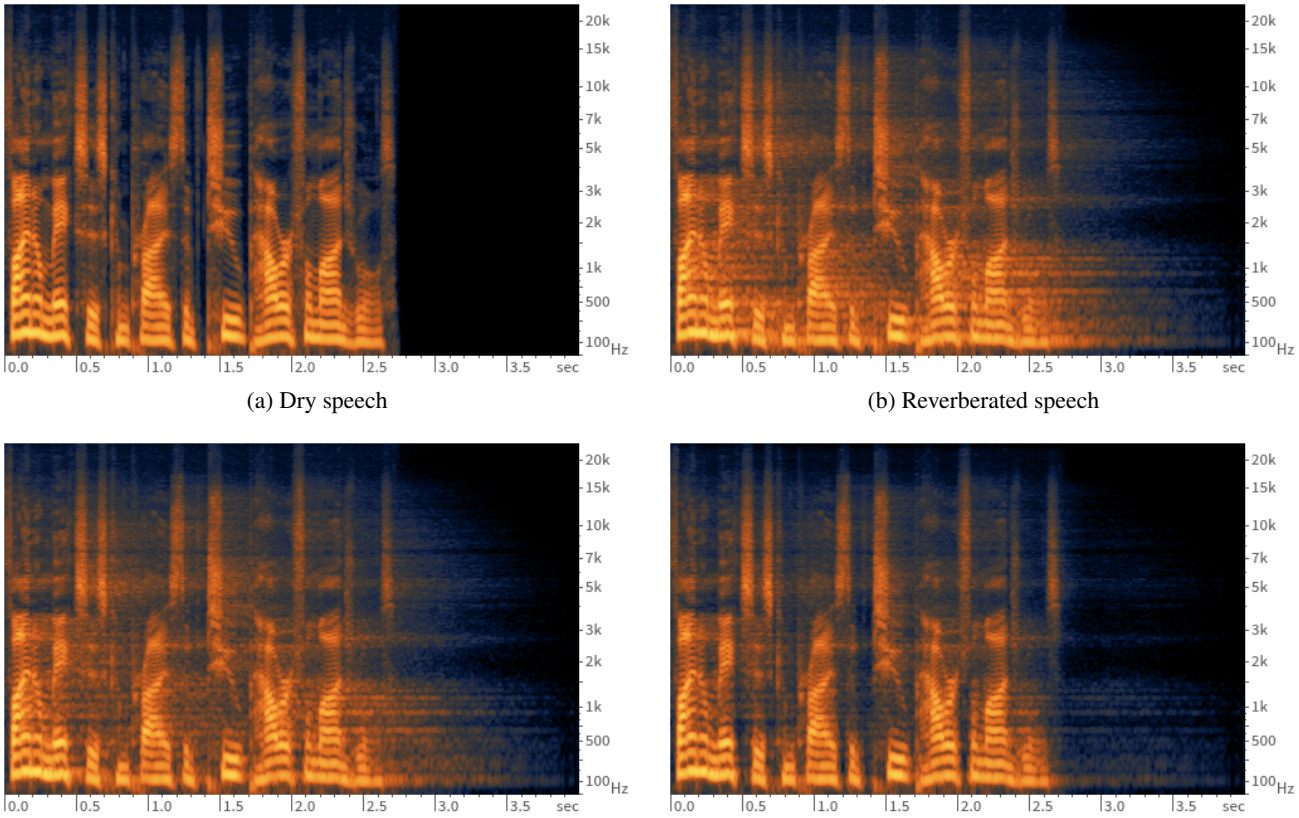
We quantitatively compared the performance of the baseline and our proposed neural network dereverberation algorithms on a test dataset consisting of audio from held-out speakers and RIRs not seen during training. Algorithm performance was evaluated at a number of different SNRs (i.e. dry-to-wet ratios), ranging from -5 to $+20$ dB. To evaluate performance in terms of reverberation reduction, we used the signal-to-distortion ratio (SDR) and signal-invariant signal-to-noise ratio (SI-SNR) [16]. For completeness, we computed metrics on the original mixture signals, as well as the results of speech separation using oracle (ground truth) magnitude ratio masks, essentially specifying the expected lower and upper performance bounds for our methods. Additionally, we used the difference of the short-time objective intelligibility measure (STOI) [17] between the processed output and the original reverberant mixtures converted to a percentage ΔSTOI. This measures the overall percent improvement in speech quality and intelligibility relatively to the original reverberant mixtures. Our choice of metrics attempt to quantify algorithm performance through both "standard" and perceptually-driven means.

Table 1 and 2 illustrates our quantitative performance evaluation in terms of reverb reduction and speech quality improvement, respectively. We can observe that the baseline approach clearly improves upon the original reverberant mixture. Our proposed neural network solution consistently outperforms the baseline approach, and as to be expected, performs a few dB worse than the oracle mask solution. The ΔSTOI confirms that speech intelligibility is not as degraded at higher SNRs. These results suggest that our proposed neural network solution can recover about half of the possible of improvement in speech intelligibility relative to the oracle solution, and it outperforms the baseline approach by a large margin in this regard.

We performed informal listening tests, both on our synthesized evaluation set and on real-world speech signals that are naturally reverberated. We observe that our neural network solution can reduce more reverberation than the baseline approach, while remaining rather transparent in its processing and reducing "pumping" artifacts often heard in the baseline approach. The spectrograms in Figure 2 provide a visual comparison between the baseline and proposed approaches on a speech sample with synthetically applied reverberation. We can see that while the baseline approach improves upon the reverberated speech, the proposed approach yields an output that more closely resembles the underlying dry speech. For audio examples and additional spectrograms, please visit http://www.izotope.com/tech/dafx_dereverb .

3.3. Generalization to non-speech signals

Though we have explicitly trained our dereverberation network on speech signals, we have informally noticed that the system can generalize to some classes of non-speech signals. This is particularly fortuitous for our post-production application, where there may be other sound effects, laughter, etc. that may be desirable to salvage in a given performance.



(a) Dry speech

(b) Reverberated speech

(c) Dereverberation of (b) using the baseline approach

(d) Dereverberation of (b) using the proposed approach

Figure 2: Speech dereverberation comparison.

4. CONCLUSIONS

In this paper, we proposed a novel application of source separation to the problem of speech dereverberation, and trained a BRNN to this end. Our proposed solution outperformed a baseline approach based on spectral subtraction through both qualitative and quantitative means. In addition to its improved performance, a benefit of our deep learning approach is that we no longer need to formulate an explicit model for reverberation, and can avoid hand-tuned estimation of reverb parameters altogether. In the future, we would like to research effective low-latency solutions, and additionally consider time-domain architectures which may be able to more accurately observe and remedy early reflections.

5. ACKNOWLEDGMENTS

We would like to thank the anonymous reviewers, whose comments greatly improved the quality of this manuscript.

6. REFERENCES

[1] P. Naylor and N. D. Gaubitch, Eds., *Speech Dereverberation*, Springer, New York, NY, USA, 2010.

[2] B. Yegnanarayana and P. Satyanarayana, "Enhancement of reverberant speech using lp residual signal," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 8, pp. 267–281, May 2000.

[3] E. A. P. Habets and J. Benesty, "A two-stage beamforming approach for noise reduction and dereverberation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 945–957, May 2013.

[4] P.S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Transactions on Speech and Language Processing*, vol. 23, pp. 2136–2147, Dec. 2015.

[5] F. Weninger, J. Le Roux, J. R. Hershey, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *IEEE GlobalSIP 2014 Symposium on Machine Learning Applications in Speech Processing*, 2014, pp. 577–581.

[6] Z. Q. Wang and D Wang, "Recurrent deep stacking networks for supervised speech separation," in *International Conference on Acoustics, Speech and Signal Processing*, New Orleans, LA, USA, Mar. 5-9 2017.

[7] G. Wichern and A. Lukin, "Removing lavalier microphone rustle with recurrent neural networks," in *Proc. Digital Audio Effects (DAFx-2018)*, Aviero, Portugal, Sept. 4-8 2018, pp. 19–25.

[8] A. Lukin and J. Todd, "Suppression of musical noise artifacts in audio noise reduction by adaptive 2d filtering," in *123rd Audio Engineering Society Convention*, New York, NY, USA, Jan. 2007.

[9] G. Wichern and A. Lukin, "Low-latency approximation of bidirectional recurrent networks for speech denoising," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, USA, Oct. 15-18 2017.

[10] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, Apr. 1984.

[11] K. Cho et al., "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Conference on Empirical Methods in Natural Language Processing*, 2014.

[12] Y. Wang, A. Narayanan, and D. L. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, pp. 1849–1858, 2014.

[13] G. Pirker, M. Wohlmayr, S. Petrik, and F. Pernkopf, "A pitch tracking corpus with evaluation on multipitch tracking scenario," in *Interspeech*, 2011, pp. 1509–1512.

[14] G. J. Mysore, "Can we automatically transform speech recorded on common consumer devices in real-world environments into professional production quality speech? A dataset, insights, and challenges," *IEEE Signal Processing Letters*, vol. 22, pp. 1006–1010, 2015.

[15] P. Kabal, "TSP speech database," Tech. Rep., Department of Electrical & Computer Engineering, McGill University, Montreal, Quebec, Canada, 2002.

[16] J. Le Roux, J. R. Hershey, S. T. Wisdom, and H. Erdogan, "SDR: half-baked or well done?," Tech. Rep., Mitsubishi Electric Research Laboratories, Cambridge, MA, USA, 2018.

[17] J. Chen et al., "Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises," *Journal of the Acoustical Society of America*, vol. 139, pp. 2604–2612, 2016.