INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Competing speaker count estimation on the fusion of the spectral and spatial embedding space

Chao Peng, Xihong Wu, Tianshu Qu

Key Laboratory on Machine Perception (Ministry of Education), Speech and Hearing Research Center, Peking University, Beijing, China

qutianshu@pku.edu.cn

Abstract

This paper presents a method for estimating the competing speaker count with deep spectral and spatial embedding fusion. The basic idea is that mixed speech can be projected into an embedding space using neural networks where embedding vectors are orthogonal for different speakers while parallel for the same speaker. Therefore, speaker count estimation can be performed by computing the rank of the mean covariance matrix of the embedding vectors. It is also a feature combination method in speaker embedding space instead of simply combining features at the input layer of neural networks. Experimental results show that embedding-based method is better than classification-based method where the network directly predicts the count of speakers and spatial features help to speaker count estimation. In addition, the features combined in the embedding space can achieve more accurate speaker counting than features combined at the input layer of nueral networks when tested on anechoic and reverberant datasets.

Index Terms: speaker count estimation, embedding space, competing speaker

1. Introduction

Picking up the target speaker's voice has always been a difficult problem in an environment where multiple speakers are talking simultaneously while other perturbation sources are present. However, Most speech separation methods have one major problem that the competing speaker count has to be known. Therefore, the number of sound sources present in an acoustic scene is crucial information for not only speech separation but also sound localization [1], sound surveillance [2] and multi-talker speech recognition [3].

However, the competing speaker count in the real environment cannot be directly obtained. Fortunately, there are several strategies to solve this problem in more recent works. The first one is counting by speaker diarization [4, 5]. It counts by detecting who starts and ends speaking in a period of time. It appears to be a very complex problem and when sources are simultaneously active, the existing segmentation strategies will fail in real cocktail party environments. The second one is counting by direction of arrival (DOA) [6, 7, 8]. It is determined by manually determining thresholds or detecting the number of peaks. The main weakness of this method is that it cannot deal with the situation where multiple speakers are close in space. The third one is counting by clustering in the time-frequency (TF) domain [6, 9, 10]. However, the method is generally limited to the anechoic setting and often require additional maximum number of competing speakers. The fourth one is the multi-channel approach based on the eigenvalue analysis of the estimated spatial covariance matrix (SCM) [11, 12] but cannot be used in an underdetermined setup. The last strategy is counting using deep neural networks [13, 14, 15, 16]. Built upon powerful machine learning technology, it directly maps input representations into speaker count. Although this strategy is feasible, it lacks mathematical interpretability.

Motivated by the Deep Clustering (DC) [17, 18] in speaker counting and separation [19], this work proposes a multi-channel method with deep spectral and spatial embedding space fusion (EBFMC) for competing speaker count estimation. Spectral features represent speakers' content information while spatial features represent speakers' spatial orientations. Since the embedding space trained by spectral features and spatial features may be complementary, two embedding spaces are firstly pre-trained and then mapped into a new embedding space. This method is capable to achieve a good estimation of the competing speaker count.

The rest of the paper is organized as follows. The next section presents the proposed method EBFMC. The section three reports the generation of the experimental data, experimental environment and experimental results. Finally, the last section summarizes the paper.

2. The proposed counting system

In this section, we present the proposed multi-channel deep spectral and spatial embedding fusion based competing speaker count estimation system EBFMC, which is shown in Figure 1. Firstly, speech signals of multiple speakers are collected from a microphone array. Spectral and spatial features are then extracted from those multi-channel signals. Next the spectral features and the spatial features are mapped into their respective embedding spaces by using multi-layer Bi-directional Long Short-Term Memory networks (BLSTMs). Since different distribution of their respective embeddings, two embedding spaces are finally mapped into one embedding space through a fully connected layer. Finally, the competing speaker count is obtained by estimating the rank of the covariance matrix of the embedding vectors.

2.1. Input features

Our system uses the spectral and spatial features as the input features. The difference between single-channel and multi-channel method is the input features. As for spectral features, only the mixture log magnitude spectrum $log(|Y_p|)$ of the reference microphone p is used.

As for spatial features such as Interaural Phase Difference (IPD), they have been typically used in a number of works in order to train network [20, 21] and can be a combination of cosIPD and sinIPD for the reference microphones p and another microphone q splicing along the frequency axis.

$$cosIPD = cos(\angle Y_p - \angle Y_q) \tag{1}$$

http://dx.doi.org/10.21437/Interspeech.2020-1781

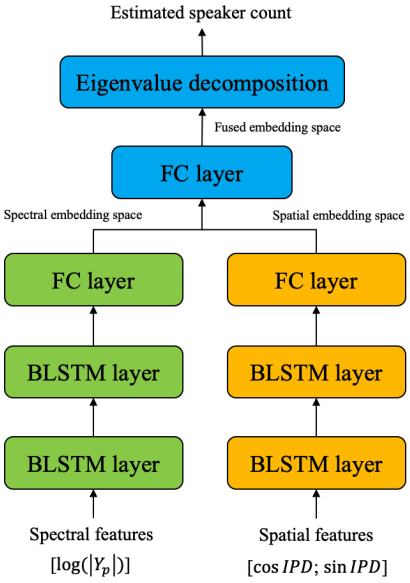

Figure 1: *Schematic diagram of the proposed EBFMC.*

$$sinIPD = sin(\angle Y_p - \angle Y_q) \tag{2}$$

Besides, when the number of microphones $M \geqslant 2$, cosIPD and sinIPD are both $F(M-1) \times T$ matrix where F and T are the number of frequency points and frames respectively.

2.2. Joint embedding space learning and fusion

Joint learning is a kind of multi-task learning, which has been successfully applied in the fields of natural language processing [22] and computer vision [23]. The joint learning-based method EBFMC firstly learns the corresponding embedding space from spectral features and spatial features obtained from multi-channel signals. Since the distribution of those two feature embedding space is complementary, they're mapped into a new embedding space by a fully connected layer.

Bidirectional long-short term memory (BLSTM) networks are trained to map features into an embedding space. In the embedding space, embedding vectors are parallel in the same direction for time-frequency bins dominated by the same speaker, or orthogonal for those dominated by different speakers. This nature of the embedding vectors allows us to perform source counting by performing the eigenvalue decomposition of the covariance matrix of the embedding vectors. The input of BLSTM is features of the speech signals X and the output is the D-dimensional deep embedding features V.

$$V = f_\theta(X) \in R^{N \times D}, \tag{3}$$

where N is the number of time-frequency bins after removing silence and $f_\theta(*)$ is a mapping function based on the BLSTM network.

The representation V mapped into the high-dimensional space should still describe the similarity of the time-frequency bins, that is, the matrix $\hat{A} = VV^T$ should be equal to $A = YY^T$ where $Y \in R^{N \times C}$ is the one-hot representation of time-frequency bins and C is the number of speakers. So the loss function can be obtained as:

$$C_Y(V) = ||\hat{A} - A||_F^2 = ||VV^T - YY^T||_F^2, \tag{4}$$

where V represent the speaker embedding vectors as shown in Figure 1 and $||*||_F^2$ is the squared Frobenius norm.

To avoid explicitly constructing the $N \times N$ affinity matrix, it is usually efficiently implemented as follows:

$$C_Y(V) = ||V^TV||_F^2 - 2||V^TY||_F^2 + ||Y^TY||_F^2 \tag{5}$$

The element at (n, n') is 0 if time-frequency bin n and n' are dominated by the different speaker, otherwise the element at (n, n') is 1 in the matrix $\hat{A}$ or A. In other words, the embedding vector for the n-th time-frequency bin $v_n = (v_{n,1}, ..., v_{n,D})^T$ would be orthogonal to v'_n if time-frequency bins n and n' are dominated by the different speaker, otherwise $v_n^T v'_n$ would be one.

2.3. Speaker count estimation

The competing speaker count can be obtained by eigenvalue decomposition of the mean covariance matrix of embedding vectors because the embedding vectors of C speakers are orthogonal to each other ideally. Suppose the mean covariance matrix of the embedding vector is R_e.

$$R_e = \frac{1}{N}\sum_{n=1}^{N} v_n v_n^T \tag{6}$$

Eigen-decomposition of the mean covariance matrix R_e is given as follows:

$$R_e = U\Lambda U^H, \tag{7}$$

where U denotes the matrix of eigenvectors and Λ is the matrix of eigenvalues, denoted as:

$$\Lambda = diag(\lambda_1, \lambda_2, ..., \lambda_D) \tag{8}$$

The rank of the covariance matrix is equal to the competing speaker count theoretically, therefore we assume that the number of eigenvalues λ larger than a threshold th represents the number of speakers.

$$\hat{C} = n(\{\lambda_i | \lambda_i \geq th, i = 1, 2, ..., D\}), \tag{9}$$

where $n(*)$ is the operation to compute the number of elements and $\hat{C}$ is the estimated speaker count.

3. Experiments

3.1. Dataset

The experimental data of simulation was generated from the Wall Street Journal (WSJ0) corpus. We created a mixture dataset of 1-5 speakers. The dataset was divided into a training set (20,000 utterances, about 30 hours), a development set (5000 utterances, about 10 hours) and a test set (3000 utterances, about 5 hours) where the number of 1-5 speakers was averaged. The length of utterances varied from about 0.8s to 16s. Both of the training set and the development set were randomly mixed from the source audio files in the folder "si_tr_s", while the test set was randomly mixed from the source audio files in the remaining two folders. All audio data was downsampled to 8 kHz in order to reduce memory and computational costs. The data was then randomly mixed them at the signal-to-noise ratio (SNR) varied from 0 to 5 dB.

To simulate multi-channel mixtures, we convolved impulse responses with the speech signals. We used a room impulse response (RIR) generator [24] to spatialize the datasets. As shown

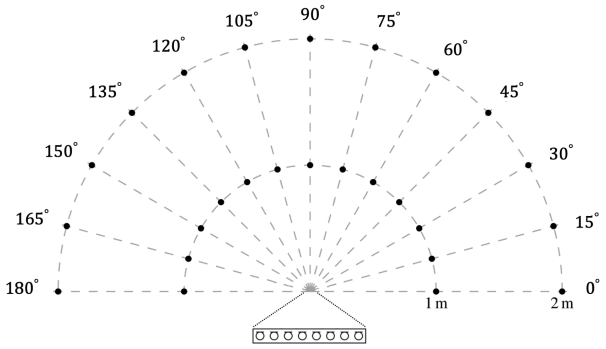

Figure 2: *Illustration of experimental setup.*

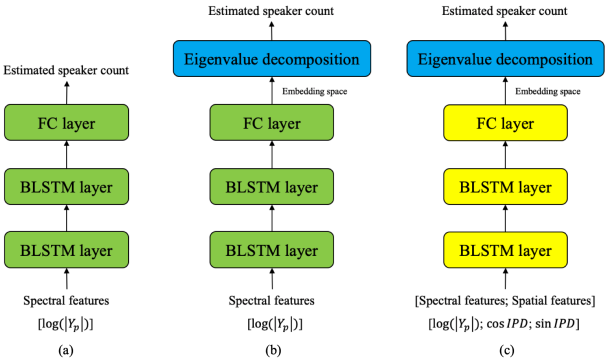

Figure 3: *Schematic diagram of the: (a) CLSC; (b) EBSC; (c) EBMC.*

in Figure 2, our study considered a linear array setup, using eight-microphone linear arrays with 2-2-2-2-2-2-2 cm. Speakers were randomly located in steps of 15° from 0° to 180° and at a distance of 1 m and 2 m to the array center. The length and width of the room were both randomly selected from 5.0 m to 10.0 m, while the height was randomly selected from 3.0 m to 4.0 m. In addition, a T60 value for each mixture was randomly drawn in the range [0.2, 0.7] s for the reverberant dateset, while the T60 value was 0.0 s for the anechoic dataset.

3.2. Settings for BLSTM training

In this work, the deep embedding network had two BLSTM layers with 600 units. The initial learning rate of the Adam learning algorithm was 0.0005, while all models contained random dropouts with a dropout rate 0.5. A tanh activation function was followed by the embedding layer. Our models were implemented using PyTorch deep learning framework. The window length and window shift of STFT were 256 ms and 64 ms respectively.

In addition, the silence regions of the time-frequency bins were ignored in the cost computation during the training process. The silence regions were defined as time-frequency bins where the magnitude was smaller than -40 dB of the maximum mixture magnitude.

3.3. Evaluation metrics

Success Rate [25] was applied for speaker counting performance in this paper. Assuming that $N_c(k)$ is the number of scenarios that the estimated speaker count $\hat{C}$ equals the true speaker count C and $N_t(k)$ is the total number per class k.

$$SR(k) = \frac{N_c(k)}{N_t(k)} \times 100 \tag{10}$$

$SR(k)$ of 100% means that the number of speakers in the experimental setup for the class k is all counted correctly.

3.4. Compared Methods

As shown in Figure 5, the baseline method was re-implemented with our experimental setup, which directly mapped spectral features to a speaker count as described in [14] where $STFT + Classification$ performed best (CLSC). In contrast to the classification-based method CLSC, the embedding-based method (EBSC) did not directly map the speaker count, but firstly mapped each time-frequency bin to an embedding space and then the speaker count could be obtained by eigenvalue decomposition as described in part 2.3. Then a multi-channel method combining spectral and spatial features at the input of networks (EBMC) were introduced into the EBSC. The only difference between the EBMC and the EBSC was that the input of the EBMC was a combination of log magnitude spectrum, cosIPD and sinIPD while the input of the EBSC was only the mixture log magnitude spectrum. What's more, the difference between the EBFMC and the EBMC was that the EBFMC was a feature combination method in speaker embedding space instead of simply combining features at the input layer of neural networks for the EBMC.

3.5. Results and discussion

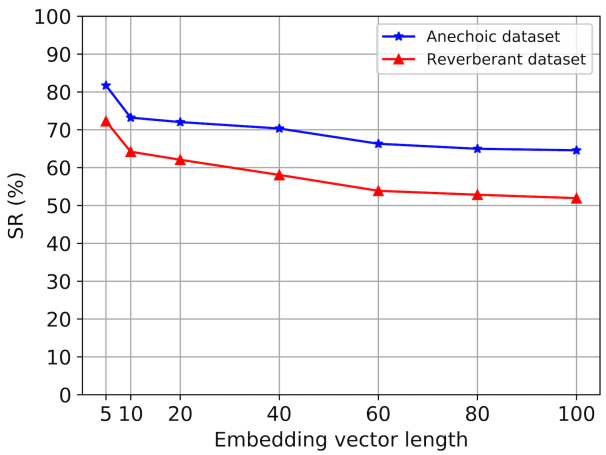

Figure 4: *Success rates with different embedding vector length for the EBSC.*

The only difference between the CLSC and EBSC was the output of the network. The output of the CLSC was one-hot encodings of the number of speakers while the EBSC outputed embedding vectors. To explore the effect of embedding vector length D on speaker counting accuracy, we tuned the embedding vector length of 5,10,20,40,60,80 and 100 on the train set. Figure 3 showed that accuracies decreased as the length of the embedding vector increased. The best SR was 81.7% and 72.3% when the embedding vector length is 5, while the accuracy of CLSC was 70.6% and 57.4% on the anechoic and reverberant dataset respectively.

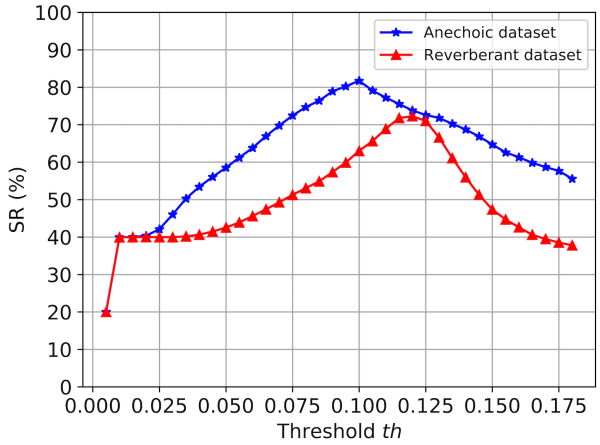

Figure 5: *Success rates with different threshold values for the EBSC.*

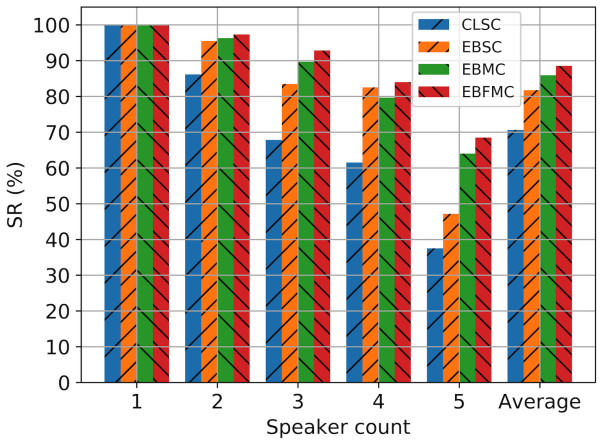

Figure 6: *Success rates for 1- to 5-speaker mixtures on the anechoic test set.*

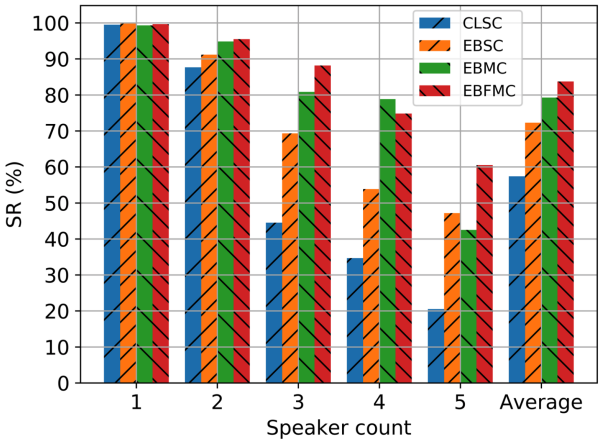

Figure 7: *Success rates for 1- to 5-speaker mixtures on the reverberant test set.*

Then, to find the best parameter th in equation (9), we tuned th with steps of 0.005 from 0.005 to 0.180 on the train set in Figure 4. The best SRs of EBSC were 81.7% with $th = 0.100$ and 72.3% with $th = 0.120$ on the anechoic and reverberant dataset respectively.

Furthermore, Figure 6 and Figure 7 showed SRs for the test set of 1-5 speakers. The average SRs of the CLSC, EBSC, EBMC and EBFMC were 70.6%, 81.7%, 85.9% and 88.5% respectively on the anechoic test set, while 57.4%, 72.3%, 79.3% and 83.7% respectively on the reverberant test set. We could see that the more speakers, the more difficult it was to estimate the speaker count and the EBSC always performed better than the CLSC. It could be obviously seen that the embedding-based method was better than the method based on classification using neural networks. Compared with EBSC, the EBMC achieved better accuracy on average due to the introduction of spatial features, increasing by 4.2% and 7.0% on the anechoic and reverberant dataset respectively. The reason was that a spectrum embedding space represented the information of speaking contents while a spatial embedding space represented the information of spatial orientations of speakers and they might be complementary to each other. It could be also shown that when the dataset contained reverberation, all speaker count estimation methods significantly performed worse.

Based on the EBMC, the proposed EBFMC counted more accurate than EBMC except for 4-speaker mixtures on the reverberant dataset. EBMC and EBFMC were two multi-channel methods. It could be obviously observed that methods of the feature combination in the embedding space had better SR than the method of the feature combination at the input layer of the network on average.

4. Conclusion

In this paper, we proposed a multi-channel competing speaker count estimation method with joint deep spectral space and spatial embedding space learning and fusion. We firstly trained two BLSTM layers to extract deep embedding features of the spectrum information and spatial information respectively. Then these features were mapped into a new embedding space by a fully connected layer. Finally, We performed source counting by computing the rank of the mean covariance matrix of embedding vectors in the new embedding space. Results showed that the embedding-based method EBSC outperformed the baseline classification-based method CLSC, which meant that compared with directly outputting the speaker count, the eigenvalue decomposition was conducive to more accurate estimation of competing speaker count. The EBMC outperformed the EBSC, which meant that spatial features could help to improve the accuracy of speaker count estimation. In addition, the EBFMC performing better than the EBMC meant that combination of features in the embedding space could achieve a higher speaker counting accuracy. In the future, we will explore how to dereverberate based on the proposed method and apply it to multi-talker speech separation and multi-source localization tasks.

5. Acknowledgements

This work is supported by the National Key Research and Development Program (No.2019YFC1408501), the National Natural Science Foundation of China (No.61175043, No.61421062), and the High-performance Computing Platform of Peking University.

6. References

[1] L. Wang, T. Hon, J. Reiss, and A. Cavallaro, "An iterative approach to source counting and localization using two distant microphones," *IEEE/ACM transactions on audio, speech, and language processing (TASLP)*, vol. 24, no. 6, pp. 1079–1093, 2016.

[2] Y. Li and G. Liu, "Sound classification based on spectrogram for surveillance applications," in *2016 IEEE International Conference on Network Infrastructure and Digital Content (IC-NIDC)*, 2016, pp. 293–297.

[3] N. Kanda, Y. Fujita, S. Horiguchi, R. Ikeshita, K. Nagamatsu, and S. Watanabe, "Acoustic modeling for distant multi-talker speech recognition with single-and multi-channel branches," in *Acoustics, Speech and Signal Processing (ICASSP), 2019 IEEE International Conference on*, 2019, pp. 6630–6634.

[4] I. Gebru, S. Ba, X. Li, and R. Horaud, "Audio-visual speaker diarization based on spatiotemporal bayesian fusion," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 5, pp. 1086–1099, 2017.

[5] G. Sell, A. McCree, and D. Garcia-Romero, "Priors for speaker counting and diarization with ahc." in *INTERSPEECH*, 2016, pp. 2194–2198.

[6] B. Yang, H. Liu, C. Pang, and X. Li, "Multiple sound source counting and localization based on tf-wise spatial spectrum clustering," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 27, no. 8, pp. 1241–1255, 2019.

[7] D. Pavlidi, A. Griffin, M. Puigt, and A. Mouchtaris, "Source counting in real-time sound source localization using a circular microphone array," in *2012 IEEE 7th Sensor Array and Multichannel Signal Processing Workshop (SAM)*, 2012, pp. 521–524.

[8] B. Yang, H. Liu, and C. Pang, "Multiple sound source counting and localization based on spatial principal eigenvector." in *INTERSPEECH*, 2017, pp. 1924–1928.

[9] S. Arberet, R. Gribonval, and F. Bimbot, "A robust method to count and locate audio sources in a multichannel underdetermined mixture," *IEEE Transactions on Signal Processing*, vol. 58, no. 1, pp. 121–133, 2009.

[10] R. Balan, "Estimator for number of sources using minimum description length criterion for blind sparse source mixtures," in *International Conference on Independent Component Analysis and Signal Separation*. Springer, 2007, pp. 333–340.

[11] R. Nadakuditi and A. Edelman, "Sample eigenvalue based detection of high-dimensional signals in white noise using relatively few samples," *IEEE Transactions on Signal Processing*, vol. 56, no. 7, pp. 2625–2638, 2008.

[12] K. Yamamoto, F. Asano, W. van Rooijen, E. Ling, T. Yamada, and N. Kitawaki, "Estimation of the number of sound sources using support vector machines and its application to sound source separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2003 IEEE International Conference on*, vol. 5, 2003, pp. V–485.

[13] W. Hu, R. Liu, X. Lin, Y. Li, X. Zhou, and X. He, "A deep learning method to estimate independent source number," in *2017 4th International Conference on Systems and Informatics (ICSAI)*, 2017, pp. 1055–1059.

[14] F. Stöter, S. Chakrabarty, B. Edler, and E. Habets, "Classification vs. regression in supervised learning for single channel speaker count estimation," in *Acoustics, Speech and Signal Processing (ICASSP), 2018 IEEE International Conference on*, 2018, pp. 436–440.

[15] F. Stöter, S. Chakrabarty, B. Edler, and E. Habets, "Countnet: Estimating the number of concurrent speakers using supervised learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 27, no. 2, pp. 268–282, 2018.

[16] P. Grumiaux, S. Kitic, L. Girin, and A. Guérin, "High-resolution speaker counting in reverberant rooms using crnn with ambisonics features," *arXiv preprint arXiv:2003.07839*, 2020.

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

[18] Y. Isik, J. L. Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," in *INTERSPEECH*, 2016, pp. 545–549.

[19] T. Higuchi, K. Kinoshita, M. Delcroix, K. Zmolíková, and T. Nakatani, "Deep clustering-based beamforming for separation with unknown number of sources." in *INTERSPEECH*, 2017, pp. 1183–1187.

[20] T. Yoshioka, H. Erdogan, Z. Chen, and F. Alleva, "Multi-microphone neural speech separation for far-field multi-talker speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2018 IEEE International Conference on*, 2018, pp. 5739–5743.

[21] Z. Wang and D. Wang, "Combining spectral and spatial features for deep learning based blind speaker separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 27, no. 2, pp. 457–468, 2018.

[22] X. Chen, L. Xu, Z. Liu, M. Sun, and H. Luan, "Joint learning of character and word embeddings," in *Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI)*, 2015, pp. 1236–1242.

[23] F. Wang, W. Zuo, L. Lin, D. Zhang, and L. Zhang, "Joint learning of single-image and cross-image representations for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1288–1296.

[24] E. Habets, "Room impulse response generator," *Technische Universiteit Eindhoven, Tech. Rep*, vol. 2, no. 2.4, p. 1, 2006.

[25] S. Pasha, J. Donley, C. Ritz, and Y. Zou, "Towards real-time source counting by estimation of coherent-to-diffuse ratios from ad-hoc microphone array recordings," in *2017 Hands-free Speech Communications and Microphone Arrays (HSCMA)*, 2017, pp. 161–165.