LOW LATENCY SOUND SOURCE SEPARATION USING CONVOLUTIONAL RECURRENT NEURAL NETWORKS

Gaurav Naithani[1] , *Tom Barker*[1], *Giambattista Parascandolo*[1]†
Lars Bramsløw[2], *Niels Henrik Pontoppidan*[2], *and Tuomas Virtanen*[1]

[1]Tampere University of Technology, Department of Signal Processing, Tampere, Finland
Email:{gaurav.naithani, giambattista.parascandolo, thomas.barker, tuomas.virtanen}@tut.fi

[2]Eriksholm Research Centre, Oticon A/S, Snekkersten, Denmark
Email:{labw, npon}@eriksholm.com

ABSTRACT

Deep neural networks (DNN) have been successfully employed for the problem of monaural sound source separation achieving state-of-the-art results. In this paper, we propose using convolutional recurrent neural network (CRNN) architecture for tackling this problem. We focus on a scenario where low algorithmic delay ($\leq$ 10 ms) is paramount, and relatively little training data is available. We show that the proposed architecture can achieve slightly better performance as compared to feedforward DNNs and long short-term memory (LSTM) networks. In addition to reporting separation performance metrics (i.e., source to distortion ratios), we also report extended short term objective intelligibility (ESTOI) scores which better predict intelligibility performance in presence of non-stationary interferers.

Index Terms— Source Separation, Low-latency, Deep Neural Networks, Convolutional Recurrent Neural Networks.

1. INTRODUCTION

Source separation is a field of research which aims to solve the problem of separating an audio mixture into its constituent sounds originating from different sources. It is useful for various applications like automatic speech recognition [1, 2], fundamental frequency estimation [3], etc., where it acts as an intermediary step which aids in the final objective of the task.

In this paper, we focus on monaural sound source separation problem for applications where low processing latency is paramount, e.g., hearing aids [4], and cochlear implants [5, 2]. It is postulated that window duration of around 20-40 ms in short time Fourier transform (STFT) processed signals is preferred for speech processing [6]. But for low-latency systems we need to work with shorter window durations. In the context of digital hearing aid applications, low processing delay is regarded as a critical design feature [7]. For such applications, latencies as low as 3 ms have been found to be detectable and anything larger than 10 ms have been found to be objectionable [8]. There is therefore a need for sound source separation algorithms which are suitable for algorithmic delays $\leq$ 10 ms.

Recently deep neural network (DNN) based approaches for sound source separation have become quite popular [9, 10, 11] and different types of architectures for the task have been reported. Feedforward DNNs, e.g., reported in [9], are unable to utilize long temporal contexts and whatever temporal context is deemed useful, has to be explicitly fed to the input in the form of stacked frames, e.g., in [12]. They are also unable to explicitly utilize the spectro-temporal structure present in time-frequency representation of audio signal which is lost if stacked frames are used as input. The need to infuse information from long temporal context motivates the use of recurrent neural networks, as has been reported in [10, 11]. The need to preserve spectro-temporal structure motivates the use of convolutional neural networks, which have been applied in conventional form e.g., in [13] , or as convolutional encoder-decoder networks, e.g, in [14, 15]. However most of these methods have been reported for algorithmic latencies $\geq$ 20 ms.

Previously, we have investigated the problem of low -latency sound source separation for algorithmic latencies $\leq$ 20 ms using non-negative matrix factorization [16] and feedforward deep neural networks [12]. This paper takes that work forward and investigates the potential of convolutional recurrent networks (CRNN) [17, 18] for this problem. Such architectures have been successfully employed, e.g., in polyphonic sound event detection [18], and music classification [19].

In this paper we compare the performance of convolutional recurrent networks (CRNN) with feedforward DNNs and long short-term memory (LSTM) networks [20]. Typically, deep learning techniques have been shown to directly benefit from very large quantities of training data [21], although in some cases either this is not always available, or costly to obtain. We therefore consider the scenario where training data is limited, and attempt to maximize performance in these cases. This is also motivated by the constraints of a project where we are using the techniques reported here for hearing impaired (HI) listeners (see Section 3.1). This user-centric application also motivates the use of an estimated intelligibility metric, as well as purely energy based separation measures. In addition to the conventionally used BSS-EVAL [22] metrics of separation, we evaluate the performance of our approach using the extended short time objective intelligibility metric (ESTOI) [23]. It is an extension of the popular short time objective intelligibility (STOI) metric [24], and is postulated to be a better predictor of intelligibility in presence of modulated noise/speech interferers (see Section 3.3).

The paper is organized as follows: Section 2 describes the CRNN architecture along with the time-frequency masking scheme utilized in this paper. Section 3 describes the evaluation procedure, experimental design along with the acoustic material used in the experiments. And finally, Section 4 concludes the paper.

[*] The authors wish to thank CSC-IT Centre of Science Ltd., Finland, for providing computational resources used in experiments reported in this paper.

[†] The author is currently with Max Planck Institute for Intelligent Systems.

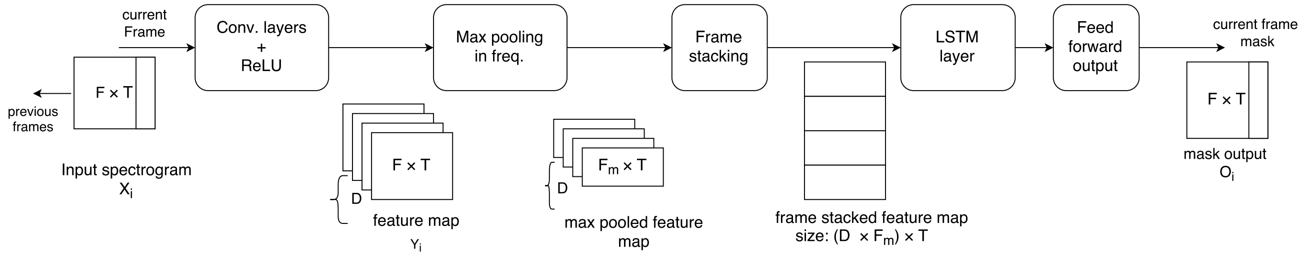

Figure 1: Block diagram of the proposed convolutional recurrent neural network architecture.

2. PROPOSED METHOD

The proposed method uses the time-frequency masking based source separation paradigm [4] which involves prediction of time-frequency masks corresponding to constituent sources present in the mixture. These time-frequency masks when applied to mixture spectrograms yield spectra of the constituent sources. Magnitudes of short time Fourier transform (STFT) coefficients are used as features. The neural network is trained in a supervised fashion and mean square error (MSE) between target and estimated masks, $\sum_{t,f}(M(t,f) - M_{est}(t,f))^2$ is used as the training objective function, where t and f are time and frequency indices, respectively.

In order to maintain low latency operation, the proposed neural network utilizes only past temporal context to predict mask corresponding to the current frame. The algorithmic latency is limited by the window size used for STFT processing.

2.1. Proposed neural network architecture

Convolutional recurrent neural networks (CRNN) are a combination of two kinds of neural network topologies: convolutional and recurrent networks. The motivation behind combining the two is to take advantage of the feature extraction capability of the former with temporal modeling capability of the latter.

The proposed architecture is shown in Figure 1. Convolutional layers form the front end of the network to which spectral features X_i , $i = 1, 2, ...n$, are fed. Each input X_i is a $(F \times T\ )$ matrix consisting of a sequence of T temporally continuous spectral feature vectors, each of size F. The convolutional layer act as feature extractor employing $(k_r \times k_c)$ size convolutional kernel to efficiently extract spectro-temporal features from the input. This procedure can be thought of as convolving the input feature matrix with the convolutional kernel and yields a $(F \times T)$ size matrix called an *activation map* or *feature map*. D such convolutional kernels are used give a three dimensional output, Y_i, of size $(D \times F \times T)$. Note that the convolution operation here is such that, for t^{th} frame of the input, the convolution operation utilizes only previous time frames in order to maintain the low latency operation. Figure 2 depicts this for one dimensional convolution example. Y_i is then fed to a pooling layer where max pooling operation is done only over the frequency axis. Max pooling operation reduces the size of the frequency axis of each of the D feature maps from F to F_m, reducing the number of parameters fed into the next layer. Several such convolutional layers in combination of pooling layers can be employed.

The second component of a CRNN architecture is a recurrent layer. Before feeding max pooled Y_i into the recurrent layer, D feature maps are stacked along the frequency axis such that for each time step we now have $D \times F_m$ size feature vector. This operation preserves the temporal continuity of the T frames and expands the feature set for each time frame by a factor of D. In this paper, we have used long short-term memory (LSTM) [20] units in the recurrent layer. The recurrent layer takes a sequence of T frames as input and produces a sequence of T output frames. Several such recurrent layers can be employed. The output from the recurrent layer is fed to a feedforward layer which is a timedistributed layer, i.e., it processes each of the T frames independently of the other frames, and serves as the output layer of the network.

2.2. Mask outputs and source reconstruction

We utilize soft time-frequency mask in this paper. For an acoustic mixture of two sources, mask corresponding to *source 1* can be expressed as,

$$M^1(t,f) = \frac{|S_1(t,f)|}{|S_1(t,f)| + |S_2(t,f)|}, \tag{1}$$

where S_1 and S_2 are STFT magnitudes corresponding to the constituent sources. The model is trained with respected to *source 1* hence the estimated mask $M^1_{est}(t,f)$ corresponds to *source 1* and mask corresponding to the other source is given by its complement, i.e., $M^2_{est} = 1 - M^1_{est}$. These estimated masks are then used to get complex STFT spectra of the separated sources utilizing the mixture phase. The time domain constituent sources are then reconstructed using inverse discrete Fourier transform (IDFT) and overlap add processing.

3. EVALUATION

The section describes the acoustic material used in the experiments, metrics used for evaluating separation performance of neural network architectures, experimental design, and finally the results obtained.

3.1. Acoustic Material

The Danish hearing in noise test (HINT) dataset was used for evaluation which is an extended version of the original HINT dataset (described in [25]). The dataset consists of 13 lists, each consisting of 20 natural sentences. Each sentence consists of 5 words. All audio files were recorded with a sampling rate of 44.1 kHz. Three speaker pairs: M1 and F1, M1 and M2, and, F1 and F2, were chosen for evaluation to cover all gender combinations. This study is a part of a larger project to investigate potential intelligibility benefits of sound separation methods for HI listeners. Danish HINT dataset is used for subjective listening experiments which involve

Table 1: Best hyperparameters and number of trainable parameters for the selected FDNN, LSTM and CRNN architectures.

	FDNN			LSTM			CRNN					
	hidden layers	hidden neurons	previous context	hidden layers	hidden neurons	sequence length	conv. layers	recurr. layers	recurr. neurons	conv. filters	pooling scheme	sequence length
	4	1024	8	3	512	64	3	1	256	256	*1 by 2*	128
par.	3,904,593			5,208,581			4,089,681					

repeated presentation of sentences to HI listeners which often leads to word memorization in test subjects. It is thus desirable to reserve only a small part of the dataset for training the DNNs and keep the larger chunk for testing. This study thus is useful for all situations where training data is not easily available. We use lists L1 and L2 for training, lists L3 and L4 for validation, and lists L7 and L8 for testing. This amounts to 40 sentences (around 80 seconds) of audio data each for training and validation.

3.2. Training data generation

Each separation model was trained with data from the target speaker in the corresponding speaker pair on which it was to operate. All audio data was resampled to 16 kHz before any further processing.

A limited amount of data was used for training (2 lists of 20 utterances per speaker from the Danish HINT corpus), and so careful construction of training material from the audio data was required. The available training material was systematically mixed in various permutations to attempt to make maximum use of the available data, by repeatedly summing the entire training data for each of the speakers within a training/test pair with a varying offset. The offset was applied in the STFT domain, by circularly shifting one source spectrogram (over time), with respect to the other, and summing. Hence for each offset shift, a novel set of training examples was produced. Initial validation experiments showed that although not as effective at boosting performance as increasing the quantity of initial training data, (e.g. using a greater fraction of the training corpus), modest improvements could be achieved by augmenting the training data in this way, when faced with a restricted quantity of training data. We used 50 shifts of length $\frac{T_s}{50}$, where T_s is the number of frames in the longer of the two training spectrograms to produce our training data spectrograms.

Training features were produced from the training data spectrograms, such that the target output for a particular sequence of input frames was the mask given by Equation 1. For the production of all STFT features, an analysis window length 5 ms (80 samples at 16 kHz) was used with 50% overlap, resulting in a 5 ms algorithmic latency.

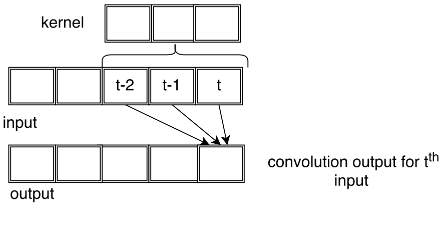

Figure 2: 1-dimensional depiction of the convolution operation for a single kernel as used in CNN layers.

3.3. Metrics

The separation performance of different neural network architectures were evaluated using BSS-EVAL toolkit [22]. It consists of three metrics: Source to Interference Ratio (SIR) and Source to Artifacts Ratio (SAR), for interference and artifact suppression, respectively; and Source to Distortion ratio (SDR) for overall separation performance. In addition, extended short time intelligibility metric (ESTOI) has been reported, which, unlike STOI [24], does not assume mutual independence between frequency bands of speech and interfering signals [23] and is hence a better predictor of intelligibility for experiments reported in this paper.

3.4. Experimental design

We consider two baselines for the proposed CRNN architecture: 1) A feedforward deep neural network architecture (FDNN) described in [12], where previous temporal context of N frames was used for predicting the output corresponding to current frame. 2) A long short term memory (LSTM) network which has been used for sound source separation, e.g, in [11]. The experimental design consists of two stages: 1) Hyperparameter selection using grid search for the three types of neural network architectures: FDNN, LSTM, and CRNN, using validation set of lists L3 and L4 , and 2) Evaluation of the chosen best versions of each topology selected in step 1 on a common test set (i.e., lists L7 and L8 here). The hyperparameter selection was performed using only speaker pair M1F1 in order to limit the degrees of freedom in the search space, for a reasonable number of experiments.

Hyperparameter selection for feedforward DNN involved selection of appropriate number of hidden layers $\{1, 2, 3, 4\}$, number of neurons in each hidden layer $\{$ 128, 256, 512, 1024, 2048$\}$, and number of previous frames $\{4, 8, 16, 32\}$ to be used in the prediction of the current frame. Similarly, for LSTM network, the hyperparameters consisted of number of hidden layers $\{1, 2, 3, 4\}$ and number of neurons in each hidden layer $\{128, 256, 512, 1024\}$. Finally, CRNN hyperparameter search consisted of number of convolutional layers $\{1, 2, 3, 4\}$, number of LSTM layers $\{$ 0, 1, 2, 3$\}$, and number of convolutional filters $\{64, 96, 128, 256\}$. Frequency max pooling arrangements $\{$ *1 by 2*, *1 by 3*, *1 by 4* $\}$ were experimented with, the first and second dimensions denoting time and frequency axes, respectively. Here, e.g., a max pooling scheme of *1 by 2*, implies that the frequency dimension is reduced by a factor of 2. Note that max pooling is done only in the frequency dimension. Moreover, for the LSTM and CRNN, we also investigated appropriate sequence length $\{$8, 16, 32, 64, 128, 256 frames$\}$. The sequence length here refers to number of frames that are unrolled for backpropagation through time. For CRNN the following convolutional

Table 2: Performance metrics for the three neural network architectures.

		Network topology		
Speaker pair	Metric	FDNN	LSTM	CRNN
M1 F1	SIR	10.39	10.68	10.99
	SAR	10.69	10.69	10.7
	SDR	7.23	7.4	7.54
	ESTOI	0.76	0.78	0.79
M1 M2	SIR	9.56	9.85	10.22
	SAR	9.96	9.98	9.92
	SDR	6.42	6.55	6.74
	ESTOI	0.76	0.77	0.79
F1 F2	SIR	8.23	8.55	9.16
	SAR	9.8	9.53	9.78
	SDR	5.49	5.51	6.01
	ESTOI	0.70	0.71	0.74

kernels shapes were investigated , { (3, 3), (5, 5) , (7, 7)}. Here, e.g., (3, 3) denotes kernel size in time and frequency axes.

The best possible hyperparameters for the three topologies was selected based on the separation performance (i.e., SDR) on the validation data. Table 1 shows the final hyperparameters selected for final evaluation. A max pooling layer was used after each convolutional layer. A convolutional kernel of size (3, 3) and max pooling scheme of *1 by 2* was used for final evaluation. Interestingly, in our tests, each of the network architectures giving best performance on the validation set had similar order of trainable parameters, with roughly 3.9, 5.2 and 4.1 million parameters across the FDNN, LSTM and CRNN architectures respectively.

Some choices of network parameters were not included in the hyperparameter search and were kept constant for all networks. These choices were empirically determined on the basis of preliminary experiments, observing the performance on validation data. For feedforward DNNs, the sigmoid activation function was used for both hidden and output neurons, and dropout regularization [26] of 0.4 was used to counter overfitting; for CRNNs rectified linear units (ReLU) activation for hidden layers and sigmoid activation for the output layer was used, with a dropout rate of 0.4; for recurrent layers, the standard LSTM cells as described in [27] were used. Other design choices include using batch normalization [28] after each feedforward/convolutional layer in FDNN/CRNN. The Adam algorithm [29] was used for gradient descent optimization. An early stopping criterion [30] was used to stop training when no further improvement in validation loss occurred for 25 epochs. The Librosa [31] library was used in feature extraction and Keras [32] neural network library was used for training the neural networks.

3.5. Results

The separation and intelligibility performance metrics were computed over 400 mixtures produced from the evaluation set lists L7 and L8, for the three speaker pairs: M1F1, M1M2, and F1F2. Table 2 shows the evaluation scores for the three speaker pairs. Best performance was achieved on the M1F1 speaker pair most likely due to the greater spectral difference between between male and female speakers, whilst F1F2 yielded the lowest separation metrics. CRNNs performed slightly better than the baseline architectures for speaker pairs M1F1 and M1M2. For speaker pair F1F2, CRNNs perform notably better than the other architectures showing 0.5 dB improvement over the baseline. Similar improvement in terms of intelligibility are indicated by ESTOI scores as well, going from 0.70 to 0.74.

For each of the various network topologies, quite different configurations produced optimal results. For the feedforward DNN, only 8 frames of previous context were used at the input, whereas in LSTM and CRNN architectures, significantly greater previous temporal context was used in calculation of the output, with sequence lengths of 64 and 128 frames respectively, yet still all network architectures had similar order of parameters to be trained, despite varying contexts.

4. CONCLUSION AND FUTURE WORK

In this paper, we showed the potential of convolutional recurrent neural networks (CRNN) for the task of monaural sound source separation for applications requiring low algorithmic latency. We compared the proposed neural network architecture to feedforward DNNs and LSTM networks and showed that the proposed method performs slightly better than these networks, whilst using far fewer parameters to model the same temporal context.

CRNNs thus offer a promising alternative to the baseline architectures. In this work, only square convolutional kernels were considered, (as it is generally used in image processing domain) but that might not be optimal for the the task of source separation. Future work would involve a more thorough investigation of the effects of possible hyperparameters on the proposed CRNN architecture for source separation. Additionally, the relationship between training data quantity, and network architecture and parameters is an interesting problem warranting further investigation. The observed objective improvements achieved through the proposed method should be further verified through listening experiments.

5. REFERENCES

[1] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen, “Exemplar-based sparse representations for noise robust automatic speech recognition,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2067–2080, 2011.

[2] K. Kokkinakis and P. C. Loizou, "Using blind source separation techniques to improve speech recognition in bilateral cochlear implant patients," *The Journal of the Acoustical Society of America*, vol. 123, no. 4, pp. 2379–2390, 2008.

[3] E. Gómez, F. J. Cañadas-Quesada, J. Salamon, J. Bonada, P. V. Candea, and P. C. Molero, "Predominant fundamental frequency estimation vs singing voice separation for the automatic transcription of accompanied flamenco singing." in *International Society for Music Information Retrieval Conference*, 2012, pp. 601–606.

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

[5] J. Hidalgo, "Low latency audio source separation for speech enhancement in cochlear implants," Master's thesis, Universitat Pompeu Fabra, 2012.

[6] K. K. Paliwal, J. G. Lyons, and K. K. Wójcicki, "Preference for 20-40 ms window duration in speech analysis," in *International Conference on Signal Processing and Communication Systems*. IEEE, 2010, pp. 1–4.

[7] L. Bramsløw, "Preferred signal path delay and high-pass cut-off in open fittings," *International Journal of Audiology*, vol. 49, no. 9, pp. 634–644, 2010.

[8] J. Agnew and J. M. Thornton, "Just noticeable and objectionable group delays in digital hearing aids," *Journal of the American Academy of Audiology*, vol. 11, no. 6, pp. 330–336, 2000.

[9] E. M. Grais, M. U. Sen, and H. Erdogan, "Deep neural networks for single channel source separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2014, pp. 3734–3738.

[10] P. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2014, pp. 1562–1566.

[11] F. Weninger, J. R. Hershey, J. Le Roux, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *IEEE Global Conference on Signal and Information Processing*. IEEE, 2014, pp. 577–581.

[12] G. Naithani, G. Parascandolo, T. Barker, N. H. Pontoppidan, and T. Virtanen, "Low-latency sound source separation using deep neural networks," in *IEEE Global Conference on Signal and Information Processing*, 2016.

[13] P. Pertilä and E. Cakir, "Robust direction estimation with convolutional neural networks based steered response power," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017.

[14] S. R. Park and J. Lee, "A fully convolutional neural network for speech enhancement," *arXiv preprint arXiv:1609.07132*, 2016.

[15] P. Chandna, M. Miron, J. Janer, and E. Gómez, "Monoaural audio source separation using deep convolutional neural networks," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2017, pp. 258–266.

[16] T. Barker, T. Virtanen, and N. H. Pontoppidan, "Low-latency sound-source-separation using non-negative matrix factorisation with coupled analysis and synthesis dictionaries," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015.

[17] P. Pinheiro and R. Collobert, "Recurrent convolutional neural networks for scene labeling," in *International Conference on Machine Learning*, 2014, pp. 82–90.

[18] E. Cakır, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," *arXiv preprint arXiv:1702.06286*, 2017.

[19] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Convolutional recurrent neural networks for music classification," *arXiv preprint arXiv:1609.04243*, 2016.

[20] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[21] M. Banko and E. Brill, "Scaling to very very large corpora for natural language disambiguation," in *Proceedings of the 39th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2001, pp. 26–33.

[22] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[23] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.

[24] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *IEEE International Conference on Acoustics Speech and Signal Processing*. IEEE, 2010, pp. 4214–4217.

[25] J. B. Nielsen and T. Dau, "The Danish hearing in noise test." *International journal of audiology*, vol. 50, no. 3, pp. 202–8, mar 2011.

[26] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[27] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," *Neural Computation*, pp. 2451–2471, 2000.

[28] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*, 2015, pp. 448–456.

[29] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[30] R. Caruana, S. Lawrence, and C. Giles, "Overfitting in neural networks: Backpropagetion. conjugate gradient, and early stopping," *Neural Information Processing Systems*, pp. 402–408.

[31] B. McFee *et al.*, "librosa 0.5.0," Feb. 2017. [Online]. Available: https://doi.org/10.5281/zenodo.293021

[32] F. Chollet *et al.*, "Keras," 2015. [Online]. Available: https://github.com/fchollet/keras