A Performance Evaluation of Several Deep Neural Networks for Reverberant Speech Separation

Qingju Liu, Wenwu Wang, Philip JB Jackson, Saeid Safavi
Centre for Vision, Speech and Signal Processing
University of Surrey, UK

***Abstract*—In this paper, we compare different deep neural networks (DNN) in extracting speech signals from competing speakers in room environments, including the conventional fully-connected multilayer perception (MLP) network, convolutional neural network (CNN), recurrent neural network (RNN), and the recently proposed capsule network (CapsNet). Each DNN takes input of both spectral features and converted spatial features that are robust to position mismatch, and outputs the separation mask for target source estimation. In addition, a psychacoustically-motivated objective function is integrated in each DNN, which explores perceptual importance of each TF unit in the training process. Objective evaluations are performed on the separated sounds using the converged models, in terms of PESQ, SDR as well as STOI. Overall, all the implemented DNNs have greatly improved the quality and speech intelligibility of the embedded target source as compared to the original recordings. In particular, bidirectional RNN, either along the temporal direction or along the frequency bins, outperforms the other DNN structures with consistent improvement.**

I. Introduction

Deep neural networks (DNN) have been prevailing in the audio source separation field in the past few years, formulating the conventional "blind" problem with supervised learning [1]. Many different DNN structures have been considered and shown advantages over traditional statistically-characterised source separation algorithms, such as the classic multilayer perception (MLP) [2]–[4], recurrent neural networks (RNN) [5]–[7], convolutional neural networks (CNN) [8], [9]. However, little work has been done in systematic comparisons and evaluations of different DNN structures in source separation tasks for a common setup, e.g. with the same dataset, mixing scenarios and input/output features.

To address this limitation, we implement a variety of DNN-based speech separation methods with different structures, to recover a target speech in the presence of another competing speaker in room conditions. In this paper, we will focus on speaker-independent source separation from recordings collected by a pair of microphones, with one target at the azimuth of 0 degree and an interference at unknown positions. This is a common setup for stereo or binaural recordings, and yet challenging due to the complex nature of both speaker and content-dependent speech, as well as the various mixing scenarios. Note the target azimuth constraint can be compensated via delay-and-sum beamforming. Four DNN topologies will be investigated, including the aforementioned MLP, RNN and CNN, as well as a recently-proposed capsule network (CapsNet) [10]. Particularly for the RNN structure, the bidirectional long short-term memory (BiLSTM) [11] will be applied to input features along both temporal and frequency directions, to exploit the audio correlation along temporal frames and frequency bins respectively.

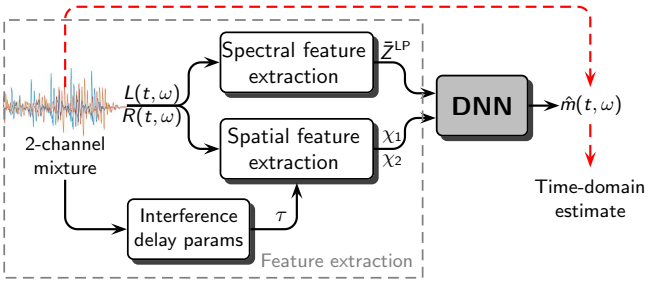

Fig. 1: Diagram of using DNNs for speech separation from 2-channel recordings.

In addition, we have integrated psychoacoustics and robust audio features into our DNN framework. A perceptually weighted objective function [4] has been proved effective in extracting sound sources corrupted by background noise from mono channel recordings, and a similar principle will be adopted here. The non-linearly wrapped spatial features [12] are robust to mismatched position combinations, which together with spectral features, will be fed into the DNN to output the separation mask for each TF point, which is then applied to the mixture for time-domain reconstruction.

The remainder of the paper is organized as follows. Section II introduces the overall DNN-based source separation system, followed by experimental results and analyses in Section III. Conclusions and insights for future work are given in Section IV.

II. DNNs for Source Separation

Treating each DNN as a black box as shown in Fig. 1, input features with groundtruth/labels are extracted from the 2-channel recordings and fed into the DNN. Each DNN aims to output the separation mask in the TF domain for target source estimation. In the training stage, we impose a loss/objective function on the DNN output and the groundtruth, to estimate the separated signal via the separation mask as perceptually close to the groundtruth as possible. Following, we will introduce the feature extraction process as well as the imposed objective function respectively.

A. Feature Extraction

Denoting $L(t,\omega)$ and $R(t,\omega)$ as the short time Fourier transform (STFT) of the 2-channel recordings at TF location (t,ω), both spectral and spatial features are extracted as DNN input. Log-power (LP) spectral features are first extracted as

$$Z^{\text{LP}}(t,\omega) = \max(\log(|L(t,\omega)|^2), \log(|R(t,\omega)|^2)), \quad (1)$$

which is then normalised, denoted as $\bar{Z}^{\text{LP}}(t,\omega)$. Non-linearly transformed spatial features [12] are exploited here,

$$\begin{cases} \chi_1(t,\omega) = \exp\left(-\left\|(\phi(t,\omega))|_{-\pi}^{\pi}\right\|^2\right), \\ \chi_2(t,\omega) = \exp\left(-\left\|(\phi(t,\omega) - 2\pi f_\omega \tau)|_{-\pi}^{\pi}\right\|^2\right), \end{cases} \quad (2)$$

where $\phi(t,\omega) = \angle\frac{L(t,\omega)}{R(t,\omega)}$ is the interchannel phase difference (IPD) between the two channels, and $|_{-\pi}^{\pi}$ wraps the phase residue in the range of $[-\pi,\pi]$. In the above equation, $2\pi f_\omega \tau$ is approximately the unwrapped IPD mean associated with the interference, where f_ω is the ω-th frequency, and τ is the delay of the interference arriving at the two channels, which can be estimated by the generalized cross-correlation phase transform method (GCC-PHAT) [13]. The above converted spatial features have yielded better performance as compared to the use of raw IPD features, as well as robustness to mismatch between the training and testing conditions [12].

We aim to output the ideal binary mask (IBM) directly, which can be obtained in the training stage by comparing the spectra of the target and the interference, denoted as $m(t,\omega)$.

B. Psychacoustically-Motivated Objective Function

In our previous work [4], we have proposed a perceptually-weighted objective function, which is a weighted modification of the mean square error (MSE) where perceptual importance of each TF point is considered from psychoacoustic point of view. This objective function is an empirical balance between boosting high energy components and suppressing any distortion that might cause perceptual changes. With the same principle, we propose an objective function to minimise the perceptual difference between the groundtruth mask m and the estimated separation mask $\hat{m}$ ignoring the TF index, using a modified weight defined on the normalised feature space:

$$\mathcal{L} = \frac{1}{N}\sum_{(t,\omega)}\left(\sigma(\bar{S}^{\text{LP}}) + (\mathbf{1} - \sigma(\bar{S}^{\text{LP}}))\sigma(\bar{\hat{S}}^{\text{LP}})\right)(\hat{m} - m)^2, \quad (3)$$

where N is the number of frames, $\sigma(\cdot)$ is the sigmoid function, $\bar{S}^{\text{LP}} = (S^{\text{LP}} - \mu)/\delta$ is the normalised LP feature of the groundtruth target, with μ and δ being the normalisation parameters. $\hat{S}^{\text{LP}} = 2\log(\hat{m})/\delta + \bar{Z}^{\text{LP}}$ is the normalised LP feature of the source estimate, extracted by applying the separation mask to the mixture spectrum directly. In the training stage, the loss will be minimised via the backpropagation, during which the DNN parameters that non-linearly map the input features to the final output will be updated.

III. Experiments

A. Data and Setup

Recordings from 22 speakers (11 male and 11 female) in the TSP Speech Database [14] were resampled at 8 kHz. For each speaker, 50 sequences were used for training and the left 10 for testing. The training mixtures were simulated with room impulse responses (RIRs) recorded by a pair of microphones distanced at 21 cm in a reverberant room [15] with RT60 of 325 ms. For training data generation, we randomly chose two sequences from two speakers, and convolved them with associated RIRs, with one target fixed at 0° azimuth, and the interference drawn from $[-90^\circ, -60^\circ, -30^\circ, 30^\circ, 60^\circ, 90^\circ]$. For each of the 3 gender combinations ("MM", "MF", "FF") and 6 position combinations, 1000 pairs of mixtures were generated for training. Similarly for testing data generation, 40 pairs of mixtures were simulated with the same mixing process, resulting 720 *matched* testing stereo mixtures in total. In addition, we also investigated *unmatched* conditions when the interference was drawn from $[-135^\circ, -110^\circ, 110^\circ, 135^\circ]$, with 480 unmatched testing mixtures in total. The matched and unmatched scenarios are illustrated in Fig. 2. To generate DNN input features, 256-point STFT with 0.75-overlapped Hanning window was employed.

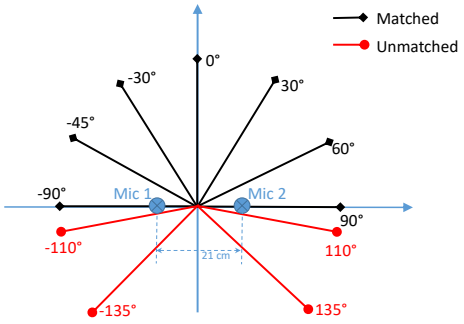

Fig. 2: The stereo mixtures were simulated as the superposition of the spatial images of a target speech located at 0° and a competing speech at other angles highlighted in square (except 0°) in the training stage. One group of *matched* testing data were generated using the same mixing process, and another *unmatched* group used new positions highlighted in dots.

B. DNN Implementation

We implemented five different DNN algorithms with different structures, detailed as follows.

The first one has the classic MLP structure, denoted as "MLP". Input features $(\bar{Z}^{\text{LP}},\chi_1,\chi_2) \in \mathcal{R}^{127\times 11\times 3}$ of consecutive 11 frames centred by the t-th frame were vectorised, which went through 4 fully-connected hidden layers with size of 1024 each. Moreover, each hidden layer is followed by batch normalisation (BN) [16] (to accelerate convergence) and leaky rectified linear units (ReLU) activation. The output layer is also a fully-connected layer with sigmoid activation, to output the separation mask at the t-th frame $\hat{\boldsymbol{m}}(t) = [\hat{m}(t,1), \hat{m}(t,2), \ldots, \hat{m}(t,127)]^T$.

The second one employs mainly the CNN structure, denoted as "CNN", which contains a convolutional encoder as well as a fully-connected decoder. The same concatenated features as for "MLP" were used without vectorisation. The encoder contains three convolutional layers without zero-padding, whose kernel size and number are respectively $(5,5)\times 64$, $(3,3)\times 128$, $(8,1)\times 256$, and maxpooling is followed with pool size of $(2,2)$, $(2,1)$ and $(4,1)$. The decoder shares the same structure as the last two layers in "MLP", for a fair comparison. BN and leaky ReLU were also applied to each hidden layer.

The third DNN is based on the capsule network structure [10], denoted as "CapsNet", which is also a modification of "CNN" by replacing the third convolutional layer in the encoder with one capsule layer. Firstly, the hidden output after the first two conventional layers were reshaped and squashed to capsules with a length of 8. Secondly, three iterations of routing process were employed to obtain 16 capsules each with a length of 64. Finally, all the capsules were forced with zero elements except the biggest capsule (with the largest Frobenius norm), which were vectorised and fed into the decoder for the separation mask $\hat{\boldsymbol{m}}(t)$ generation.

The above three methods are block-based, where temporal information is limited in the consecutive 11 frames, spanning in total 112 ms. However, take a speech signal as a sequence, longer temporal correlations exist. To address this problem, we also exploited RNN for source separation as follows.

For the forth method, input features were extracted from 100 consecutive frames ($\in \mathcal{R}^{127\times 100\times 3}$), lasting in total 824 ms[1]. At each frame, the features were vectorised with dimension of 381. We then applied two stacked bidirectional LSTM with size of 256 along temporal frames. The feed forward dropout and recurrent dropout were set to 0.5 and 0.2 respectively. Afterwards, the LSTM output at each frame was fed into a frame-independent fully-connected layer to generate $\boldsymbol{m}(t)$. This method is denoted as "RNN-T".

Similarly, speech spectrum also yields strong correlation across frequency such as harmonics, and we therefore also implemented RNN along the frequency direction denoted as "RNN-F". First, we segmented all the frequency bins to 32 bands with each containing 4 bins. Considering the three features ($\bar{Z}^{\text{LP}}$,χ_1,χ_2), each band has the feature dimension of 12. Then two stacked BiLSTM also with size of 256 were applied along each band. The same dropout as for "RNN-T" was applied except the feed forward dropout for the first layer BiLSTM, which was set to 0 due to the low dimensional input. A fully-connected layer followed BiLSTM with a size of 4 is used to generate the separation mask associated with the 4 bins of each band.

We summarised the above DNNs in Table I.

[1]RNN-based methods are designed to work with sequences, such more frames were used here as compared to aforementioned block-based DNN methods. For fair comparisons, we also tested feeding 100 frames to block-based methods. Yet, the limited DNN size could not model the complex input features and worse results were obtained. As a result, short temporal length spanning 11 frames was used for block-based methods.

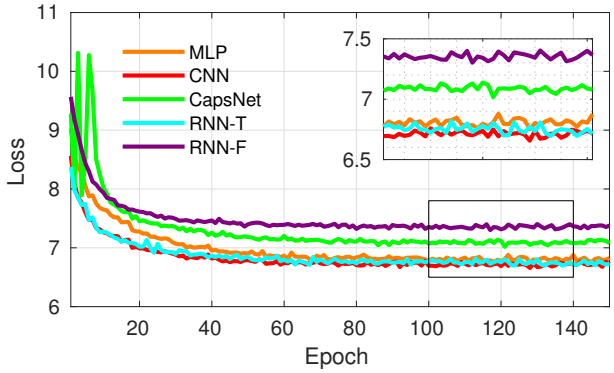

Fig. 3: Loss convergence for the five DNN-based source separation methods.

C. Results and Analysis

For each implemented method, one speaker-independent model was built on the training data, of which 20% were used for validation. Learning rate of 0.001 was initialised, with a decay rate of 0.95 after each epoch, and 150 epochs were enforced in total. Adaptive moment estimation (Adam) optimiser [17] was employed in the backpropagation.

We first show the loss convergence on the validation dataset in Fig. 3. All the DNN methods converged after around 50 epochs, with "MLP", "CNN" and "RNN-T" gaining similar results. We have zoomed in part of the converged results (highlighted) in the embedded rectangle, and it can be observed "RNN-F" and "CapsNet" have slightly higher losses.

We then applied the converged models on the matched testing data, and evaluated the separated targets using metrics of perceptual evaluation of speech quality (PESQ) [18], signal-to-distortion ratio (SDR) as well as short-time objective intelligibility (STOI) [19], as shown on the left panel of Fig. 4. The same evaluations were performed on the mixtures directly as a benchmark, denoted as "Input". In addition, the ideal binary masks (IBM) were extracted from the groundtruth signals, whose evaluations are denoted as "Ideal".

It can be observed that all the tested DNN methods have significantly improved the perceptual quality and speech intelligibility, and have reduced the distortion as compared to the original target speech embedded in competing speech. Average improvement over "Input" was obtained as 0.84, 6.3 dB and 0.18 for PESQ, SDR and STOI respectively. The total number of separated sounds that have shown worse performance than "Input" was less than 1%, and most of these degraded samples were estimated via "MLP". Moreover, the two RNN-based methods, "RNN-T" and "RNN-F" gained similar results as "IBM", especially for PESQ evaluations. They also outperformed the other three block-based DNN methods, showing average of 0.54, 2.92 dB, and 0.06 increase with the three metrics. Take "RNN-T" and "CNN" for example, paired sample t-tests were applied to their separated sources, and we got $p \approx 0.00$ for all the three evaluation matrices, which shows the statistical significance of the performance improvement. The three block-based DNN methods, "MLP",

TABLE I: Layer specification summary for the five DNN methods. The same processing between different methods is highlighted in boxes.

MLP	CNN	CapsNet	RNN-T	RNN-F
L1: FC1024+BNLR	L1: Conv $(5 \times 5) \times 64$ +BNLR+MP (2,2)	L1: Conv $(5 \times 5) \times 64$ +BNLR+MP (2,2)	L1: BiLSTM256 + D (0.5,0.2)	L1: BiLSTM256 + D (0,0.2)
L2: FC1024+BNLR	L2: Conv $(3 \times 3) \times 128$ +BNLR+MP (2,1)	L2: Conv $(3 \times 3) \times 128$ +BNLR+MP (2,1)	L2: BiLSTM256 + D (0.5,0.2)	L2: BiLSTM256 + D (0.5,0.2)
L3: FC1024+BNLR	L3: Conv $(8 \times 1) \times 256$ +BNLR+MP (4,1)	L3: Capsule (16×64)	L3:FC127+sigmoid	L3:FC4+sigmoid
L4: FC1024+BNLR	L4: FC1024+BNLR	L4: FC1024+BNLR		
L5: FC127+sigmoid	L5: FC127+sigmoid	L5: FC127+sigmoid		

L=layer, FC=fully connected, BNLR=BN+Leaky ReLU, Conv=convolutional, MP = max pooling, D = dropout

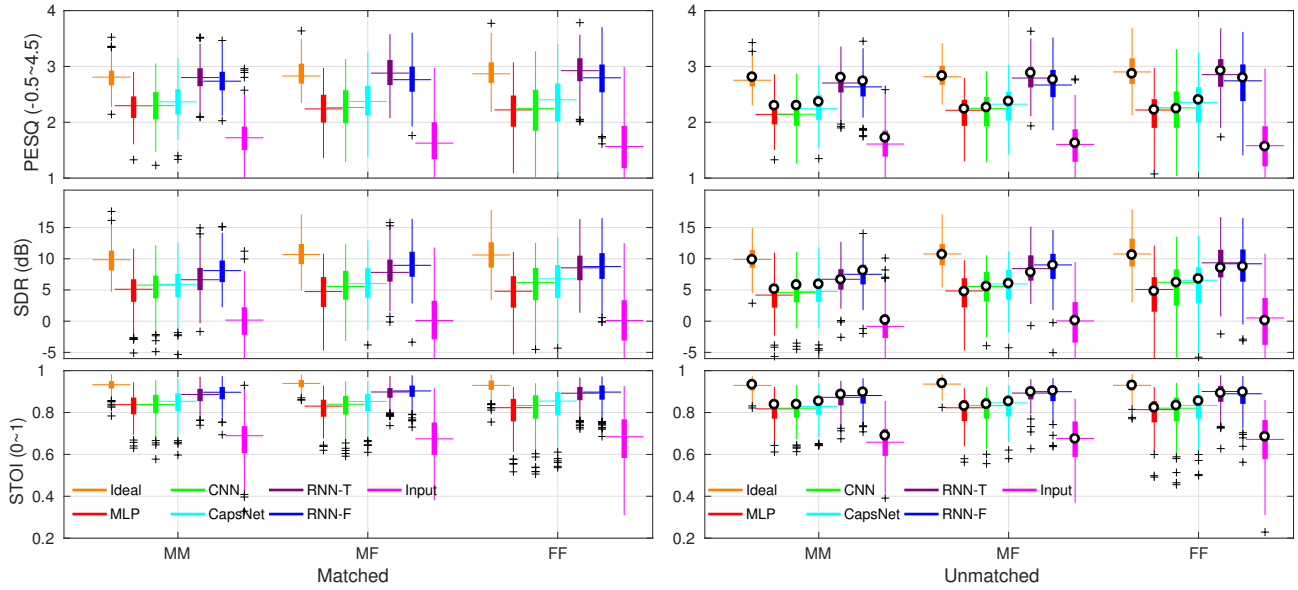

Fig. 4: Quantitative evaluations on *matched* testing data (left) and *unmatched* testing data (right) in terms of PESQ (top), SDR (middle) and STOI (bottom). The box has edges with 25th and 75th percentiles, as well as the median value (the central cross mark), and the whiskers extend to the most extreme data points. Outliers are shown in black pluses. The associated median values from the matched situations are highlighted in black circles in the unmatched conditions for comparisons.

"CNN" and "CapsNet" yielded similar performance. This limitation might be caused by its less effective modelling of long temporal correlation (as compared to "RNN-T") and whole frequency band relationship (as compared to "RNN-F"). To investigate this, we compared the spectrum of the recovered source estimate with that of the groundtruth in Fig. 5, where the spectra were either extracted from one frequency bin and spanning different time frames (top), or extracted in one frame spanning all the frequency bins (bottom). It can be observed that "RNN-T" exhibited high correlation with the groundtruth over time while "RNN-T" yielded strong correlation over frequency. Interestingly, "RNN-T" showed better results than "RNN-F" in PESQ while worse results in SDR and similar results in STOI. This is because the three involved evaluation metrics measure different aspect of sounds with very different mechanisms. For instance, with the same distortion levels (SDR), the competing speech with overlapped spectra to the target is more likely to affect human perception than isolated spectra without overlap.

In addition, "CapsNet", which has shown advantages over "CNN" in classification [10], had only very slight improvement over "CNN" in terms of PESQ and STOI. This might be because in our regression model, there does not exist groundtruth category labels that can be applied to the capsule layer directly, to indicate the ownership of each capsule as in the classification model, thus no additional learning rule was enforced on the capsule layer output. Yet, this improvement, on the other hand, shows the capability of "CapsNet" in representing the underlying structure of speech signals.

Moreover, no obvious performance difference between different gender combinations was observed, possibly due to the following two reasons. First, the essential spectral difference (e.g. timbrel, pitch) between male and female groups could be relaxed by the strong individual difference in our 22 training

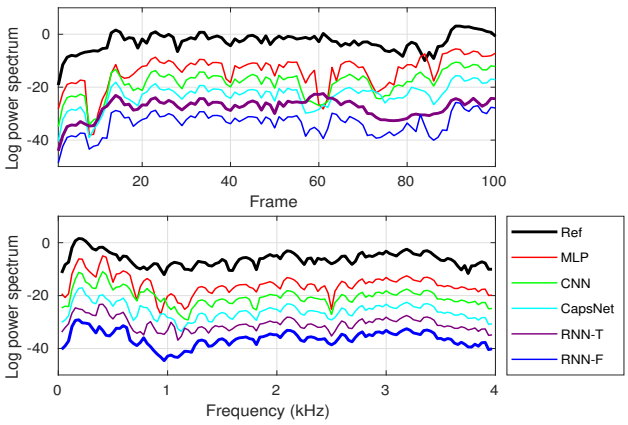

Fig. 5: Spectra illustration in one frequency bin (top) and in one frame (bottom). "RNN-T" shown more consistent similarity with the groundtruth (Ref) over temporal frames, while "RNN-F" showed strong correlation over frequency bins, as highlighted in thick lines. DNN-separated spectra were shifted downward by steps of 5 for illustration purposes.

speakers. Second, contributions from the spatial features could reduce the importance of the spectral features.

To test the robustness of these DNN-based methods, we also applied these methods on the mismatched testing data, as shown in Fig. 4 (right). The highlighted median values of the associated matched conditions almost overlapped with that of the unmatched scenarios. This consistent performance proves that our DNN implementation exploiting the proposed features is robust to position mismatch.

IV. Conclusions

To systematically evaluate DNN-based source separation methods from 2-channel recordings in room environments, we implemented several methods with different DNN topologies, including MLP, CNN, CapsNet and RNN. Objective evaluation metrics of PESQ, SDR and STOI were performed on the extracted target sounds for speech quality and speech intelligibility measurements. We found that exploiting the psychacoustically-motivated objective function and position-robust spatial features, the RNN-based methods with BiLSTM structure showed consistent and better performance than other DNN structures. In the future, we are interested in implementing more state-of-the-art DNN architectures for comparisons and also consider other audio features to augment DNN input. Moreover, we will generalise these methods to more than two speaker scenarios and also take background environment noise into account.

Acknowledgement

The authors of the paper would like to acknowledge the support of the EPSRC Programme Grant S3A: Future Spatial Audio for an Immersive Listener Experience at Home (EP/L000539/1) and the BBC as part of the BBC Audio Research Partnership.

References

[1] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *CoRR*, vol. abs/1708.07524, 2017.

[2] Y. Jiang, D. Wang, R. Liu, and Z. Feng, "Binaural classification for reverberant speech segregation using deep neural networks," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 22, no. 12, pp. 2112–2121, December 2014.

[3] X. Zhang and D. Wang, "Deep learning based binaural speech separation in reverberant environments," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 1075–1084, May 2017.

[4] Q. Liu, W. Wang, P. JB Jackson, and Y. Tang, "A perceptually-weighted deep neural network for monaural speech enhancement in various background noise conditions," in *European Signal Processing Conference*, August 2017.

[5] P. S. 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*, 2014, pp. 1562–1566.

[6] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *International Conference on Latent Variable Analysis and Signal Separation*, 2015, pp. 91–99.

[7] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, March 2016, pp. 31–35.

[8] P. Chandna, M. Miron, J. Janer, and E. Gómez, "Monoaural audio source separation using deep convolutional neural networks," in *IEEE International Conference on Latent Variable Analysis and Signal Separation*, 02 2017.

[9] E. M. Grais and M. D. Plumbley, "Single channel audio source separation using convolutional denoising autoencoders," in *2017 IEEE Global Conference on Signal and Information Processing*, November 2017, pp. 1265–1269.

[10] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Annual Conference on Neural Information Processing Systems*, 2017.

[11] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Networks*, vol. 18, no. 5, pp. 602 – 610, 2005.

[12] Q. Liu, Y. Xu, P. Coleman, P. JB Jackson, and W. Wang, "Iterative deep neural networks for speaker-independent binaural blind speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, April 2018.

[13] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, August 1976.

[14] P. Kabal, "TSP Speech Database," Tech. Rep., McGill University, 2002.

[15] L. Remaggi, P. J. B. Jackson, P. Coleman, and W. Wang, "Acoustic reflector localization: Novel image source reversion and direct localization methods," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 2, pp. 296–309, Feburary 2017.

[16] 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.

[17] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.

[18] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2001, vol. 2, pp. 749–752.

[19] 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*, March 2010, pp. 4214–4217.