Blind Source Extraction Based on Multi-channel Variational Autoencoder and X-vector-based Speaker Selection Trained with Data Augmentation

Zhaoyi Gu¹, Lele Liao¹, Jing Lu¹

¹Key Laboratory of Modern Acoustics, Institute of Modern Acoustics, Nanjing University, Nanjing 210093, China

guzhaoyi@mail.nju.edu.cn, LiaoLele@mail.nju.edu.cn, lujing@nju.edu.cn

Abstract

Multi-channel variational autoencoder (MVAE) has been proven to be a promising method for blind source separation (BSS), which is an elegant combination of the strong modeling power of deep neural networks (DNN) and interpretable BSS algorithm based on independence assumption. However, the success of MVAE is limited to the training with very few speakers and the speeches of desired speakers are usually included. In this paper, we develop a sequential approach for blind source extraction (BSE) by combining MVAE with the x-vector based speaker recognition (SR) module. Clean speeches of 500 speakers are utilized in training MVAE to verify its capability of generalization to unseen speakers, and an augmented dataset is constructed to train the SR module. The efficacy of the proposed BSE approach in terms of extraction accuracy, signal-to-interference ratio (SIR) and signal-to-distortion ratio (SDR) are validated with test data consisting of unseen speakers under varied environments.

Index Terms: blind source extraction, multi-channel variational autoencoder

1. Introduction

Multi-channel blind source extraction (BSE) solves a similar but more challenging problem than blind source separation [1] (BSS) since BSS focuses only on the recovery of all source signals, whereas BSE further aims to effectively extract the source(s) of interest (SOI) from the multi-channel output. BSE is a more appropriate choice as an automatic speech recognition (ASR) front-end because it can directly output the enhanced target speech signal.

Signal-processing based BSE has been widely investigated, in which traditional methods such as beamforming and time-frequency masking are combined with BSS methods to obtain the SOI [2–3]. Recently, two novel BSE algorithms based on a reformulated BSS mixing model are proposed to deal with the extraction of a non-Gaussian signal from a Gaussian background [4, 5]. The method is further extended to enhance the performance under a non-Gaussian background [6], and the extraction accuracy can be further improved by introducing a partially supervised structure with pilot signals in an online implementation [7, 8]. However, these methods usually assume a fixed signal model for speech, which restricts their performance in complex situations.

Leveraging the advances in deep learning, DNN-based BSE has also attracted much attention. In particular, several speaker aware neural networks have been proposed to guide an extraction network towards the learning of a target time-frequency mask by using an adaptation utterance of the target speaker [9–11]. However, as a pure supervised approach, end-to-end DNN-based BSE faces the challenge of generalization. Furthermore, most of the end-to-end methods focus on mono-channel application, and cannot be extended to microphone arrays by effectively exploiting the spatial information.

Multi-channel variational autoencoder (MVAE) has been recently proposed as an intriguing hybrid BSS method that combines data-driven DNN-based signal model and interpretable BSS algorithm with separation matrix updated by iterative projection. The conditional variational autoencoder (CVAE) [12, 13] is utilized as the generative model of speech (ACVAE), a variant of the standard CVAE that strengthens dependencies between the decoder outputs and the conditional variables. Although theoretically the MVAE effectively combines the benefits of both data-driven and rule-based methods, its success has only been validated with training on very few speakers, and its capability of generalizing to unseen speakers remains to be verified on larger datasets.

In this paper, we focus on the study of BSE by sequentially combining MVAE with the x-vector based speaker recognition (SR) module [15]. Clean speeches of 500 speakers are utilized in training the CVAE network to verify the generalization performance of MVAE to unseen speakers. In addition, an augmented dataset consisting of separated utterances under a variety of simulated configurations is constructed to train the SR module. Comprehensive results regarding extraction accuracy, signal-to-interference-ratio (SIR) and signal-to-distortion ratio (SDR) are presented to validate the efficacy of the proposed BSE system.

2. MVAE based multi-channel blind source separation

2.1. Problem formulation

The signal model of a determined situation is assumed, where an array of M microphones is utilized to capture the signals from M sources. After transforming the signals into short-time Fourier transform (STFT) domain and ignoring the noise, the mixing model can be represented with an instantaneous model as

\[ \mathbf{x}_f = \mathbf{A}_f \mathbf{s}_f, \]

where \( \mathbf{s}_f = [s_{f1}, s_{f2}, \ldots, s_{fM}]^T \), \( \mathbf{x}_f = [x_{f1}, x_{f2}, \ldots, x_{fM}]^T \) are the multi-channel vectors containing the source and observed signals at frequency bin \( f \) and time frame \( t \) respectively, with \( f = 1, \ldots, F \) and \( t = 1, \ldots, T \). Subscript \( m \) denotes the index of either sources or microphones based on the context, and \( [\cdot]^T \) is
a notation for non-conjugate transposition. \(A_r\) is an \(M \times M\) complex-valued mixing matrix containing the information of the room impulse responses in the frequency domain. When \(A_r\) is invertible, estimated source signals \(y_\beta = [y_{\beta,1}, y_{\beta,2}, \ldots, y_{\beta,M}]^T\) can be obtained by multiplying a demixing matrix \(W_f\) as \(W_f, \ldots, W_{FS}\) as

\[
y_\beta = W_f x_\beta ,
\]

where \([\cdot]^H\) is a notation for conjugate transposition.

Let \(S_m = \{S_{\beta,m}\}_{\beta,m}\) be the vector containing the \(m\)th source signals at all \((t, f)\) bins, which is assumed in MVAE to follow a local Gaussian model (LGM)

\[
p_x(s_m) = \prod_\beta \mathcal{N}(s_{\beta,m} | 0, \nu_{\beta,m}) ,
\]

where signals from different \((t, f)\) bins are independently characterized using a zero-mean circular symmetric Gaussian distribution with \(s_{\beta,m}\) being the estimated power spectral density. Assuming the independence among different sources, the signal model for the observed signals \(x_\beta\) at time-frequency bin \((t, f)\) can be derived as

\[
p_x(x_\beta) = \left| \text{det} W_f \right| \prod_\beta \mathcal{N}(y_{\beta,m} | 0, \nu_{\beta,m}) ,
\]

where \(\text{det}(\cdot)\) is the determinant notation. Hence, the log likelihood function of the demixing matrix \(W_f\) and the source model parameters \(\nu_{\beta,m}\) given the observed signals \(x_\beta\) becomes

\[
L_{W_f} = -\sum_{\beta,m} \log v_{\beta,m} + \frac{|y_{\beta,m} - \nu_{\beta,m}|^2}{v_{\beta,m}} + 2T \sum f \text{det} W_f ,
\]

where \(\nu\) denotes the whole set of \(\{s_{\beta,m}\}_{\beta,m}\). Dependencies among time-frequency bins in \(y_{\beta,m}\) given source \(m\) are additionally characterized with a CVAE decoder model using a CNN architecture, so that theoretically the permutation ambiguity can be simultaneously eliminated during the separation process [16].

### 2.2. Conditional variational autoencoder

CVAE learns a multi-modal deep generative model of the speech spectrograms \(S\) by introducing a hierarchical structure with a latent vector \(z\) and a condition vector \(c\) that contains class information. Specifically, a disentangled representation is represented by maximizing the following log likelihood function

\[
L_{\text{CVAE}} = \sum_j \log p_j(S | z, c)p(z)dz ,
\]

where \(p_j(S | z, c)\) is the conditional probability distribution of speech spectrogram and \(p(z)\) represents the prior distribution of the latent vector. Subscript \(j\) is used instead of \(m\) to denote an arbitrary CVAE training example. Since (6) cannot be directly optimized, a lower bound is derived using Jenson’s inequality as

\[
Q_{\text{CVAE}} = \sum_j \left[ q_j(z | S, c) \log p_j(S | z, c)p(z) \right] - \sum_j \text{KL}(q_j(z | S, c) || p(z)) ,
\]

where \(\text{KL}(\cdot || \cdot)\) denotes the Kullback-Leibler divergence between two probability density functions (PDF). To optimize (7), CVAE employs an encoder network to estimate the posterior distribution \(q_j(z | S, c)\) of the latent vector, and a decoder network to produce the generative distribution \(p_j(S | z, c)\), where \(\phi, \psi\) are symbols denoting the parameters of the two networks respectively.

In this work, \(c\) is assigned as a one-hot vector indicating the speaker identity, and \(p_j(S | x, c) = q_j(z | S, c), p(z)\) are assumed to be Gaussian distributions. Specifically,

\[
p_j(S | z, c) = \prod_\beta \mathcal{N}(z_{\beta,m} | 0, \sigma_{z_{\beta,m}}^2(f, t, z, c)) ,
\]

\[
q_j(z | S, c) = \prod_\beta \mathcal{N}(d_S^{(j)}, \nu_{\beta,m}) ,
\]

\[
p_j(z | c) = \mathcal{N}(z | 0, 1) .
\]

where \(\sigma_{z_{\beta,m}}^2(f, t, z, c)\) denotes the generated source spectrogram, and \(\mu_d^{(j)}(S, c), \sigma_d^{(j)}(S, c)\) are the posterior mean and variance of the \(d\)th element in the latent vector. Note that (8) possesses the form of an LGM, so that the training objective of CVAE is in consistent with the MVAE objective depicted in (5). Figure 1 shows the configuration of the CVAE network.

### 2.3. MVAE optimization process

After training the CVAE model, the decoder network is utilized as a deep source model in MVAE. Since the CVAE training utterances might be different from the source signals in terms of global energy, a scale factor \(g_m\) is introduced to compensate for the difference. Therefore, the source model used in MVAE can be specified as

\[
p_x(s_m) = \prod_\beta \mathcal{N}(s_{\beta,m} | 0, g_m \sigma_{\beta,m}^2(f, t, z, c)) .
\]

Incorporating (11) into the MVAE objective function (5), updating rules for \(g_m, W_f, z\) and \(c\) can be derived as follows

\[
g_m = \frac{1}{T} \sum_j \frac{|y_{\beta,m}|^2}{v_{\beta,m}} ,
\]

\[
V_f = \frac{1}{T} \sum_j \sigma_{\beta,m}^2(f, t, z, c) ,
\]

\[
w_f = \frac{1}{T} \left| \frac{wx_{\beta,m}}{V_f} \right| e_m ,
\]

where (12) is derived by setting the derivative of (5) with respect to \(g_m\) to zero and (13) – (15) are derived using the iterative-projection (IP) method [17]. \(e_m\) is the \(m\)th column of
Table 1: Extraction accuracy using the simulated RIRs.

| BSS Algorithm | Performance | SDR [dB] | Average/ Sum |
|---------------|-------------|----------|---------------|
|               | accuracy rate [%] |          |               |
| ILRMA         | 73.33       | 90.08    | 95.31         | 98.20 | 99.58 | 99.91 | 100 | 100 | 97.03 |
|               | No. utterances | 255      | 484           | 1259 | 1727 | 1654 | 1123 | 517 | 181 | 7200 |
| MVAE 1        | accuracy rate [%] | 69.94    | 81.65         | 86.05 | 96.02 | 98.75 | 99.42 | 100 | 100 | 93.21 |
|               | No. utterances | 469      | 594           | 1118 | 1431 | 1519 | 1205 | 603 | 261 | 7200 |
| MVAE 2        | accuracy rate [%] | 70.39    | 83.51         | 91.81 | 97.92 | 99.26 | 99.53 | 100 | 100 | 95.63 |
|               | No. utterances | 331      | 479           | 1075 | 1537 | 1616 | 1276 | 626 | 260 | 7200 |

an $M \times M$ identity matrix, and $(\cdot)^{-1}$ is the notation for matrix inversion. Both $z_a$ and $c_\alpha$ are updated via back propagation.

3. Target speech extraction

After the multi-channel outputs are obtained using the MVAE algorithm as described in section 2, the target speech can be identified and selected by matching the x-vectors of all channel candidates with that of the enrollment utterance.

3.1. x-vector based SR system

The x-vector embedding system used in the paper has the same network structure as illustrated in [18], and the embeddings are extracted from the affine component of layer segment 6. The input features are 30-dimensional MFCCs with a frame length of 64 ms and a frame hop of 16 ms. For each batch of data, the input length is randomly chosen within a range of 160 to 200 frames, and each training example is mean-and-variance-normalized. Energy-based VAD is computed as the average of x-vectors extracted from each window.

The x-vectors are centered and the dimensionality is reduced to 128 using linear discriminant analysis (LDA). After that, the embeddings are length normalized and a PLDA backend is trained using the Bob toolbox [19, 20]. The BSS output whose x-vector has the highest joint probability with the enrolled x-vector is extracted as the target signal.

4. Experimental Setup

In our experiments, the independent low-rank matrix analysis (ILRMA) [21], regarded as the SOTA rule-based BSS algorithm, is also combined with the SR module to form a BSE system for comparison. Both the CVAE source model and the x-vector embedding system are trained using two “clean” training subsets in the Librispeech dataset [22]. Specifically, 500 and 1168 different speakers are selected for the training purpose of the two tasks respectively. To improve the robustness of target speech extraction described in section 3, artificial separated speeches are employed as augmented data to help train the x-vector based SR system.

4.1. Data augmentation

In order to obtain separated signals for data augmentation, binaural mixed signals of 5 to 30 seconds are generated by convolving the clean training utterances with simulated room impulse responses (RIRs). In our experiment, a total number of 4410 RIRs are created so as to cover a wide range of mixing scenario. Specifically, we consider seven different reverberation times (RT), ranging from 0.1 s to 0.7 s with a 0.1 s interval. Under each RT, 21 regular sized rooms are created, with the length, width and height lying in the range of [3 m, 15 m], [3 m, 15 m] and [2 m, 4 m] respectively. For each room, we generate 30 RIRs using the image model method [23] according to the following configurations. Microphone arrays are randomly located in the room with a margin of [1 m, 1.5 m] to the wall and an average element interval of 0.2 m. Both sources lie on the same side of the array with the direction of arrival (DOA) interval set in the range of [20°, 160°]. The height of sources and microphones are set to be 1.2 m, and the distances between array center and sources are randomly chosen from [0.5 m, $r_c + 0.5$ m], where $r_c$ denotes the critical distance.

From our numerous tests, although the MVAE has comparatively better performance on average than the ILRMA, it suffers from a more serious block permutation problem when the training set includes speech from a large number of speakers. Therefore, we use the ILRMA to produce the augmented dataset. In total, two million training examples are generated by randomly picking utterances and offsets from each training speaker, and the number of examples extracted from the clean and augmented datasets are roughly the same.

4.2. Evaluation configuration

We evaluate the extraction performance of the proposed BSE method using both simulated and real RIRs under the $2 \times 2$ scenario, where there are two microphones and two sources, with one source containing the target speech signal. The evaluation mixtures are created by convolving source signals with either simulated or real RIRs. Both the sources and the corresponding enrollment utterances are also selected from the database of Librispeech, containing 40 speakers different from those in the training dataset. In other words, we investigate the situation where both the x-vector SR system and the MVAE algorithm are generalized to unseen speakers. The length of all evaluation mixtures is set to be 5 s, and the durations of enrollment utterances used in extraction are 30 s. In addition, we make sure that the utterances used for the enrollment are different from those used to generate evaluation mixtures.

The simulated RIRs are created with a room size of $8 m \times 7 m \times 3 m$ and a reverberation time varying from 0.15 s to 0.65 s. The center of a dual-channel microphone array with an element interval of 10 cm is located at [4.8 m, 4.3 m, 1.2 m]. Both two sources are placed a meter away from the array.
corresponding SDR range, and the accuracy rates are calculated as the ratio between the number of correctly extracted utterances and the total number of trials, i.e. No. utterances. The last column shows the average accuracy rates and the total number of test utterances respectively. It can be seen that the accuracy rates of all investigated systems increase monotonically with the SDR values, implying a positive correlation between the separation performances of front-end BSS system and the extraction accuracy of the tandem BSE method. Focusing on the No. utterances in SDR larger than 16 dB, one can see that the MVAE algorithms produces more high-quality speeches than ILRMA, validating that the MVAE generalizes well to unseen speakers. MVAE 2 has comparatively more higher SDR output than MVAE 1, indicating using ILRMA to initialize MVAE does help to further improve the separation performance. However, it should be noted that the ILRMA based BSE outperforms the MVAE based BSE regarding the extraction accuracy. This is also attributed to more block permutation (BP) problems observed in the MVAE-extracted speeches. An implicit evidence of this problem can be inferred by looking at the No. utterances corresponding to SDR ≤ 0 dB, where MVAE 1 produces the most low-quality speeches. Note that the implementation of ILRMA initialization only alleviates but not fully solves this problem since the No. utterances of MVAE 2 under SDR ≤ 0 still exceeds that of the ILRMA.

Experimental results using the real RIR dataset is illustrated in Table 2, where the performance is assessed under different reverberation times. The effectiveness of data augmentation is validated by comparing the extraction accuracy performed with and without data augmentation. It can be seen that data augmentation increases the extraction accuracy in all configurations. Columns indexed with ‘Extracted speech quality’ presents the average SIR and SDR values of correctly extracted speeches, and the best results are obtained with the MVAE using ILRMA initialization.

### Table 2: Extraction performances using the real RIR dataset.

| Time [s] | Algorithm | Extraction accuracy | Extracted speech quality |
|----------|-----------|---------------------|--------------------------|
|          |           | Without data augmentation | With data augmentation | SDR [dB] | SIR [dB] |
| 0.16     | ILRMA     | 97.42%              | 98.79%                   | 6.97     | 13.77    |
|          | MVAE 1    | 92.27%              | 94.55%                   | 6.46     | 12.24    |
|          | MVAE 2    | 96.82%              | 97.58%                   | 7.59     | 15.67    |
| 0.36     | ILRMA     | 95.61%              | 97.95%                   | 4.79     | 11.86    |
|          | MVAE 1    | 90.38%              | 93.03%                   | 4.75     | 11.57    |
|          | MVAE 2    | 95.08%              | 97.35%                   | 5.20     | 13.37    |
| 0.61     | ILRMA     | 93.33%              | 96.67%                   | 3.20     | 10.93    |
|          | MVAE 1    | 88.03%              | 91.26%                   | 3.38     | 10.91    |
|          | MVAE 2    | 92.27%              | 95.35%                   | 3.71     | 12.28    |

### 4.3 Results and discussion

We evaluate the extraction performance in terms of the extraction accuracy, SIR and SDR. The SIR and SDR are calculated using the BSS_EVAL toolbox [25].

Table 1 shows the results of our experiment using the simulated RIR, where the three BSE systems are summarized separately within eight non-overlapping SDR intervals. Rows labeled with ‘No. utterances’ presents the total number of separated utterances whose SDR falls into the corresponding SDR range, and the accuracy rates are calculated as the ratio between the number of correctly extracted utterances and the total number of trials, i.e. No. utterances. The last column shows the average accuracy rates and the total number of test utterances respectively. It can be seen that the accuracy rates of all investigated systems increase monotonically with the SDR values, implying a positive correlation between the separation performances of front-end BSS system and the extraction accuracy of the tandem BSE method. Focusing on the No. utterances in SDR larger than 16 dB, one can see that the MVAE algorithms produces more high-quality speeches than ILRMA, validating that the MVAE generalizes well to unseen speakers. MVAE 2 has comparatively more higher SDR output than MVAE 1, indicating using ILRMA to initialize MVAE does help to further improve the separation performance. However, it should be noted that the ILRMA based BSE outperforms the MVAE based BSE regarding the extraction accuracy. This is also attributed to more block permutation (BP) problems observed in the MVAE-extracted speeches. An implicit evidence of this problem can be inferred by looking at the No. utterances corresponding to SDR ≤ 0 dB, where MVAE 1 produces the most low-quality speeches. Note that the implementation of ILRMA initialization only alleviates but not fully solves this problem since the No. utterances of MVAE 2 under SDR ≤ 0 still exceeds that of the ILRMA.

Experimental results using the real RIR dataset is illustrated in Table 2, where the performance is assessed under different reverberation times. The effectiveness of data augmentation is validated by comparing the extraction accuracy performed with and without data augmentation. It can be seen that data augmentation increases the extraction accuracy in all configurations. Columns indexed with ‘Extracted speech quality’ presents the average SIR and SDR values of correctly extracted speeches, and the best results are obtained with the MVAE using ILRMA initialization.

### 5. Conclusion

This paper proposes a two-stage BSE method by combining the MVAE BSS method with an x-vector based speaker recognition module. The generalization of the MVAE to unseen speakers is verified with a training set of a large number of speakers. An augmented dataset using artificial separated signals is constructed to train the x-vector embedding network and the PLDA backend. Experimental results confirm the effectiveness of the proposed scheme and demonstrate that the BSE based on MVAE initialized with ILRMA produces the best extracted speech quality. However, it should be noted that the MVAE based BSE has a slightly less extraction accuracy than ILRMA based BSE due to comparatively more serious block permutation problem.

### 6. Acknowledgement

The National Science Foundation of China supported this work with grant number 11874219.
7. References

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

[2] L. Wang, H. Ding, and F. Yin, “Target speech extraction in cocktail party by combining beamforming and blind source separation,” Acoustics Australia, vol. 39, no. 2, pp. 64-68, 2011.

[3] H. Sawada, S. Araki, R. Mukai, and S. Makino, “Blind extraction of dominant target sources using ICA and time-frequency masking,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, no. 6, pp. 2165-2173, 2006.

[4] Z. Koldovský, P. Tichavský, and V. Kautský, “Orthogonally constrained independent component extraction: Blind MPDR beamforming,” in EUSIPCO 2017 – 26th European Signal Processing Conference, August 28-September 2, Kos island, Greece, Proceedings, 2017, pp. 1155–1159.

[5] Z. Koldovský and P. Tichavský, “Gradient algorithms for complex non-gaussian independent component/vector extraction, question of convergence,” IEEE Transactions on Signal Processing, vol. 67, no. 4, pp. 1050–1064, 2019.

[6] Z. Koldovský, P. Tichavský, and N. Ono, “Orthogonally-constrained extraction of independent non-Gaussian component from non-Gaussian background without ICA,” in LVA/ICA 2018 – 14th International Conference on Latent Variable Analysis and Signal Separation, July 2-5, Guildford, UK, Cham, Switzerland: Springer, 2018, pp. 161–170.

[7] F. Nesta and Z. Koldovský, “Supervised independent vector analysis through pilot dependent components,” in ICASSP 2017 – 42nd IEEE International Conference on Acoustics, Speech and Signal Processing, March 5-9, New Orleans, LA, USA, Proceedings, 2017, pp. 536–540.

[8] J. Janský, J. Málek, J. Čmejla, T. Kounovský, Z. Koldovský, and J. Žďánský, “Adaptive blind audio source extraction supervised by dominant speaker identification using x-vectors,” in ICASSP 2020 – 45th IEEE International Conference on Acoustics, Speech and Signal Processing, March 4-8, Virtual Barcelona, Spain, Proceedings, 2020, pp. 676-680.

[9] J. Wang et al., “Deep extractor network for target speaker recovery from single channel speech mixtures,” in INTERSPEECH 2018 – 19th Annual Conference of the International Speech Communication Association, September 2-6, Hyderabad, India, Proceedings, 2018, pp. 307–311.

[10] Q. Wang et al., “VoiceFilter: Targeted voice separation by speaker-conditioned spectrogram masking,” in INTERSPEECH 2019 – 20th Annual Conference of the International Speech Communication Association, September 15-19, Graz, Austria, Proceedings, 2019, pp. 2728-2732.

[11] K. Žmolíková et al., “SpeakerBeam: Speaker aware neural network for target speaker extraction in speech mixtures,” IEEE Journal of Selected Topics in Signal Processing, vol. 13, no. 4, pp. 800–814, 2019.

[12] H. Kameoka, L. Li, S. Inoue, and S. Makino, “Supervised determined source separation with multichannel variational autoencoder,” Neural Computation, vol. 31, no. 9, pp. 1891–1914, 2019.

[13] D. P. Kingma and D. J. Rezende, S. Mohamed, and M. Welling, “Semi-supervised learning with deep generative models,” in NIPS 2014 – 28th Conference on Neural Information Processing Systems, December 8-13, Montreal, QC, Canada, Proceedings, 2014, pp. 3581–3589.

[14] L. Li, H. Kameoka, and S. Makino, “Fast MVAE: Joint separation and classification of mixed sources based on multichannel variational autoencoder with auxiliary classifier,” in ICASSP 2019 – 44th IEEE International Conference on Acoustics, Speech and Signal Processing, May 12-17, Brighton, UK, Proceedings, 2019, pp. 546–550.

[15] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, “Deep neural network embeddings for text-independent speaker verification,” in INTERSPEECH 2017 – 18th Annual Conference of the International Speech Communication Association, August 26-24, Stockholm, Sweden, Proceedings, 2017, pp. 999–1003.

[16] T. Kim, T. Eltoft, and T. W. Lee, “Independent vector analysis: An extension of ICA to multivariate components,” in ICA 2006 – 6th International Conference on Independent Component Analysis and Signal Separation, March 5-8, Charleston, SC, USA, Proceedings, 2006, pp. 165 – 172.

[17] N. Ono, “Stable and fast update rules for independent vector analysis based on auxiliary function technique,” in WASPAA 2011 – 12th IEEE Workshop Applications of Signal Processing to Audio and Acoustics, October 16-19, New Palz, NY, USA, Proceedings, 2011, pp. 189–192.

[18] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, “X-vectors: Robust DNN embeddings for speaker recognition,” in ICASSP 2018 – 43rd IEEE International Conference on Acoustics, Speech and Signal Processing, April 15-20, Calgary, AB, Canada, Proceedings, 2018, pp. 5329–5333.

[19] A. Anjos, L. El-Shafey, R. Wallace, M. Günther, C. McCool, and S. Marcel, “Bob: a free signal processing and machine learning toolbox for researchers,” in ACM Multimedia 2012 – 20th ACM international conference on Multimedia, October 29-November 2, Nara, Japan, Proceedings, 2012, pp. 1449–1452.

[20] S. J. D. Prince and J. H. Elder, “Probabilistic linear discriminant analysis for inferences about identity,” in ICCV 2007 – 11th IEEE International Conference on Computer Vision, October 14-20, Rio de Janeiro, Brazil, 2007, pp. 1–8.

[21] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, “Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization,” IEEE Transactions on Audio Speech & Language Processing, vol. 24, no. 9, pp. 1626–1641, 2016.

[22] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An ASR corpus based on public domain audio books,” in ICASSP 2015 – 40th IEEE International Conference on Acoustics, Speech and Signal Processing, April 19-24, Brisbane, QLD, Australia, Proceedings, 2015, pp. 5206–5210.

[23] E. A. P. Habets, “Room impulse response generator,” Technische Universiteit Delft, The Netherlands, 2007, pp. 1-17, 2010.

[24] E. Hadad, F. Heese, P. Vary, and S. Gannot, “Multichannel audio database in various acoustic environments,” in IWAENC 2014 – 14th International Workshop on Acoustic Signal Enhancement, September 8-11, Juan-les-Pins, Alpes Maritimes, France, Proceedings, 2014, pp. 313–317.

[25] E. Vincent, R. Grinblat, 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, 2006.