LOW-LATENCY APPROXIMATION OF BIDIRECTIONAL RECURRENT NETWORKS FOR SPEECH DENOISING

Gordon Wichern and Alexey Lukin

iZotope, Inc.
Cambridge, MA 02139, USA
gwichern, alukin@izotope.com

ABSTRACT

The ability to separate speech from non-stationary background disturbances using only a single channel of information has increased significantly with the adoption of deep learning techniques. In these approaches, a time-frequency mask that recovers clean speech from noisy mixtures is learned from data. Recurrent neural networks are particularly well-suited to this sequential prediction task, with the bidirectional variant (e.g., BLSTM) achieving strong results. The downside of bidirectional models is that they require offline operation to perform both a forward and backward pass over the data. In this paper we compare two different low-latency bidirectional approximations. The first uses block processing with a regular bidirectional network, while the second uses the recently proposed lookahead convolution layer. Our results show that using just 1000 ms of backward context can recover approximately 75% of the performance improvement gained from using bidirectional as opposed to forward-only recurrent networks.

Index Terms— Speech enhancement, source separation, time-frequency masking, bidirectional recurrent networks, lookahead convolution

1. INTRODUCTION

While traditional speech enhancement techniques, such as spectral subtraction [1], work well on stationary background noise, removal of non-stationary disturbances remains extremely challenging. This is especially true when working with only a single channel of audio where spatial techniques are not available. In the single-channel case, a source separation approach is typically adopted to separate speech from the non-stationary background. Popular approaches include non-negative matrix factorization (NMF) [2] and time-frequency mask estimation [3].

The recent explosion of research in deep neural networks has led to several techniques that learn from data how to create non-linear functions that estimate time-frequency masks for separating speech from difficult non-stationary background disturbances, such as music or even other speech [4, 5, 6, 7, 8, 9, 10]. In [5] pre-processed log spectral magnitude coefficients are passed through a fully-connected feedforward multi-layer neural network. A similar feedforward architecture using a collection of extracted speech features as input is proposed in [4, 6]. The feedforward architectures of [5, 6] deal with the sequential nature of input audio by concatenating multiple surrounding frames into a single input vector to predict the separation mask for the frame of interest.

A more natural approach to incorporate context from a sequential input audio stream is to use a *recurrent neural network* (RNN) [8, 9, 10]. The RNN architecture stores an internal hidden state, and is, in essence, trained to remember past network inputs when making predictions. In [8] a basic RNN architecture is successfully used for speech denoising in addition to other source separation tasks. Backpropagation through time is necessary to train a RNN, and a special gated hidden unit architecture is required to avoid vanishing or exploding gradients for long sequences. The most popular of these gated hidden unit architectures is the *long short-term memory* (LSTM) [11], which formed the basis of the speech denoising systems proposed in [9, 10]. Recently the *gated recurrent unit* (GRU) [12] has begun to show better performance with fewer parameters than the LSTM on sequential prediction tasks such as speech recognition [13].

While forward-only RNNs use only past inputs when making decisions, *bidirectional RNNs* incorporate future context [14]. Bidirectional RNNs have demonstrated state-of-the-art performance in speech recognition [13] and speech/noise separation tasks [9]. However, because they perform both a forward and backward pass over the data, they require offline operation. This increase in buffering time can be unacceptable in many speech denoising applications, such as hearing aid development, a front end for speech recognition, or audio post production. The desire to maintain the performance improvements of a bidirectional RNN, while operating at low latency, motivated the development of the *lookahead convolution layer* [15] for speech recognition systems. This layer is inserted after a stack of forward-only RNN layers, and learns the optimal weights for incorporating future context using a convolution-like operation, which was found to generalize better than a regular convolution layer.

In this paper we evaluate the source separation performance of the lookahead layer in separating speech from non-stationary background noise. The lookahead layer is incorporated in a deep recurrent architecture using stacked GRU layers. We also evaluate an approach using regular bidirectional GRU layers operating on blocks of audio in order to lower latency. Furthermore, by evaluating performance degradation as a function of latency, we can glean insight into the amount of future context learned by the bidirectional network during training.

We begin detailing our approach by describing ratio mask estimation using bidirectional recurrent networks in Section 2. The block-based realtime approximation and lookahead convolution layers are discussed in Section 3. We evaluate the performance of the proposed low-latency denoising algorithms in Section 4 on speech corrupted by noise using several publicly available datasets. Finally, conclusions and discussions of future work are provided in Section 5.

2. SPEECH SEPARATION NETWORKS

Given a single channel additive mixture $y(t) = s(t) + n(t)$, where $s(t)$ represents the clean speech signal, and $n(t)$ the non-stationary noise, our goal is to obtain an estimate of the clean speech signal $\hat{s}(t)$ from the noisy mixture $y(t)$. We take a source separation approach operating on STFT magnitude spectrograms, of $y(t)$, $s(t)$, and $n(t)$ defined as $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_T] \in \mathbb{R}^{d \times T}$, $\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_T] \in \mathbb{R}^{d \times T}$, and $\mathbf{N} = [\mathbf{n}_1, \mathbf{n}_2, ..., \mathbf{n}_T] \in \mathbb{R}^{d \times T}$, respectively, where d is the number of frequency bins. We also define the magnitude ratio mask for frame t as

$$\mathbf{m}_t = \frac{\mathbf{s}_t}{\mathbf{s}_t + \mathbf{n}_t} \tag{1}$$

and separation is achieved from

$$\hat{\mathbf{s}}_t = \hat{\mathbf{m}}_t \odot \mathbf{y}_t \tag{2}$$

$$\hat{\mathbf{n}}_t = (\mathbf{1} - \hat{\mathbf{m}}_t) \odot \mathbf{y}_t \tag{3}$$

where $\odot$ represents an element-wise product, and $\hat{\mathbf{m}}_t$ is an estimate of (1) obtained from the noisy mixture $\mathbf{Y}$. The separated speech $\hat{s}(t)$ and noise waveforms $\hat{n}(t)$ are obtained from (2) and (3) with inverse STFT using the phase from the noisy mixture $y(t)$.

In this work we estimate $\hat{\mathbf{m}}_t$ using deep recurrent neural networks, where the output $\mathbf{h}_t^\ell$ of the ℓ-th recurrent layer in a stack of L layers for frame t is defined as

$$\hat{\mathbf{m}}_t = \sigma(\mathbf{W}^L \mathbf{h}_t^{L-1} + \mathbf{b}^L) \tag{4}$$

$$\mathbf{h}_t^\ell = f(\mathbf{h}_t^{\ell-1}, \mathbf{h}_{t-1}^\ell) \tag{5}$$

$$\mathbf{h}_t^1 = f(\mathbf{y}_t, \mathbf{h}_{t-1}^1) \tag{6}$$

where $\sigma(\cdot)$ represents the sigmoid nonlinearity, and $\mathbf{W}^L$ and $\mathbf{b}^L$ are the parameters of the fully connected output layer. The nonlinear mapping of the recurrent layers is signified by $f(\cdot)$. Several types of recurrent units have been proposed in the literature, with gated units such as LSTM [11] and GRU [12] being the most widely used. In this work we use GRU units for $f(\cdot)$, as we found them to provide slightly better performance with fewer parameters than LSTM, a result consistent with speech recognition work [13].

We also use bidirectional recurrent layers [14] defined by

$$\tilde{\mathbf{h}}_t^\ell = f(\mathbf{h}_t^{\ell-1}, \vec{\mathbf{h}}_{t-1}^\ell) + f(\mathbf{h}_t^{\ell-1}, \overleftarrow{\mathbf{h}}_{t+1}^\ell) \tag{7}$$

where the output of a bidirectional layer is the sum of a forward recurrent layer $\vec{\mathbf{h}}_{t-1}^\ell$ and backward recurrent layer $\overleftarrow{\mathbf{h}}_{t+1}^\ell$, which receives its input in time-reversed order and hence requires the network to operate in an offline fashion.

2.1. Training Objective

As described in [4, 8, 10], better separation performance is typically achieved when the error is computed on the output spectra of separated speech $\hat{\mathbf{s}}_t$, as opposed to mask $\hat{\mathbf{m}}_t$, leading to the objective

$$J_{MSE} = \frac{1}{2} \sum_{t=1}^{T} ||\hat{\mathbf{s}}_t - \mathbf{s}_t||_2^2 \tag{8}$$

Following [8], we found that separation performance could be improved by adding a KL divergence regularizer to our objective function that penalizes interference between sources

$$J_{DIS} = \frac{1}{2} \sum_{t=1}^{T} \Big(||\hat{\mathbf{s}}_t - \mathbf{s}_t||_2^2 + ||\hat{\mathbf{n}}_t - \mathbf{n}_t||_2^2 - \gamma ||\mathbf{s}_t - \hat{\mathbf{n}}_t||_2^2 - \gamma ||\mathbf{n}_t - \hat{\mathbf{s}}_t||_2^2 \Big) \tag{9}$$

In this work we set the regularization constant to $\gamma = 10^{-3}$ which worked well in our experiments.

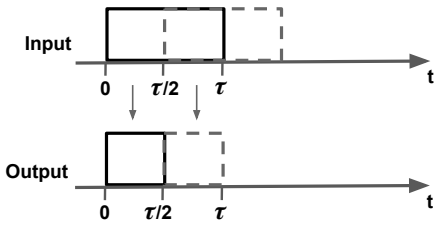

Figure 1: Illustration of the half-overlapping block procedure for bidirectional approximation.

3. LOW-LATENCY BIDIRECTIONAL APPROXIMATIONS

The ability of bidirectional recurrent layers to incorporate future context provides performance benefits at the cost of offline operation. In this section we investigate block-based approaches operating on spectrogram sub-blocks $\mathbf{Y}_{t:t+\tau} = [\mathbf{y}_t, \mathbf{y}_{t+1}, ..., \mathbf{y}_{t+\tau}] \in \mathbb{R}^{d \times (\tau+1)}$ with fixed latency τ.

3.1. Block processing

One approach for operating with low latency is to use a trained bidirectional network, but operate on fixed size-blocks. Assuming no overlap between blocks, the forward portion of each bidirectional layer can be used without modification and the hidden state $\vec{\mathbf{h}}_t^\ell$ from (7) can propagate between blocks. However the hidden state for the backward portion of each bidirectional layer $\overleftarrow{\mathbf{h}}_t^\ell$ from (7) is reset to its initial value every block.

With no overlap between blocks, the amount of backward context available to each frame varies from no backward context at frame $t+\tau$ to τ frames of backward context at frame t. By overlapping blocks we can assure a more consistent amount of backward context between frames, at the cost of additional processing. For overlapping blocks, we also need to store the forward hidden state for each layer corresponding to the start of the next block. We have also found cross-fading between blocks to be unnecessary (assuming the forward layer hidden states are properly propagated), so for half-overlapping blocks we update the mask in blocks like

$$\hat{\mathbf{M}}_{t:t+\frac{\tau}{2}} = \Phi(\mathbf{Y}_{t:t+\tau}) \tag{10}$$

where $\Phi(\cdot)$ represents the mapping learned by the trained bidirectional network. The half-overlapping block procedure is also illustrated graphically in Figure 1.

3.2. Lookahead layer

The lookahead convolution layer proposed in [15] uses only forward recurrent layers and inserts a lookahead convolution layer after the

final recurrent layer to approximate a bidirectional architecture. The lookahead convolution layer is defined as

$$\mathbf{h}_t = \phi\left(\sum_{j=1}^{\tau+1} \mathbf{w}_j \odot \mathbf{x}_{t+j-1}\right) \quad (11)$$

where $\phi(\cdot)$ is the layer nonlinearity, a hyperbolic tangent in this work, and $\mathbf{x}_t \in \mathbb{R}^n$ is the lookahead layer input, which corresponds to the output from a stack of forward recurrent layers. The lookahead layer is parameterized by weight matrix $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_{\tau+1}] \in \mathbb{R}^{n \times \tau+1}$. The lookahead convolution layer has been shown to generalize better than a fully convolutional layer for speech recognition in [13].

4. RESULTS

4.1. Experimental setup

To compare the performance of the low-latency bidirectional approximations investigated in this paper we've gathered training and testing data from multiple sources. For training clean speech we use the pitch tracking corpus from [16] and the reverberant speech from the Chime challenge [17]. For training non-stationary background noise we use that available from the Chime corpus [17], which was recorded in a home and contains sounds such as children speaking, television, etc. We also use the audio scene classification data from the DACSE 2016 challenge [18], which contains 15 different acoustic environments, with sounds such as traffic noise and babble recorded in a cafe. For testing we use the processed speech utterances from the DAPS experiment [19] and noise from the first DCASE challenge [20], which was an entirely different collection experiment from the 2016 test set used for training.

For training, we created mixtures up to 12 seconds in length by randomly concatenating multiple speech utterances, prior to mixing with the background scenes. The mixture SNR for our training and test sets varied from -6 to $+9$ dB. The input and output of our network are 1025 spectral magnitude coefficients obtained from 48-kHz sampled audio. All of our testing mixtures were 12 seconds in length.

Our network architecture is based on GRU layers as we found them to perform similarly to LSTM layers, but with fewer parameters making them cheaper to train and evaluate. Our bidirectional architecture consists of two bidirectional hidden layers with 512 units followed by the fully connected output layer which produces the magnitude ratio mask $\hat{\mathbf{m}}_t$. As a forward-only benchmark, we use four 512-unit GRU layers, which have an equivalent number of parameters since the bidirectional layer defined in (7) is actually a sum of a forward and backward layer each with their own parameters. We stop training after 20,000 mini-batches of 16 sequences.

We evaluate speech separation performance using the source-to-distortion ratio (SDR), source-to-interference ratio (SIR), and source-to-artifact ratio (SAR) metrics from BSS-EVAL [21]. The metrics were implemented using the package from [22]. The SDR is equivalent to the SNR of the separated speech, SIR measures the background noise remaining in the separated speech signal, and SAR quantifies artifacts introduced by the separation process. The metrics are defined in a way that distortion equals interference plus artifacts.

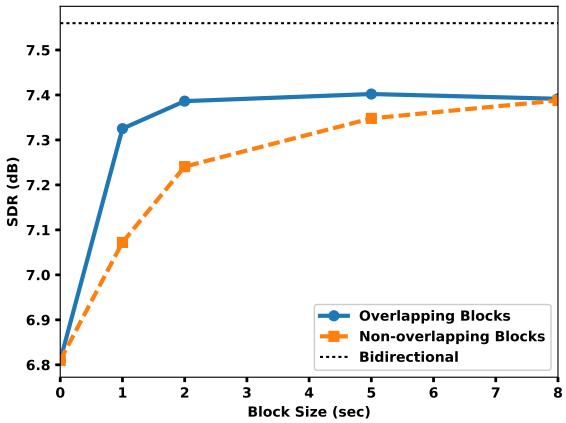

Figure 2: SDR of block processing on a previously trained bidirectional RNN with non-overlapping and half-overlapping blocks. Input mixtures are 12 seconds long and have SNR of 3dB.

4.2. Block size analysis

We evaluate the approach described in Section 3.1 using block-based processing with a previously trained bidirectional network. Figure 2 displays the average SDR of the separated speech as a function of block size for 3-dB SNR test set mixtures. A block size of zero seconds is the forward-only four-layer GRU architecture. For convenience, the offline bidirectional architecture which processes the entire 12-second mixtures is shown as a dotted line, even though it does not depend on the block size. For the overlapping blocks we use 50% overlap between blocks.

From Figure 2, we notice that for smaller block sizes overlapping blocks noticeably improve SDR, at the cost of increased computational load from performing the backward layer computations two times. Frames near the edges between non-overlapping blocks have little backward context available, which accounts for this drop in performance. We also note that the performance curve for overlapping blocks seems to level-off for block sizes larger than two seconds. This helps to quantify the amount of backward context learned by our network. For the two-second half-overlapping block case, every frame has at least one second of backward context. Since increasing the overlapping block size beyond two seconds does not improve performance greatly, we can conclude that the backward layers exploit approximately one second of backward context. Given that a single word in a speech signal is on the order of one second in length, it appears that the network is learning that this amount of future context is important. However, the offline bidirectional architecture still has benefits, which may be due to the background noise evolving at much longer time scales, or the network learning to exploit sentence-level context.

4.3. Lookahead layer evaluation

We now evaluate adding a lookahead layer as described in Section 3.2 between a stack of forward-only recurrent layers and the fully connected output layer. We explore future context of $\tau = 20$ and $\tau = 100$ frames, which correspond to approximately 213 ms and 1067 ms, respectively. Figure 3 displays the average SDR per-

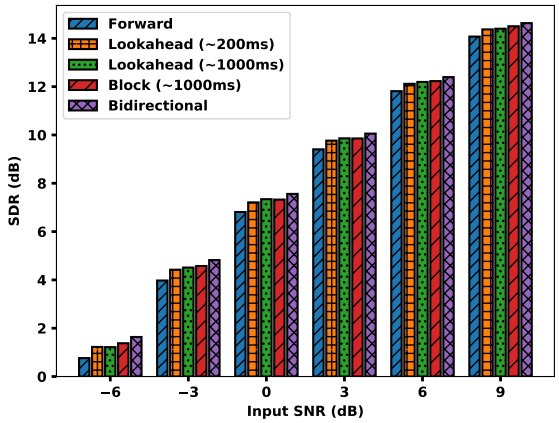

Figure 3: SDR as a function of input mixture SNR for several network configurations.

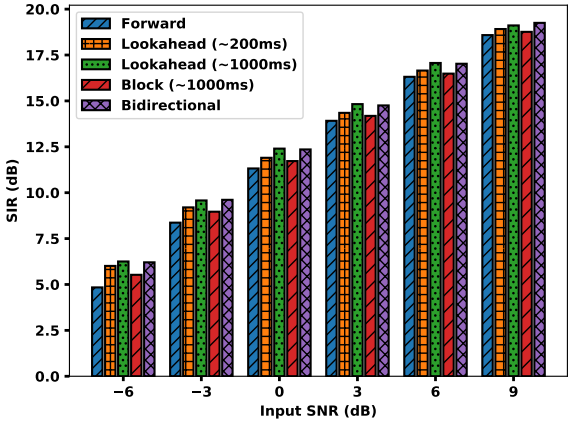

Figure 4: SIR as a function of input mixture SNR for several network configurations.

formance on our test set for input mixture SNR values between -6 to $+9$ dB in 3 dB steps. For comparison, we also include the forward-only architecture and the bidirectional architecture operating both in offline fashion and evaluated with 1000 ms overlapping blocks at test time. From Figure 3, relative performance between network architectures is consistent over the input mixture SNR range, since our network was trained over this range of SNR. As expected, the forward-only network exhibits the smallest SDR for separated speech, while the offline bidirectional network has the highest output SDR.

From Figure 3, we see that both lookahead layer architectures exhibit similar performance, even though one is incorporating five times the amount of future context. This is similar to the finding in [15] for speech recognition. We also note that using the bidirectional network with 1000 ms overlapping blocks exhibits similar performance to the lookahead layer architecture. We can also see from Figure 3 that the low-latency approximations recover approximately 75% of the performance lost when going from a bidirectional to a forward-only recurrent speech denoising architecture.

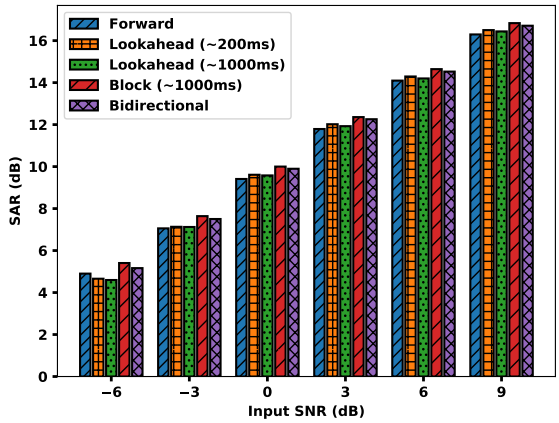

Figure 5: SAR as a function of input mixture SNR for several network configurations.

Figures 4 and 5 evaluate performance in terms of SIR and SAR, respectively. From Figure 4 we see that the larger lookahead layer with future context of approximately 1000 ms performs almost equivalently to the offline bidirectional network in terms of SIR, while the one-second overlapping block approach performs the worst. The results are reversed for the SAR shown in Figure 5, with the 1000 ms overlapping block approach exhibiting the highest values and the 1000 ms lookahead layer configuration exhibiting the lowest. These results suggest that the overlapping block-based approximation to bidirectional networks should be preferred in applications where artifacts are not desirable, but more noise can remain in the separated speech. In applications where maximum interference removal is desired, the lookahead layer approach may be preferable at the cost of increased artifacts.

5. CONCLUSIONS AND FUTURE WORK

Source separation approaches to single-channel speech denoising in non-stationary backgrounds have benefited tremendously from advances in deep recurrent neural networks. Bidirectional architectures provide performance gains, but don't meet the low-latency requirements necessary for realtime applications. This paper has evaluated approximations of offline bidirectional networks with block-based processing or by adding a lookahead convolution layer, and showed that performance close to offline bidirectional architectures can be obtained at latencies of 1000 ms or less. In the future we plan to investigate why bigger lookahead convolution layers do not provide significant performance improvements, and close the gap in understanding how deep bidirectional recurrent architectures exploit future context.

6. REFERENCES

[1] P. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL: CRC Press, 2013.

[2] B. Raj, T. Virtanen, S. Chaudhuri, and R. Singh, "Non-negative matrix factorization based compensation of music for automatic speech recognition," in *Interspeech*, 2010, pp. 717–720.

[3] D. Wang, "Time-frequency masking for speech separation and its potential for hearing aid design," *Trends in Amplification*, vol. 12, pp. 332–353, 2008.

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

[5] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, pp. 7–19, 2015.

[6] D. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, pp. 483–492, 2016.

[7] J. Chen, Y. Wang, S. Yoho, D. Wang, and E. Healy, "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.

[8] P. 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.

[9] H. Erdogan, J. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 708–712.

[10] F. Weninger, J. Le Roux, J. 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.

[11] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 23, no. 8, pp. 1735–1780, 1997.

[12] K. Cho, B. van Merrinboer, C. Glehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," in *Conference on Empirical Methods in Natural Language Processing*, 2014.

[13] D. Amodei, R. Anubhai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, J. Chen, M. Chrzanowski, A. Coates, G. Diamos, and E. Elsen, "Deep speech 2: End-to-end speech recognition in English and Mandarin," in *Proceedings of The 33rd International Conference on Machine Learning*, 2016, pp. 173–182.

[14] M. Schuster and K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, pp. 2673–2681, 1997.

[15] C. Wang, D. Yogatama, A. Coates, T. Han, A. Hannun, and B. Xiao, "Lookahead convolution layer for unidirectional recurrent neural networks," in *Workshop Extended Abstracts of the 4th International Conference on Learning Representations*, 2016.

[16] 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.

[17] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, "The PASCAL CHiME speech separation and recognition challenge," *Computer Speech and Language*, vol. 27, pp. 621–633, 2013.

[18] A. Mesaros, T. Heittola, and T. Virtanen, "Tut database for acoustic scene classification and sound event detection," in *24th European Signal Processing Conference (EUSIPCO 2016)*, 2016.

[19] G. 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.

[20] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and classification of audio scenes and events," *IEEE Transactions on Multimedia*, vol. 17, pp. 1733–1746, 2015.

[21] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 1462–1469, 2006.

[22] C. Raffel, B. McFee, E. Humphrey, J. Salamon, O. Nieto, D. Liang, and D. Ellis, "mir_eval: A transparent implementation of common mir metrics," in *Proceedings of the 15th International Conference on Music Information Retrieval*, 2014.