ITERATIVE DEEP NEURAL NETWORKS FOR SPEAKER-INDEPENDENT BINAURAL BLIND SPEECH SEPARATION

*Qingju Liu, Yong Xu, Philip JB Jackson, Wenwu Wang**

Centre for Vision, Speech and Signal Processing
University of Surrey, UK

Philip Coleman

Institute of Sound Recording
University of Surrey, UK

ABSTRACT

In this paper, we propose an iterative deep neural network (DNN)-based binaural source separation scheme, for recovering two concurrent speech signals in a room environment. Besides the commonly-used spectral features, the DNN also takes non-linearly wrapped binaural spatial features as input, which are refined iteratively using parameters estimated from the DNN output via a feedback loop. Different DNN structures have been tested, including a classic multilayer perception regression architecture as well as a new hybrid network with both convolutional and densely-connected layers. Objective evaluations in terms of PESQ and STOI showed consistent improvement over baseline methods using traditional binaural features, especially when the hybrid DNN architecture was employed. In addition, our proposed scheme is robust to mismatches between the training and testing data.

Index Terms— Deep neural network, binaural blind speech separation, spectral and spatial, iterative DNN

1. INTRODUCTION

Deep neural networks (DNN) [1] have recently been exploited in the field of blind source separation [2], e.g., to extract target speech corrupted by background noise [3–7] or to jointly estimate multiple sound sources [8–11]. Monaural source separation methods often employ spectral features such as time-frequency (TF) domain representative features and filterbank features [3–5, 7–9, 11], while multiple channel source separation methods can exploit additional spatial information, e.g., to directly feed the DNN [6] or to refine the Wiener filtering for recovering sources [10].

In this paper, we focus on the speaker-independent binaural source separation problem where two talkers are speaking concurrently at unknown positions. Existing DNN-based multi-channel source separation methods often aim to recover one target (after delay-and-sum beamforming) at the azimuth of 0 degree, in the presence of uncorrelated ambient noise [6, 10]. In addition, the DNN models trained in [10] use only spectral cues as DNN input. The fixed DNN training in [6] does not consider the large number of spatial combinations of two speakers, where distributions of the commonly-used spatial cues differ under these combinations. To address these limitations, we propose to transform the spatial features to a uniquely-distributed space that is robust to different spatial combinations. This transformation process requires position-associated delay information, which can be obtained from the binaural mixture and refined from the DNN output with an iterative process. The refinement process uses similar optimisation principles of stochastic depth [12]. The pre-trained neural network outputs spectral features for the reconstruction of estimated sources in the time-domain.

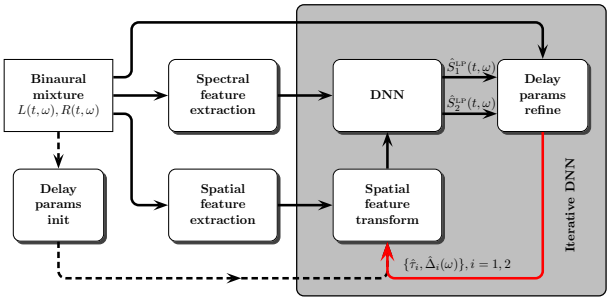

Fig. 1: Diagram of our proposed iterative DNN. Parameters $\hat{\tau}_i$ and $\hat{\Delta}_i(\omega)$ used to calculate the converted spatial features, are initialised from the binaural mixture (dashed line) and iteratively refined with the DNN output (red line).

The remainder of the paper is organized as follows. Section 2 introduces the overall proposed scheme, followed by experimental results and analyses in Section 3. Conclusions and insights for future work are given in Section 4.

2. THE PROPOSED METHOD

Our proposed binaural source separation scheme is shown in Fig. 1, where the DNN takes both spectral and spatial features as input and then outputs the spectra of two source estimates.

Suppose $L(t,\omega)$ and $R(t,\omega)$ are the short time Fourier transform (STFT) of the two channels in the binaural recordings indexed by the TF location (t,ω). Log-power (LP)

*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.

features $L^{\text{LP}}(t,\omega) = \log(|L(t,\omega)|^2)$ can be extracted, which have proven effective in monaural speech enhancement [3, 7]. The maximal value between the two channels is extracted as the spectral feature, to reduce inter-channel redundancy:

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

Similarly, the groundtruth spectral feature $S_i^{\text{LP}}(t,\omega), i = 1, 2$ can be obtained from spatial images (contributions in both channels) of each sound source.

Besides these spectral features, spatial features are also considered in our system. Commonly used binaural spatial features include interaural phase difference (IPD) and interaural level difference (ILD), which can be either statistically characterised [13] via an expectation maximisation (EM) process, or used to boost the DNN spectral input directly [6]. In this pilot study, we exploit only the IPD cue $\phi(t,\omega) = \angle\frac{L(t,\omega)}{R(t,\omega)}$. Considering the potentially large number of scenarios when two targets are combined in the binaural mixtures, the IPD distributions become complex. We illustrate in the top of Fig. 2 the distributions of $\phi(t,\omega)$ calculated from mixtures containing two sound sources under two combination scenarios: $-30° + 45°$, $-45° + 30°$. It can be observed that the IPD distributions vary with frequency and combination conditions. To represent the various IPD distributions under all possible mixing conditions, a large neural network is required, which is prone to overfitting and needs a big training set. To address this problem, we transform the IPD feature with the prior information of the unwrapped IPD mean[1] for each sound source denoted as $\beta_i(\omega), i = 1, 2$:

$$\chi_i(t,\omega) = \exp\left(-\left\|(\phi(t,\omega) - \beta_i(\omega))|_{-\pi}^{\pi}\right\|^2\right), \quad (2)$$

which is a nonlinear wrapping of the squared phase residual into the range of $(0, 1)$. The 2D vector $[\chi_1(t,\omega), \chi_2(t,\omega)]$ shows joint distributions with a sparse pattern, robust to frequency and combination conditions. In addition, data associated with different targets can be clustered to different groups, as illustrated in the bottom of Fig. 2. As a result, the converted features may facilitate DNN training.

In the supervised offline DNN training stage, the unwrapped IPD mean $\beta_i(\omega)$ can be extracted directly from spatial images of each target sound given the exact mixing process, which is however unknown in the online DNN separation stage and thus needs to be estimated. Ideally, if the right channel signal is a delayed version of the left channel by delay τ, we could obtain $\beta_i(\omega) = 2\pi f_\omega \tau$, where $f_\omega = \frac{\omega F_s}{N_{\text{FFT}}}$ is the ω-th frequency with F_s being the sampling rate and N_{FFT} the FFT size. Due to reflections within the listening environment as well as the head shadowing effect, the delay for a sound from certain direction is in practice frequency-dependent, and $\beta_i(\omega)$ can be compensated by some angle

[1]The IPD mean can be constrained in the range of $-\pi$ to π, which can also be unwrapped to a wider range to maintain frequency consistency. This unwrapping process does not affect Equation (2).

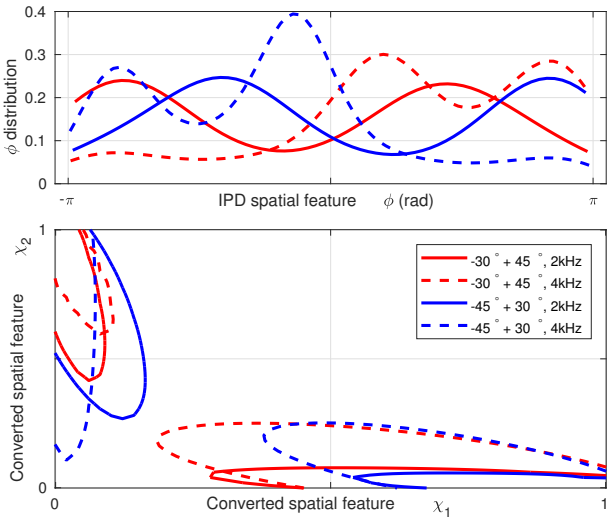

Fig. 2: IPD distributions (top) and joint distributions of the converted features (bottom) from mixtures consisting of two sound sources, at different frequencies and position combinations. The binaural mixtures were simulated by convolving speech signals with associated binaural room impulses [14] and adding them together. The IPD distribution (top) can be modelled with a Gaussian mixture model containing two kernels associated with the two sound sources, whose mean is frequency and position dependent [13]. The converted features (bottom) yield very sparse distributions and 80% of time-frequency points fall in the area enclosed by the contours.

shift $\Delta_i(\omega)$ with small values:

$$\beta_i(\omega) = 2\pi f_\omega \tau_i + \Delta_i(\omega). \quad (3)$$

When applying the trained DNN for online separation, we need to estimate the dominant delay $\hat{\tau}_i$ and $\hat{\Delta}_i(\omega)$ for calculating $\beta_i(\omega)$ associated with each sound source. The generalized cross-correlation phase transform method (GCC-PHAT) [15] can be employed to estimate $\hat{\tau}_i$, while $\hat{\Delta}_i(\omega)$ can be initialised with zeros (dashed line in Fig. 1). Afterwards, these parameters can be iteratively refined using the DNN-separated signals $\hat{S}_i(t,\omega)$ (red line in Fig. 1) as follows:

$$\begin{cases} \hat{\tau}_i = \underset{\tau}{\text{argmin}} \sum\limits_{t,\omega|(t,\omega)\in\mathcal{H}_i} \left\|(\phi(t,\omega) - 2\pi f_\omega \tau)|_{-\pi}^{\pi}\right\|^2, \\ \hat{\Delta}_i(\omega) = \dfrac{\sum_{t|(t,\omega)\in\mathcal{H}_i} (\phi(t,\omega) - 2\pi f_\omega \hat{\tau}_i)\,|_{-\pi}^{\pi}}{\sum_{t|(t,\omega)\in\mathcal{H}_i} 1}. \end{cases} \quad (4)$$

In the above equations, $\mathcal{H}_i$ is defined as

$$\mathcal{H}_i = \{(t,\omega)\}, \text{s.t. } Z^{\text{LP}}(t,\omega) > \epsilon, \hat{S}_i(t,\omega) > \hat{S}_{\bar{i}}(t,\omega), \quad (5)$$

where ϵ is the median value of $Z^{\text{LP}}(t,\omega)$, and $\bar{i}$ denotes the source index that does not equal to i. We enforce the condition $Z^{\text{LP}}(t,\omega) > \epsilon$ to avoid using components with small values that are prone to outliers; we enforce the condition $\hat{S}_i(t,\omega) > \hat{S}_{\bar{i}}(t,\omega)$ to exploit information dominated by the associated source. The minimisation can be simplified by iteratively updating $\hat{\tau}_i$ with a greedy method using candidate values close to the previous optimal point.

3. EXPERIMENTS

3.1. Data and setup

Recordings from four male speakers M_i and four female speakers $F_i, i = 1, ..., 4$, in the TSP Speech Database [16] were used to test our proposed scheme. The sampling rate is 16 kHz. For each speaker, 50 sequences were used for training and 10 for testing. We generated binaural mixtures using binaural room impulse responses (BRIRs) recorded in a reverberant room with RT60 of 640 ms [14], by convolving two randomly chosen signals with corresponding BRIRs and then adding them together. In the training stage, we considered position combinations drawn from $[-60°, -30°, 0°, 30°, 60°]$ when generating the binaural mixtures. To test the robustness of our proposed method to unmatched combination scenarios, two groups of testing data were simulated, with one group using the same position combinations as the training data, and the other group drawn from $[-45°, 5°, 45°]$. Different gender combinations of "MM", "MF", and "FF" were also investigated. Under each location combination, e.g. $-60° + 30°$, we randomly chose two speakers, e.g. M_1 and F_2, and generated 50 (10) binaural mixtures from randomised training (testing) data sequences. In total, 18000 binaural mixtures lasting about 12 hours were generated for training, 3600 for matched testing, and 1080 for unmatched testing. 20% of the training binaural mixtures were used for validation. For each method we implemented, a single speaker-independent network was trained using the whole training data.

In our DNN implementations, the spectral and spatial features (IPD) were extracted using a 512-point STFT with half-overlapped Hamming windows. At each time frame t, features from the neighbouring 11 frames were used to exploit temporal correlation. The mean square error (MSE) was used as the cost function. In the training process, the mini-batch size was set to 128 in the backpropagation, and the training data were randomized after each epoch.

We have tested two DNN architectures. The first one is the classic multilayer perception (MLP) DNN [3, 6, 7]. In our implementation, the MLP DNN has three hidden layers, and each hidden layer contains 3000 leaky rectified linear units (ReLU). We employed batch normalisation (BN) layers [17] after each densely-connected hidden layer to accelerate convergence. The second DNN has a hybrid structure as shown in Fig. 3, which contains both convolutional layers as well as densely-connected layers. Two parallel processing streams were employed. One stream (top, local stream) aims to learn local information, where convolutional layers with strides of $(1, 1)$ and zero-padding were used without any pooling process. A similar structure has been shown useful in maintaining details [18] in image super-resolution. The other stream (bottom, global stream) aims to learn global relationships, where convolutional layers with max pooling and strides were used with zero-padding, followed by a densely-connected layer. To combine the two streams together, the

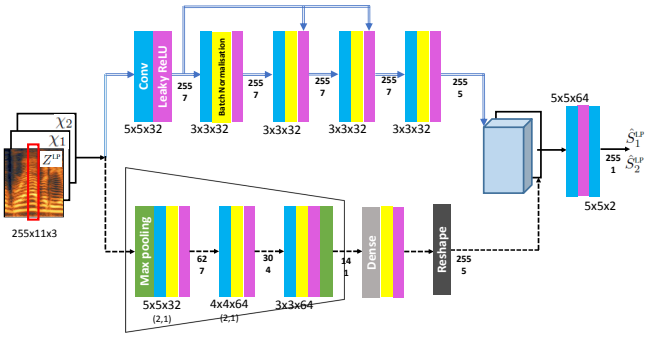

Fig. 3: The hybrid DNN architecture. The first and second dimensions in each layer output are listed. For each convolutional layer, the kernel size as well as the number, and the stride not equals to $(1, 1)$ are listed. Zero-padding (not shown) was used. Take the first convolutional layer in the local stream for example, kernel size of $(5, 5)$ was applied to input size of $(255, 11)$ and output size of $(255, 7)$, which means the amount of zero-padding of 2 was enforced in the borders of the first dimension. The two max-pooling layers have the size of $(2, 1)$ and $(2, 2)$ respectively.

global stream output was reshaped and concatenated with the local stream. More convolutional layers were tagged in the end to generate the final result. Leaky ReLUs were employed at each hidden layer. Residual learning [19] and BN were used to ease the training.

Using the converted features as spatial features, we denote the proposed iterative DNN scheme with the above two DNN structures as "Convert-MLP" and "Convert-Hybrid" respectively. Note that, the same input feature for "Convert-Hybrid" needs to be vectorised for "Convert-MLP". Three iterations were employed in the proposed DNN scheme to refine the converted spatial features. We implemented two baseline methods, by directly feeding the raw features of Z^{LP} and ϕ to the aforementioned DNNs, denoted as "Raw-MLP" and "Raw-Hybrid". In addition, we also used the method proposed in [6], denoted as "Method [6]", as a baseline, which was slightly modified as follows. Instead of outputting the ideal ratio mask associated with one target, the DNN output is the LP spectra of both estimated sources, to be consistent with all the other DNN methods. As with the other DNN methods, features spanning 11 neighbouring frames were concatenated as the DNN input.

3.2. Results and analysis

The converged DNN models were saved after 200 epochs. DNN methods using the converted spatial features, i.e. "Convert-Hybrid" and "Convert-MLP", converged to similar results as "Method [6]" with a loss around 0.40. Methods using raw features, i.e. "Raw-Hybrid" and "Raw-MLP" exhibited higher loss of approximately 0.65 and 0.60 respectively.

We then evaluated and compared the speech quality and intelligibility of signals separated by the five approaches in

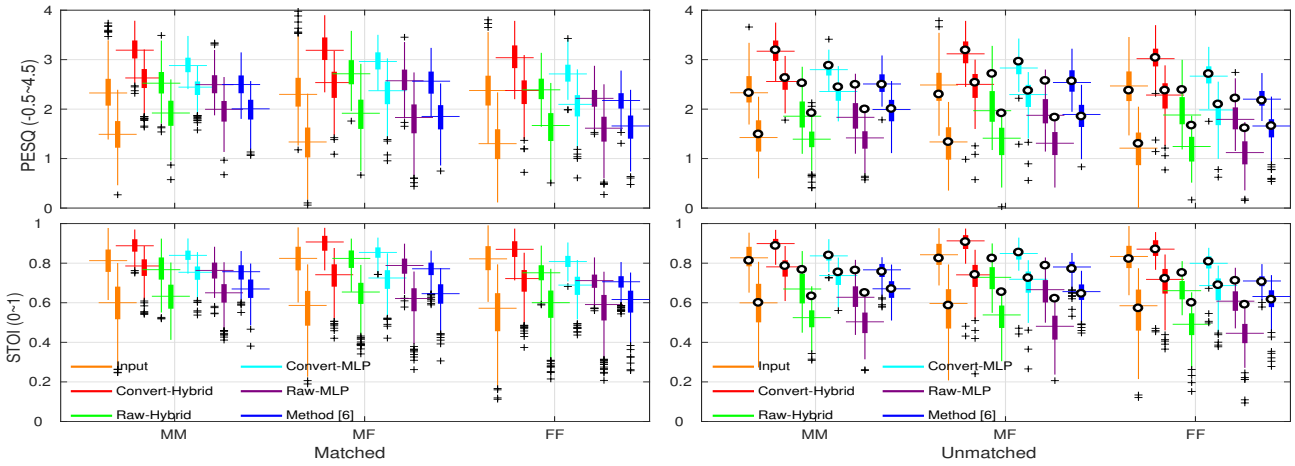

Fig. 4: Quantitative evaluations on "matched" testing data (left) and "unmatched" testing data (right) in terms of PESQ (top) and STOI (bottom). We ranked the two DNN outputs in each scenario based on their evaluated score, and plotted evaluation results for the first (best) and second source estimates. For each scenario, 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.

terms of perceptual evaluation of speech quality (PESQ) [20], and short-time objective intelligibility (STOI) [21]. We first performed evaluations on "matched" testing data, as illustrated in the box-plots in Fig. 4 (left), where the evaluation metrics were applied to each of the two source estimates. The same evaluations were performed on the binaural mixtures without processing as a benchmark, denoted as "Input", i.e. by directly comparing the mixture with each of the two groundtruth/reference signals. It can be observed that using the proposed feature, either "Convert-Hybrid" or "Convert-MLP" gained the best and consistent performance over different gender combinations, especially for "Convert-Hybrid", where the worst separated source still exhibited similar or even better scores than the other three baseline methods. As compared to "Input", methods using the converted feature showed consistent improvement in both PESQ and STOI, except very slight decrease in STOI for the first separated signal with "Convert-MLP" under the FF gender combination. The three baseline methods also outperformed "Input" in MM and MF scenarios, especially when evaluating the second separated source. Yet, this advantage did not persist for the first separated sound in STOI evaluations. It can be noted that "Method [6]" showed worse performance than our proposed methods in terms of PESQ and STOI, even though the MSE loss in the training process showed similar convergence performance. Informal listening revealed that "Method [6]" suffered from severe distortions of the target and interfering speech, compared to the methods based on the converted IPD feature, with Convert-Hybrid sounding the most natural among the methods tested. These results imply that the MSE loss might not be the best choice of cost function to drive speech quality.

To test the robustness of our proposed scheme to the mismatch between the training and testing data, we also ran the same evaluation process on "unmatched" testing data, as illustrated in Fig. 4 (right). To compare with the matched situations, we highlighted median values of the associated matched conditions in black circles. It can be seen that using raw features, i.e. "Raw-Hybrid" and "Raw-MLP", the performance dramatically decreased, and worse results than "Input" were obtained. Consistent results were observed with converted features as compared to matched situations, proving the robustness of our proposed scheme. Interestingly for "Method [6]", very similar results were gained as compared to the matched scenarios, which shows that their employed features also exhibit robustness to mismatched conditions. Overall, the proposed methods using the converted features, i.e. "Convert-Hybrid" and "Convert-MLP", still outperform all the other baseline methods as well as direct evaluations on the binaural mixtures in both PESQ and STOI.

4. CONCLUSIONS

An iterative DNN-based binaural source separation scheme using converted features refined iteratively from the DNN output has been proposed, to solve the problem of retrieving two concurrent speech signals in a room environment. The proposed scheme showed good performance in terms of perceptual speech quality and speech intelligibility, especially when the proposed hybrid DNN was employed. In addition, our proposed scheme yielded robustness to position combination mismatch between the training and testing data. In the future, we plan to use more advanced cost functions, that better reflect the speech quality, in the DNN training. Also, we need to generalise the situations for concurrently recovering more than two sound sources. More comparisons should be performed to other DNN structures, such as recurrent neural networks. More evaluation metrics should be considered such as signal to distortion ratio (SDR).

5. REFERENCES

[1] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Parallel distributed processing, vol. 1: Explorations in the microstructure of cognition: Foundations," chapter Learning Internal Representations by Error Propagation, pp. 318–362. MIT Press, Cambridge, MA, USA, 1986.

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

[3] Y. Xu, J. Du, L. R. Dai, and C. H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, January 2014.

[4] 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.

[5] Y. Wang and D. Wang, "Towards scaling up classification-based speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1381–1390, July 2013.

[6] 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.

[7] 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.

[8] 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.

[9] P. S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2136–2147, December 2015.

[10] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1652–1664, September 2016.

[11] 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.

[12] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Q. Weinberger, "Deep networks with stochastic depth," *CoRR*, vol. abs/1603.09382, 2016.

[13] A. Alinaghi, P. JB 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.

[14] C. Hummersone, "Binaural room impulse response measurements," Online, August 2011, `http://iosr.surrey.ac.uk/software/#BRIRs`.

[15] 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, Aug 1976.

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

[17] 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.

[18] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. P. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," *CoRR*, vol. abs/1609.04802, 2016.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition*, June 2016, pp. 770–778.

[20] 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.

[21] 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.