Localization Based Stereo Speech Separation Using Deep Networks

Yang Yu*, Wenwu Wang†, Jian Luo*, Pengming Feng†
*School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an, China, 710072
Email: nwpuyuy@nwpu.edu.cn
†Centre for Vision Speech and Signal Processing, University of Surrey, UK, GU2 7XH
Email: W.Wang@surrey.ac.uk

***Abstract*—Time-frequency (T-F) masking is an effective method for stereo speech source separation. However, reliable estimation of the T-F mask from sound mixtures is a challenging task, especially when room reverberations are present in the mixtures. In this paper, we proposed a new stereo speech separation system where deep networks are used to generate soft T-F mask for separation. More specifically, the deep network, which is composed of two sparse autoencoders and a softmax classifier, is used to estimate the orientations of the target and interferers at each T-F unit, based on low-level features, such as mixing vector (MV), interaural level and phase difference (IPD/ILD). The deep network is trained by a greedy layer-wise method using a dataset that was generated by convolving room impulse responses (RIRs) with clean speech signals positioned in different angles with respect to the sensors. With the trained deep networks, the probability that each T-F unit belongs to the target or interferer can be estimated based on the localization cues for generating the soft mask. Experiments based on real binaural RIRs and TIMIT dataset are provided to show the performance of the proposed system for reverberant speech mixtures, as compared with a model based T-F masking technique proposed recently.**

***Index Terms*—Deep learning; Deep networks; Source separation; Soft mask;**

I. INTRODUCTION

Speech separation provides a useful front-end for hearing aids and automatic speech recognition systems. Many methods have been applied to this problem, such as independent component analysis (ICA) [1], [2], [3], beamforming [4], and computation auditory scene analysis (CASA) [5], [6]. Time-frequency (T-F) masking is an effective method for speech source separation, where the mask can be derived from various cues such as mixing vector (MV) [7], interaural phase and level difference (IPD/ILD) [8] and their combination [9], based on a Gaussian mixture model (GMM) whose parameters are estimated iteratively using an expectation maximization (EM) algorithm. These methods provide a nice probabilistic framework for incorporating complementary information to deal with the uncertainties in T-F assignment. However, the performance of these algorithms is limited by the accuracy of model-fitting especially when room reverberation is present.

In this paper, we present a new approach for T-F assignment and mask estimation based on the emerging technique of deep neural network [10]. The network is trained with the low-level of features (i.e. MV and ILD/IPD) extracted from a training set of observed signals. In the separation stage, the trained network is used to estimate the orientation of the target and interferers which is further exploited to derive the source occupation probability (and thereby the mask) at each T-F unit of the mixture. Our experimental results show that the proposed method performs significantly better than the GMM/EM based baseline method [9] in terms of both signal to distortion ratio (SDR) and perceptual evaluation of speech quality (PESQ).

The remainder of the paper is organized as follows. Section II briefly discusses the related works. Section III outlines the proposed system. Section IV discusses the low-level features to be used as inputs to the network. Section V presents the details about the deep network, including its structure, the training method, and how it is used for separation. Section VI shows the experiments using real RIRs and TIMIT data before the conclusion is drawn in Section VII.

II. RELATION TO PRIOR WORK

Several recent works have explored the potential of using deep neural networks for speech separation. In [11], Wang et al. proposed a supervised learning approach to monaural segregation of reverberant speech with the features called multi-resolution cochleagram (MRCG) [12]. In [13], the features extracted by gammatone filters are used to train the multilayer perceptron and to generate a binary mask, where target speech and noise sources from different locations are used as training sets. The dataset and ground truth i.e. ideal binary mask (IBM) for training are generated by Praat [14], [15], and the output of the classifier i.e. the multi-layer perceptron is an estimated IBM.

Our method is built on the work in [9] where GMM is used to model the MV and IPD/ILD cues, and the EM algorithm is used to estimate the model parameters and to derive the T-F mask. Therefore, stereo source separation problem is also considered here. Instead of using GMM and EM, however, we use deep networks to estimate the source occupation likelihood at each T-F point. More specifically, we use pre-trained sparse autoencoder to extract high-level features (i.e. spatial information of the sources) from the T-F representation of the mixtures and use the softmax regression to generate the soft mask. In addition, the data set that we used for deep

978-1-4799-8058-1/15/$31.00 ©2015 IEEE.

networks training is composed of the observed speech signals from different directions with respect to the sensors.

III. SYSTEM OVERVIEW

Our proposed system consists of the following four stages: (1) extraction of the low-level features (i.e. MV and ILD/IPD) (details in Section IV); (2) training of the deep networks (details in Section V); (3) estimation of the probabilities that each T-F unit belongs to different sources and generation of the soft mask (details in Section V); and (4) reconstruction of the target signal from the soft mask and mixture signal. The system architecture is shown in Figure 1.

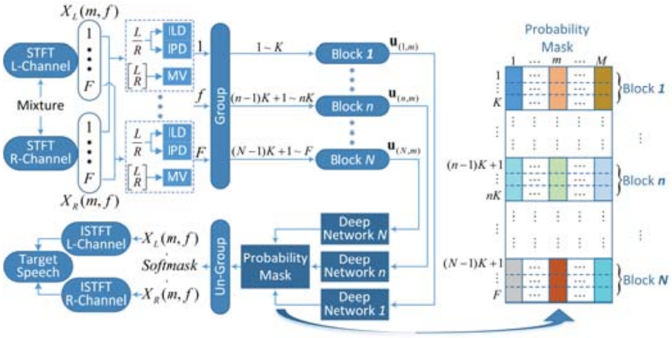

Fig. 1. The architecture of the proposed system using deep neural network based time-frequency masking for stereo source separation.

As shown in Figure 1, the inputs to the system are the stereo (left and right) channel speech mixtures. We perform short-time Fourier transform (STFT) for each channel, and obtain the T-F representation of the input signals, $X_L(m, f)$ and $X_R(m, f)$ where $m = 1, \cdots, M$ and $f = 1, \cdots, F$ are the time frame and frequency bin indices respectively. The low-level features, i.e. MV and IPD/ILD, are then estimated at each T-F unit (details in Section 4). Next, we group the low-level features into N blocks (only along the frequency bins f). The block n includes K frequency bins $((n-1)K+1, \cdots, nK)$, where $K = \frac{F}{N}$. We build N deep networks with each corresponding to one block and use them to estimate the direction of arrivals (DOAs) of the sources. Through unsupervised learning and the sparse autoencoder [16] in deep networks, high-level features (coded positional information of the sources) are extracted and used as inputs for the output layer (i.e. the softmax regression) of the networks. The output of softmax regression is a source occupation probability (i.e. the soft mask) of each block (through the un-group operation, T-F units in the same block are assigned with the same source occupation probability) of the mixtures. Then the sources can be recovered applying the softmask to the mixtures followed by the inverse STFT (ISTFT).

The deep networks are pre-trained by using a greedy layer-wise [17] training method based on a dataset containing observed speech signals (sources convolved with RIRs) from different directions with respect to the sensors.

The N deep networks in our proposed system have the same architecture, and the details about the architecture and training method can be found in Section V.

IV. THE FEATURES FOR CLASSIFICATION

The quality of speech separation can be improved using combined features of IPD/ILD and MV [9]. These low-level features are used to derive high-level features which are easy to classify through the sparse autoencoder. The MV and the IPD/ILD cues are derived from the mixtures. The MV [7] can be derived as

$$\mathbf{z}(m,f) = \frac{\mathbf{W}(f)\widetilde{\mathbf{x}}(m,f)}{\|\mathbf{W}(f)\widetilde{\mathbf{x}}(m,f)\|} \tag{1}$$

with $\widetilde{\mathbf{x}}(m,f) = \frac{[X_L(m,f), X_R(m,f)]^T}{\|[X_L(m,f), X_R(m,f)]^T\|}$, where $\mathbf{W}(f)$ is a whitening matrix, with each row being one eigen vector of $E(\widetilde{\mathbf{x}}(m,f)\widetilde{\mathbf{x}}^H(m,f))$, the superscript H is Hermitian transpose, and $\|\bullet\|$ is Frobenius norm. ILD and IPD are the phase and amplitude difference between the left and right channel, and calculated as follows [8].

$$\alpha(m,f) = 20\log_{10}\left(\left|\frac{X_L(m,f)}{X_R(m,f)}\right|\right) \tag{2}$$

$$\phi(m,f) = \angle\left(\frac{X_L(m,f)}{X_R(m,f)}\right) \tag{3}$$

where $|\bullet|$ takes the absolute value of its argument, and $\angle(\bullet)$ finds the phase angle.

Concatenating the MV and ILD/IPD features, a feature vector can be obtained at each T-F unit, which is $\widetilde{\mathbf{u}}(m,f) = \left[\mathbf{z}^T(m,f), \alpha(m,f), \phi(m,f)\right]^T \in \mathbb{R}^4$. Then we group all the feature vectors $\widetilde{\mathbf{u}}(m,f)$ into N blocks (only along the frequency bins). For each block, we get a $4K$-dimensional feature vector $\mathbf{u}_{(n,m)} = \left[\widetilde{\mathbf{u}}^T(m,(n-1)K+1), \cdots, \widetilde{\mathbf{u}}^T(m,nK)\right]^T \in \mathbb{R}^{4K}$, as the input to the deep networks.

V. THE DEEP NETWORKS

This section discusses how the network used in Section III is constructed and trained.

A. Architecture of the Deep Network

The deep network we used is shown in Figure 2, which is composed of one input layer, two hidden layers using sparse autoencoders [16] and one output layer using softmax classifier. We choose the sigmoid function as the activation function, and the number of inputs equals the dimension of the feature vectors calculated in Section IV. In order to extract high-level features from low level ones, same as in [18], we used two sparse autoencoders as the first and second hidden layer of the proposed deep network. The two hidden layers are composed of V neuron unit and 1 bias unit. For stereo speech separation, the target direction/orientation is a natural choice of the output of the network. We built a softmax classifier which contains J ranges of directions as the output layer. The output of each range (unit) s_j gives the probability $p(y = Q_j|u)$ of the orientation y of the current input block u belonging to the rang of Q_j, where j is the orientation index as shown in Figure 3. Note that, we split the whole space to J ranges with respect to the sensors and assume that target and interferers

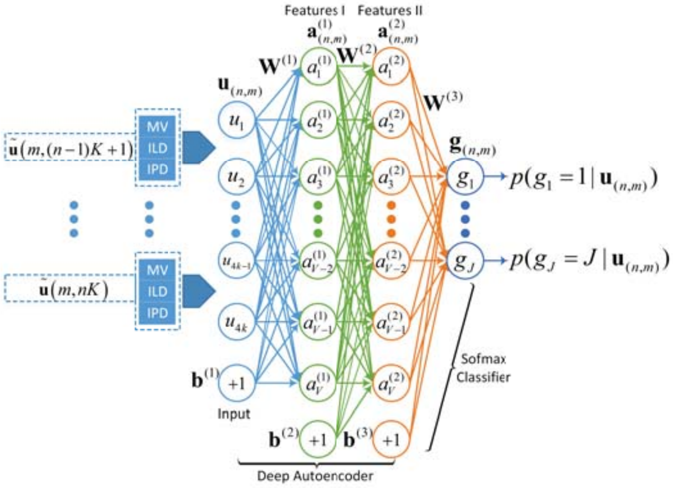

Fig. 2. The architecture of the deep neural networks in our proposed system.

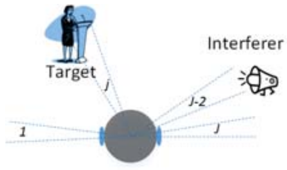

Fig. 3. The directions of arrivals of the sources are split to J ranges.

are at different ranges. Assuming that the position of the target in the current input sample remains unchanged, we can estimate the orientation (that gives the maximum probability) and the number of sources according to the output vector of the network (by using a predefined probability threshold, typically chosen as 0.1), and thus obtain the soft mask from such spatial information for separation. Each T-F unit in the same block is assigned the same probability.

B. Training Method

Similar to [18], [19], the training of the deep network is divided into two stages of pre-training and fine-tuning. In the pre-training phase, an efficient way to obtain optimized parameters for the deep network is to use greedy layer-wise training [17] and the Limited-BFGS (L-BFGS) algorithm [20]. To do this, we train the networks layer by layer, and use the output of each layer as the input for the next layer. The unlabeled data which is composed of observed signals (not speech mixtures, but the individual source convolved with RIRs) are used to train the two sparse autoencoders, and the outputs of sparse autoencoder are used to train the softmax classifier. We use the orientation information of the unlabeled data to build ground truth for the softmax classifier training. If the individual source in the observed signals belongs to rang j, we will set $p(y = Q_j|\mathbf{u}) = 1$ while others such as $p(y = Q_1|\mathbf{u})$ are all set to zero. Through the supervised learning, we can get a nonlinear function set between the observed signal and the probability - $p(y = Q_j|\mathbf{u})$, $j = 1, \cdots, J$. In greedy layer-wise training, the parameters of each layer are trained individually while fixing the parameters of the other layers of the network. After pre-training is completed, we can get the set of network parameters as the initialized parameters for the fine-tuning step. In this step, we use back-propagation algorithm [21] and the same data set used in the pre-training phase to obtain the global optimized parameters for the whole deep networks.

VI. EXPERIMENTS

In this section, we evaluate our proposed system and compare it with the GMM/EM baseline [9].

A. Experimental Setting

We use real BRIRs [22] and TIMIT dataset [23] to generate the training set and test set. As shown in Table I, one advantage of the BRIRs dataset is that they were measured in rooms with different acoustic properties which facilitate the comparison of the system over different conditions [9].

TABLE I
ROOM ACOUSTIC PROPERTIES

Room	Type	ITDG (ms)	DRR (dB)	T60 (s)
A	Medium office	8.73	6.09	0.32
B	Small class room	9.66	5.31	0.47
C	Large lecture theatre	11.9	8.82	0.68
D	Large seminar room	21.6	6.12	0.89

For each room, we randomly select 8 sentences from two speakers from TIMIT, convolve each of these sentences with BRIRs (from -90° to $+90^\circ$ with a step of 5°), and use them as the training set. Therefore, each signal in the training set is a clean speech signal convolved with RIRs. For the test set, we randomly select different speakers and 16 different sentences (8 from female and 8 from male, one female and male speaker from different dialect regions as target and another two as interferers) which are then convolved with BRIRs as the test dataset. The mixture for the test set were generated by adding the reverberant target and interferer signals which is equivalent to assuming superposition of their respective sound fields [9]. Target and interferers were 1.5m away from the dummy head and had a same height as the dummy head. Even though all the sources (including both target and interferers) at different azimuths are recovered in our proposed system, the performance of the system is reported based on the quality of the recovered target located at 0° azimuth, with the interferers azimuth varied from -90° to $+90^\circ$ with a step of 5°), similar to [9]. The sampling rate was f_s = 16kHz. We used a Hann window of 2048 (128ms) samples with 75% overlap between the neighboring windows for the STFT. The frequency grouping parameters K and N are set to 16 and 128. Hence, we use 128 deep networks to generate the soft mask, with each deep network corresponding to a block. For each deep network, the input layer includes 64 units and the output layer includes $J = 37$ ranges (correspond -90° to $+90^\circ$ with a step of 5°). For the two hidden layers, we use $V = 256$ units for both two hidden layers. With the real BRIRs and TIMIT data, we generated 875 T-F units for each orientation (totally 32375 T-F units for 37 orientations) for training. The learning parameters are set as follows, the weight decay parameter $\lambda = 1 \times 10^{-4}$, the weight of sparsity penalty term $\beta = 3$, the sparsity parameter $\rho = 0.3$. The maximum number of iterations is set to 400. The ground truth for the softmax classifier training is a $37 \times M$ matrix, and M is the number of time frames. The 37-dimensional vector in the ground truth represents the orientation of the current sample. The parameters for the softmax training are set as follows. The weight decay parameter $\lambda = 1 \times 10^{-4}$, and the maximum number of iterations is set to 200. In the fine-tuning phase, the weight decay parameter $\lambda = 3 \times 10^{-3}$. For speech separation performance evaluation, we considered SDR [24] and PESQ [25].

B. Experimental Results

We fix the target (random selected speakers) at the azimuth $0\,°$, and the interferer (random selected speakers, different from the target) at the different azimuth $-90\,°$ to $+90\,°$ (except $0\,°$) with a step of $5\,°$.

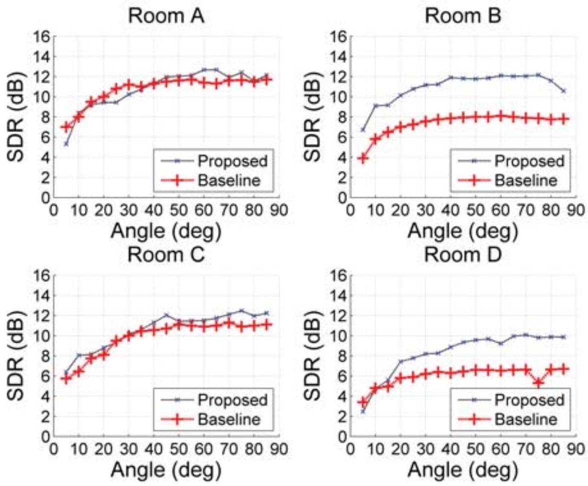

Fig. 4. SDRs comparison between the proposed system and the baseline method for rooms A, B, C, and D.

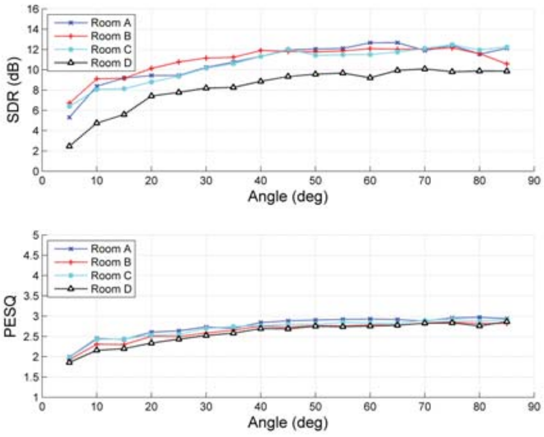

Fig. 5. SDRs and PESQs performance comparison among the four rooms.

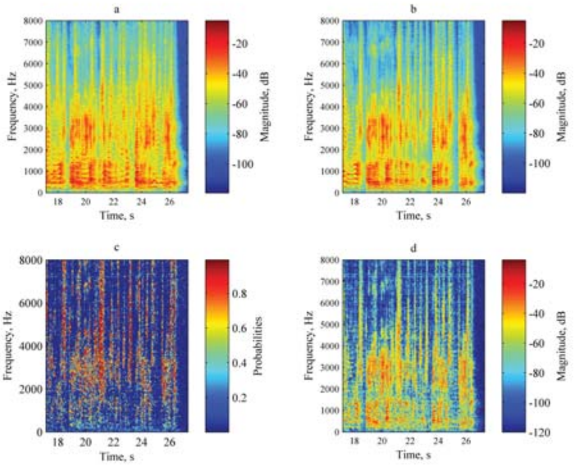

Fig. 6. A separation example for room D, including the amplitude spectrogram of the mixture signals, original target source signals, separated target source signals and soft mask for separation.

Figure 4 presents the SDRs of the separated signals with different DOAs of the interferer and different rooms. Compared with the baseline, we obtain at least 2 dB improvement in room B and D, when the interfering speech is placed far away from the target source. However, we obtain similar performance to the baseline in room A and C. It can be seen that with different reverberation times (T60s) and direction to reverberation ratios (DRRs), the proposed system performs generally more robust than the baseline method, and the performance of the proposed system does not decrease as much as the baseline method when the level of room reverberation increases. Figure 5 presents the SDRs and PESQs performance comparison for the four rooms. Similar to [9], it can be seen that, the separation quality depends on the acoustic parameters T_{60} and DRR.

Figure 6 shows a separation example for room D, including the amplitude spectrogram of the mixture signals (Figure 6a), original target source signals (Figure 6b), separated target source signals (Figure 6d), and the soft mask for separation (Figure 6c) (the interferer was located at $+15\,°$).

VII. Conclusion

We proposed a new localization based stereo speech separation system using deep networks. Compared with GMM/EM based algorithm in [9], the deep networks based technique provide better results in SRD and PESQ.

Acknowledgment

This study was performed when the first author was an academic visitor in the Center for Video Speech and Signal Processing, University of Surrey, and he wishes to thank Qingju Liu, Philip JB Jackson and Atiyeh Alinaghi for providing help for issues related to the algorithm in [9]. This research was supported partially by the Natural Science Basis Research Plan in Shaanxi Province of China (Program No.2014JQ8355).

References

[1] P. Comon and C. Jutten, *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. New York: Academic Press, 2010.

[2] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York: John Wiley & Sons, 2004, vol. 46.

[3] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural Networks*, vol. 13, no. 4, pp. 411–430, 2000.

[4] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, 1988.

[5] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. New York: Wiley-IEEE Press, 2006.

[6] G. J. Brown and D. Wang, "Separation of speech by computational auditory scene analysis," in *Speech Enhancement*. Berlin Heidelberg: Springer, 2005, pp. 371–402.

[7] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 516–527, 2011.

[8] M. I. Mandel, R. J. Weiss, and D. P. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 382–394, 2010.

[9] A. Alinaghi, P. Jackson, Q. Liu, and W. Wang, "Joint mixing vector and binaural model based stereo source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 9, pp. 1434–1448, Sept 2014.

[10] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, Nov 2012.

[11] Z. Jin and D. Wang, "A supervised learning approach to monaural segregation of reverberant speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 625–638, May 2009.

[12] J. Chen, Y. Wang, and D. Wang, "A feature study for classification-based speech separation at very low signal-to-noise ratio," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2014, pp. 7039–7043.

[13] Y. Jiang, D. Wang, R. Liu, and Z. Feng, "Binaural classification for reverberant speech segregation using deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 2112–2121, Dec 2014.

[14] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 5.3.51) [computer program]. retrieved 2 june 2013," 2009. [Online]. Available: http://www.praat.org/

[15] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2002.

[16] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, Aug 2013.

[17] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," *Advances In Neural Information Processing Systems*, vol. 19, p. 153, 2007.

[18] D. Yu and L. Deng, "Deep learning and its applications to signal and information processing [exploratory dsp]," *IEEE Signal Processing Magazine*, vol. 28, no. 1, pp. 145–154, Jan 2011.

[19] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, M. Mao, A. Senior, P. Tucker, K. Yang, Q. V. Le *et al.*, "Large scale distributed deep networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1223–1231.

[20] D. C. Liu and J. Nocedal, "On the limited memory bfgs method for large scale optimization," *Mathematical Programming*, vol. 45, no. 1-3, pp. 503–528, 1989.

[21] Y. Chauvin and D. E. Rumelhart, *Backpropagation: Theory, Architectures, and Applications*. London: Psychology Press, 1995.

[22] C. Hummersone, "A psychoacoustic engineering approach to machine sound source separation in reverberant environments," Ph.D. dissertation, University of Surrey, 2011.

[23] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon Technical Report N*, vol. 93, p. 27403, 1993.

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

[25] L. Di Persia, D. Milone, H. L. Rufiner, and M. Yanagida, "Perceptual evaluation of blind source separation for robust speech recognition," *Signal Processing*, vol. 88, no. 10, pp. 2578–2583, 2008.