symmetry

MDPI

Article

Multichannel Variational Autoencoder-Based Speech Separation in Designated Speaker Order

Lele Liao, Guoliang Cheng, Haoxin Ruan, Kai Chen and Jing Lu *

Key Laboratory of Modern Acoustics, Institute of Acoustics, Nanjing University, Nanjing 210093, China
* Correspondence: lujing@nju.edu.cn

Abstract: The multichannel variational autoencoder (MVAE) integrates the rule-based update of a separation matrix and the deep generative model and proves to be a competitive speech separation method. However, the output (global) permutation ambiguity still exists and turns out to be a fundamental problem in applications. In this paper, we address this problem by employing two dedicated encoders. One encodes the speaker identity for the guidance of the output sorting, and the other encodes the linguistic information for the reconstruction of the source signals. The instance normalization (IN) and the adaptive instance normalization (adaIN) are applied to the networks to disentangle the speaker representations from the content representations. The separated sources are arranged in designated order by a symmetric permutation alignment scheme. In the experiments, we test the proposed method in different gender combinations and various reverberant conditions and generalize it to unseen speakers. The results validate its reliable sorting accuracy and good separation performance. The proposed method outperforms the other baseline methods and maintains stable performance, achieving over 20 dB SIR improvement even in high reverberant environments.

Keywords: multichannel variational autoencoder; global permutation; instance normalization

Citation: Liao, L.; Cheng, G.; Ruan, H.; Chen, K.; Lu, J. Multichannel Variational Autoencoder-Based Speech Separation in Designated Speaker Order. *Symmetry* **2022**, *14*, 2514. https://doi.org/10.3390/sym14122514

Academic Editors: Chengshi Zheng, Xiaodong Li and Jinqiu Sang

Received: 11 October 2022
Accepted: 22 November 2022
Published: 28 November 2022

1. Introduction

Blind source separation (BSS) aims to separate out individual source signals from their mixtures and plays an important role in frontend processing for speech communication and human machine interaction. A widely used approach for convolutive mixture separation is independent vector analysis (IVA), which effectively solves the frequency components alignment problem of the independent component analysis (ICA) by modeling the inter-frequency dependence of each source with a proper joint probability distribution. The prior source models largely determine the performance of IVA [1], and the commonly utilized ones include the spherically symmetric Laplace (SSL) distribution [2], the Gaussian mixture model [3,4] with its variant [5], the multivariate Gaussian [6], and the generalized Gaussian distributions [7]. Instead of a fixed prior source model, a hybrid distribution [8] was proposed to better model the non-stationarity of speech. Furthermore, the independent low-rank matrix analysis (ILRMA) [9,10] unifies the update rules of auxiliary function-based IVA (AuxIVA) [11] with the source model based on non-negative matrix factorization (NMF) [12,13]. NMF is a flexible model good at capturing the spectral structures and approximating the spectrogram by summing a set of spectral templates scaled by time-varying activations [14]. However, it has been noted that the low-rank assumption of the source spectrogram in the ILRMA restricts its performance in practical applications [15–17].

Apart from the pure rule-based BSS methods, the supervised method has benefited from the rapid development in deep learning [18]. It has also been noted that the deep generative models can be utilized to effectively improve the multichannel BSS method. In the framework of the multichannel variational autoencoder (MVAE) [15], the rule-based iterative demixing matrix updating procedure used in AuxIVA is combined with the deep generative source model based on the conditional variational autoencoder (CVAE) [19].

The MVAE leverages the strong representation power of deep neural networks (DNN) and, in the meantime, ensures the non-decrement of the log-likelihood at each iteration [15].

The global permutation ambiguity is inherent in BSS, i.e., we cannot determine the order of the separated sources [20]. In spite of a promising separation performance, the MVAE still faces this challenge. Pragmatically, the order of the separated sources can be determined by the speaker identities so that we can choose the speech from the target speaker. Inspired by the supervised CVAE model with a speaker label, it is straightforward to specify the speaker label(s) in the MVAE during the separation process [21]. However, it has been demonstrated that the speaker label is likely to degenerate, meaning that the sources can be reconstructed regardless of the speaker label. To address this limitation, the fast MVAE method [22] uses an auxiliary classifier VAE (ACVAE) [23] to learn a more disentangled representation, but its source classification (i.e., speaker identification) accuracy rate (about 80%) is still not satisfactory. Moreover, the one-hot encoding of speaker labels in the MVAE and fast MVAE is hard to generalize to unseen speakers.

To more effectively solve the global permutation problem, in this paper, we employ two encoders with the instance normalization (IN) and adaptive instance normalization (adaIN) [24] layers to learn separately the speaker representation and the content representation, and propose a separation algorithm integrating the task of sorting the outputs in designated speaker order. Unlike the supervised MVAE algorithm, our proposed system can autonomously learn meaningful speaker representations without any speaker labels. To the best of our knowledge, the presented work is the first unsupervised unified system to handle the multitasking of separation and identification. We call the proposed algorithm IN-MVAE to distinguish it from the original MVAE. The experiments carried out in the various scenarios demonstrate its reliable sorting accuracy as well as a better separation performance than the original MVAE algorithm. Moreover, since the IN-MVAE is constructed in an unsupervised way, it also works under speaker-open conditions as indicated by our experimental results.

2. Problem Formulation

We consider the determined BSS where J source signals are captured by I microphones ($J = I$). The observed signal at the i-th microphone is represented as

$$x_i(t) = \sum_{j=1}^{J} \sum_{\tau=0}^{L-1} a_{ij}(\tau) s_j(t-\tau) \tag{1}$$

where the j-th source signal $s_j(t)$ emitted from the j-th speaker is convolved with the room impulse response $a_{ij}(t)$, and L is the length of the impulse response. Performing short-time Fourier Transform (STFT) and assuming that L is shorter than the analysis window length, Equation (1) can be transformed into instantaneous mixtures in the time-frequency domain as

$$\boldsymbol{x}(f,n) = \boldsymbol{A}(f)\boldsymbol{s}(f,n) \tag{2}$$

where f and n denote the indexes of frequencies and frames, respectively. $\boldsymbol{x}(f,n)$ is a vector of I-dimension defined as

$$\boldsymbol{x}(f,n) = [x_1(f,n),\ x_2(f,n),\ \ldots,\ x_I(f,n)]^T \tag{3}$$

$\boldsymbol{s}(f,n)$ is a vector of J-dimension defined as

$$\boldsymbol{s}(f,n) = \left[s_1(f,n),\ s_2(f,n),\ \ldots,\ s_J(f,n)\right]^{\mathrm{T}} \tag{4}$$

where $[\cdot]^{\mathrm{T}}$ denotes the transpose operation, and $\boldsymbol{A}(f)$ is the mixing matrix in the form of

$$\boldsymbol{A}(f) = \begin{bmatrix} a_{11}(f) & \cdots & a_{1J}(f) \\ \vdots & \ddots & \vdots \\ a_{I1}(f) & \cdots & a_{IJ}(f) \end{bmatrix} \tag{5}$$

where $a_{ij}(f)$ is the STFT representation of $a_{ij}(t)$ in the f-th frequency bin, and the time index is omitted due to the time-invariant mixing assumption. The goal of IVA is to figure out the demixing matrix $\boldsymbol{W}(f)$ and estimate the source signals $\boldsymbol{y}(f,n)$ with $\boldsymbol{y}(f,n) = [y_1(f,n),\ y_2(f,n),\ \ldots,\ y_I(f,n)]^{\mathrm{T}}$ by

$$\boldsymbol{y}(f,n) = \boldsymbol{W}(f)\boldsymbol{x}(f,n) \tag{6}$$

where $\boldsymbol{W}(f)$ is a $J \times I$ matrix consisting of J demixing vectors shown as

$$\boldsymbol{W}(f) = \left[w_1(f),\ w_2(f),\ \cdots w_J(f)\right]^{\mathrm{H}} \tag{7}$$

and the superscript H denotes the complex conjugate transpose operation. Under the ideal conditions, the demixing process is symmetrical to the actual signal mixing process [25], so that $\boldsymbol{W}(f) = \boldsymbol{A}^{-1}(f)$. In this paper, we aim to separate the mixture in designated speaker order, i.e., permute the elements of $\boldsymbol{y}(f, n)$ to align $y_j(f, n)$ with the given j-th speaker identity for $j = 1, \ldots, J$.

The source distribution is assumed to be the local complex Gaussian model (LGM) with zero mean and variance of $v_i(f, n)$

$$s_j(f,n) \sim N_{\mathbb{C}}\left(s_j(f,n) \mid 0, v_j(f,n)\right), j = 1, \ldots, J. \tag{8}$$

Under the source independency assumption, we have

$$\boldsymbol{s}(f,n) \sim N_{\mathbb{C}}(\boldsymbol{s}(f,n) \mid 0, \boldsymbol{V}(f,n)) \tag{9}$$

where $\boldsymbol{V}(f, n) = \mathrm{diag}[v_1(f, n), \ldots, v_J(f, n)]$. According to Equation (2) and the condition that $\boldsymbol{A}(f) = \boldsymbol{W}^{-1}(f)$, the distribution of the transformation $\boldsymbol{x}(f, n) = \boldsymbol{W}^{-1}(f)\boldsymbol{s}(f, n)$ is obtained as

$$p[\boldsymbol{x}(f,n)] = |\mathrm{det}W(f)|p[\boldsymbol{s}(f,n)] \tag{10}$$

Hence, the negative log-likelihood of the separation matrices $W = \{\boldsymbol{W}(f)\}_f$ given the observed mixture signals $\mathcal{X} = \{\boldsymbol{x}(f,n)\}_{f,n}$ is represented by

$$\begin{aligned} J(\mathbb{W}, \mathbb{V}) &= -\sum_{f,n} \log p[\boldsymbol{x}(f,n)] \\ &= \sum_{f,n,j} \left[\log v_j(f,n) + \frac{\left|w_j^{\mathrm{H}}(f)\boldsymbol{x}(f,n)\right|^2}{v_j(f,n)} \right] - 2N\sum_f \log|\mathrm{det}\boldsymbol{W}(f)|, \end{aligned} \tag{11}$$

where the last equality is obtained from Equations (9) and (10) by replacing $\boldsymbol{s}_j(f, n)$ with $\boldsymbol{w}_j^{\mathrm{H}}(f)\boldsymbol{x}(f,n)$. Note that we use $\{u_i\}_i$ to denote the set of variable u_i under all the possible values of index i. The variables W and $\mathcal{V}$ can be optimized by minimizing the cost function of Equation (11).

3. Related Work

The MVAE is the state-of-the-art (SOTA) deep learning separation method using the CVAE as the generative source model. The encoder of the CVAE, represented by ϕ, generates the latent space variable $\boldsymbol{z}$ with the conditional distribution $q_\phi(\boldsymbol{z} \mid \widetilde{\boldsymbol{S}}, \boldsymbol{c})$, where the source spectrogram can be described as $\widetilde{\boldsymbol{S}} = \{s(f,n)\}_{f,n}$, and $\boldsymbol{c}$ is the one-hot encoded class label. The decoder of the CVAE, represented by θ, generates the reconstructed source signal

parameters with the conditional distribution $p_\theta(\widetilde{\boldsymbol{S}} \mid \boldsymbol{z},\boldsymbol{c})$. Figure 1 is a schematic diagram of the CVAE. The models ϕ and θ are trained jointly using labeled training samples $\{\widetilde{\boldsymbol{S}},\boldsymbol{c}\}$ with the training objective

$$\mathcal{J}(\phi,\theta) = \mathbb{E}_{(\widetilde{\boldsymbol{S}},\boldsymbol{c})\sim p_D(\widetilde{\boldsymbol{S}},\boldsymbol{c})}\left[\mathbb{E}_{\boldsymbol{z}\sim q(\boldsymbol{z}|\widetilde{\boldsymbol{S}},\boldsymbol{c})}[\log p(\widetilde{\boldsymbol{S}}|\boldsymbol{z},\boldsymbol{c})] - \mathrm{KL}[q(\boldsymbol{z}\left|\widetilde{\boldsymbol{S}},\boldsymbol{c}\right)\|p(\boldsymbol{z})]\right] \tag{12}$$

where $\mathrm{KL}[q\|p]$ denotes the Kullback–Leibler (KL) divergence between q and p, and $p(\boldsymbol{z})$ is assumed to be the standard Gaussian distribution.

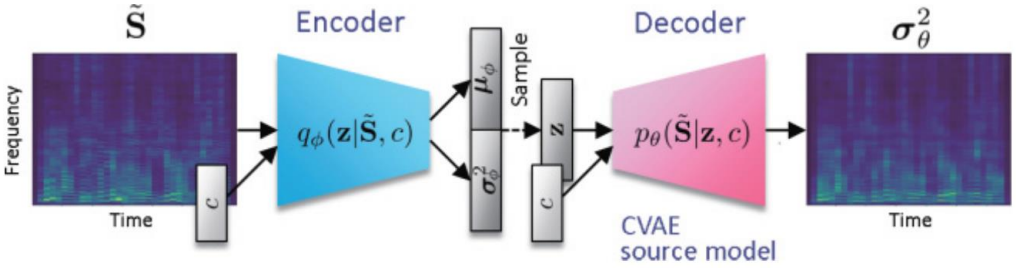

Figure 1. The conditional variational autoencoder model.

During the separation, the trained decoder distribution

$$p_\theta\left(\widetilde{\boldsymbol{S}}|\boldsymbol{z},\boldsymbol{c}\right) = \prod_{f,n} N_{\mathbb{C}}\left(s(f,n)\middle|0,\ \sigma_\theta^2(f,n|\boldsymbol{z},\boldsymbol{c})\right) \tag{13}$$

has the same form as Equation (8), which provides an effective signal model for the MVAE to directly optimize the cost function of Equation (11) with

$$v_j(f,n) = g_j\sigma_\theta^2(f,n|\boldsymbol{z}_j,\boldsymbol{c}_j) \tag{14}$$

where g_j is a global scale parameter to compensate the energy difference between the normalized training data and the test data, and σ_θ^2 is the output of the CVAE generative model. The MVAE iteratively updates the parameters $\{\boldsymbol{z}_j\}_j$ and $\{\boldsymbol{c}_j\}_j$ of the source generative model using backpropagation. The demixing matrix can be updated using iterative projection (IP) rules [11] with

$$\boldsymbol{V}_j(f) = \frac{1}{N}\sum_n \frac{\boldsymbol{x}(f,n)\boldsymbol{x}^{\mathrm{H}}(f,n)}{g_j\sigma_\theta^2(f,n;\boldsymbol{z}_j,\boldsymbol{c}_j)} \tag{15}$$

$$\boldsymbol{w}_j(f) = \left(\boldsymbol{W}(f)\boldsymbol{V}_j(f)\right)^{-1}\boldsymbol{e}_j \tag{16}$$

$$\boldsymbol{w}_j(f) = \boldsymbol{w}_j(f)/\sqrt{\boldsymbol{w}_j^{\mathrm{H}}(f)\boldsymbol{V}_j(f)\boldsymbol{w}_j(f)} \tag{17}$$

and the scale parameter $\{g_j\}_j$ can be updated with

$$g_j = \frac{1}{FN}\sum_{f,n}\frac{|y_j(f,n)|^2}{\sigma_\theta^2(f,n;\boldsymbol{z}_j,\boldsymbol{c}_j)} \tag{18}$$

Equation (18) is obtained by setting the derivative of Equation (11) w.r.t. g equal to 0. This optimization process guarantees the convergence to a stationary point [15].

The latent variable $\boldsymbol{z}$ learned by the CVAE is entangled with the speaker information; hence, the decoder can reconstruct the sources using only the latent variable $\boldsymbol{z}$, regardless of the speaker label $\boldsymbol{c}$, i.e., $q_\phi(\boldsymbol{z} \mid \widetilde{\boldsymbol{S}},\boldsymbol{c}) = q_\phi(\boldsymbol{z} \mid \widetilde{\boldsymbol{S}})$ and $p_\theta(\widetilde{\boldsymbol{S}} \mid \boldsymbol{z},\boldsymbol{c}) = p_\theta(\widetilde{\boldsymbol{S}} \mid \boldsymbol{z})$. This degeneration causes the failure of the source classification (speaker identification) using the speaker labels. Some efforts have been made to learn more disentangled representation, e.g., the fast MVAE (fMVAE) method [22] extended the CVAE model by an additional auxiliary classifier VAE (ACVAE) [23] maximizing the mutual information between $\boldsymbol{c}$ and $\widetilde{\boldsymbol{S}} \sim p_\theta(\widetilde{\boldsymbol{S}} \mid \boldsymbol{z},\boldsymbol{c})$ conditioned

on $\boldsymbol{z}$. This auxiliary classifier forces the decoder outputs to be correlated as far as possible with the class label $\boldsymbol{c}$. The mutual information is expressed as [22]

$$\boldsymbol{I}(\boldsymbol{c}, \widetilde{\boldsymbol{S}} \,|\boldsymbol{z}) = \mathbb{E}_{\boldsymbol{c}\sim p(\boldsymbol{c}),\widetilde{\boldsymbol{S}}\sim p_\theta(\widetilde{\boldsymbol{S}}|\boldsymbol{z},\boldsymbol{c}),\boldsymbol{c}'\sim p(\boldsymbol{c}|\widetilde{\boldsymbol{S}})}\left[\log p\left(\boldsymbol{c}'\middle|\widetilde{\boldsymbol{S}}\right)\right] + H(\boldsymbol{c}) \tag{19}$$

where $H(\boldsymbol{c})$ denotes the entropy of $\boldsymbol{c}$ and can be considered as a constant term. Since the direct optimization of Equation (19) is difficult, the variational lower bound of the first term is introduced using a variational distribution $r_\psi(\boldsymbol{c}\,|\,\widetilde{\boldsymbol{S}})$ for approximating $p(\boldsymbol{c}\,|\,\widetilde{\boldsymbol{S}})$. The lower bound is shown as

$$\mathcal{L}(\phi, \theta, \psi) = \mathbb{E}_{(\widetilde{\boldsymbol{S}},\boldsymbol{c})\sim p_D(\widetilde{\boldsymbol{S}},\boldsymbol{c}),\boldsymbol{z}\sim q_\phi(\boldsymbol{z}|\widetilde{\boldsymbol{S}},\boldsymbol{c})}\left[\mathbb{E}_{\boldsymbol{c}\sim p(\boldsymbol{c}),\widetilde{\boldsymbol{S}}\sim p_\theta(\widetilde{\boldsymbol{S}}|\boldsymbol{z},\boldsymbol{c})}\left[\log r_\psi(\boldsymbol{c}\middle|\widetilde{\boldsymbol{S}})\right]\right] \tag{20}$$

and it is added into the training objective as an information-theoretic regularization term. In addition, the ACVAE also considers the cross-entropy

$$\mathcal{I}(\psi) = \mathbb{E}_{(\widetilde{\boldsymbol{S}},\boldsymbol{c})\sim p_D(\widetilde{\boldsymbol{S}},\boldsymbol{c})}\left[\log r_\psi(\boldsymbol{c}\middle|\widetilde{\boldsymbol{S}})\right] \tag{21}$$

Finally, the entire training criterion combines the three terms in Equations (12), (20) and (21), shown as

$$\mathcal{J}(\phi, \theta) + \lambda_{\mathcal{L}}\mathcal{L}(\phi, \theta, \psi) + \lambda_{\mathcal{I}}\mathcal{I}(\psi) \tag{22}$$

where $\lambda_{\mathcal{L}}$ and $\lambda_{\mathcal{I}}$ are the weight factors. Maximizing Equation (22) w.r.t. ϕ, θ and ψ, the fMVAE can prompt the encoder to learn a more disentangled representation from the speaker label and, thus, promote the identification accuracy. However, the disentanglement is not complete so that the identification accuracy is still limited. According to the reported results in [22], the accuracy rate of the fMVAE is 80.00%. Moreover, its separation performance degrades considerably in high reverberant environments.

In the next section, we extend the MVAE to learn the completely disentangled representations in an unsupervised framework.

4. Proposed Method

4.1. The Proposed VAE Model

The training of the original MVAE model in Figure 1 is supervised using speaker label $\boldsymbol{c}$, and exposed to the problem of degeneration. To make the identification more reliable, we replaced the encoder network of the MVAE with two dedicated encoders for separately learning the speaker and content information. It is motivated by the strategy used in the VAE-based voice conversion (VC) [26]. The proposed model consists of three components as shown in Figure 2: a speaker encoder $\mathrm{E}^{(\mathrm{s})}$ to encode the speaker representations $\boldsymbol{z}^{(\mathrm{s})}$, a content encoder $\mathrm{E}^{(\mathrm{c})}$ to encode the content representations $\boldsymbol{z}^{(\mathrm{c})}$, and a decoder D. The instance normalization (IN) is applied to $\mathrm{E}^{(\mathrm{c})}$ to disentangle $\boldsymbol{z}^{(\mathrm{c})}$ from $\boldsymbol{z}^{(\mathrm{s})}$. The decoder reconstructs the signal spectrograms from the combination of $\boldsymbol{z}^{(\mathrm{s})}$ and $\boldsymbol{z}^{(\mathrm{c})}$ using the adaptive instance normalization (adaIN). Note that this model can learn meaningful speaker representations without any supervision (speaker labels), thus, it is feasible to generalize to unseen speakers.

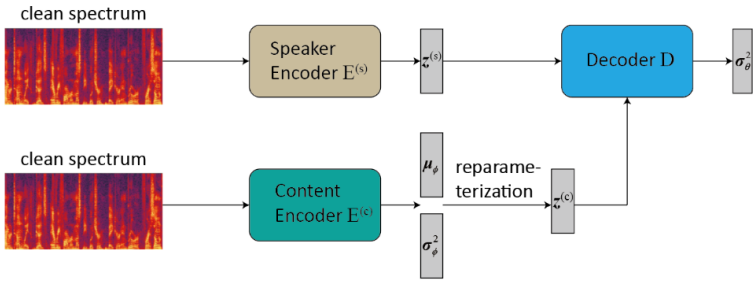

Figure 2. The proposed variational autoencoder model.

We denote the network parameters of D, $E^{(c)}$, and $E^{(s)}$ as θ, ϕ, and ψ, respectively, which are trained using the objective

$$L(\theta, \phi, \psi) = \lambda_{rec} L_{rec} + \lambda_{kl} L_{kl} \tag{23}$$

where

$$L_{rec} = -\mathbb{E}_{\widetilde{\mathbf{S}} \sim p(\widetilde{\mathbf{S}})} \left[\mathbb{E}_{\boldsymbol{z}^{(c)} \sim q_{\phi}(\boldsymbol{z}^{(c)} | \widetilde{\mathbf{S}})} \left[\log p_{\theta} \left(\widetilde{\mathbf{S}} \middle| \boldsymbol{z}^{(c)}, \boldsymbol{z}^{(s)} \right) \right] \right] \tag{24}$$

is the reconstruction error term measuring the difference between the generated data and real data, and

$$L_{kl} = \mathbb{E}_{\widetilde{\mathbf{S}} \sim p(\widetilde{\mathbf{S}})} \mathrm{KL} \left[q_{\phi} \left(\boldsymbol{z}^{(c)} \middle| \widetilde{\mathbf{S}} \right) \| p \left(\boldsymbol{z}^{(c)} \right) \right] \tag{25}$$

is the Kullback–Leibler divergence term that forces each element of the encoder output to be independent and normally distributed. λ_{rec} and λ_{kl} are the weight factors, $q_{\phi}(\boldsymbol{z}^{(c)} \mid \widetilde{\mathbf{S}})$ is the posterior distribution of $\boldsymbol{z}^{(c)}$ learned by the content encoder, $p(\boldsymbol{z}^{(c)})$ is the prior standard Gaussian distribution, and $p_{\theta}(\widetilde{\mathbf{S}} \mid \boldsymbol{z}^{(c)}, \boldsymbol{z}^{(s)})$ is the conditional distribution of the reconstructed spectrograms learned by the decoder. Once the parameters θ, ϕ, and ψ are optimized, they are fixed during the inference (separation) phase and then, the trained decoder distribution $p_{\theta}(\widetilde{\mathbf{S}} \mid \boldsymbol{z}^{(c)}, \boldsymbol{z}^{(s)})$ is used as the generative model to generate spectrograms of all the sources.

4.2. Instance Normalization

In order to learn the disentangled representations, the instance normalization is applied after each convolution layer of $E^{(c)}$ to remove the global information such as the speaker information, formulated by

$$M_c'[w] = \frac{M_c[w] - \mu_c}{\sigma_c} \tag{26}$$

where

$$\mu_c = \frac{1}{W} \sum_{w=1}^{W} M_c[w] \tag{27}$$

$$\sigma_c = \sqrt{\frac{1}{W} \sum_{w=1}^{W} (M_c[w] - \bar{}_c)^2 + "} \tag{28}$$

M_c is the c-th channel feature map of the output of the previous convolutional layer, $M_c[w]$ denotes the w-th element of M_c, μ_c and σ_c represent the mean and standard variation of the c-th channel, and ε is a small value to guarantee numerical stability. The IN layers remove the global information from $\boldsymbol{z}^{(c)}$ so that only the linguistic information is preserved. Meanwhile, the global pooling is applied after the last convolution layer of $E^{(s)}$ to enforce $\boldsymbol{z}^{(s)}$ to contain the global information only. Later in the decoder, $\boldsymbol{z}^{(c)}$ is combined with $\boldsymbol{z}^{(s)}$ by adaptive instance normalization (adaIN) as formulated by

$$M_c'[w] = \gamma_c \frac{M_c[w] - \mu_c}{\sigma_c} + \beta_c, \tag{29}$$

where M_c is the c-th channel feature map learned by the previous convolutional layer with input $\boldsymbol{z}^{(c)}$, μ_c and σ_c are calculated by Equations (27) and (28), β_c and γ_c are the c-th channel mean and standard variation of $\boldsymbol{z}^{(s)}$ learned by the Dense layers. The network architecture is detailed in Figure 3.

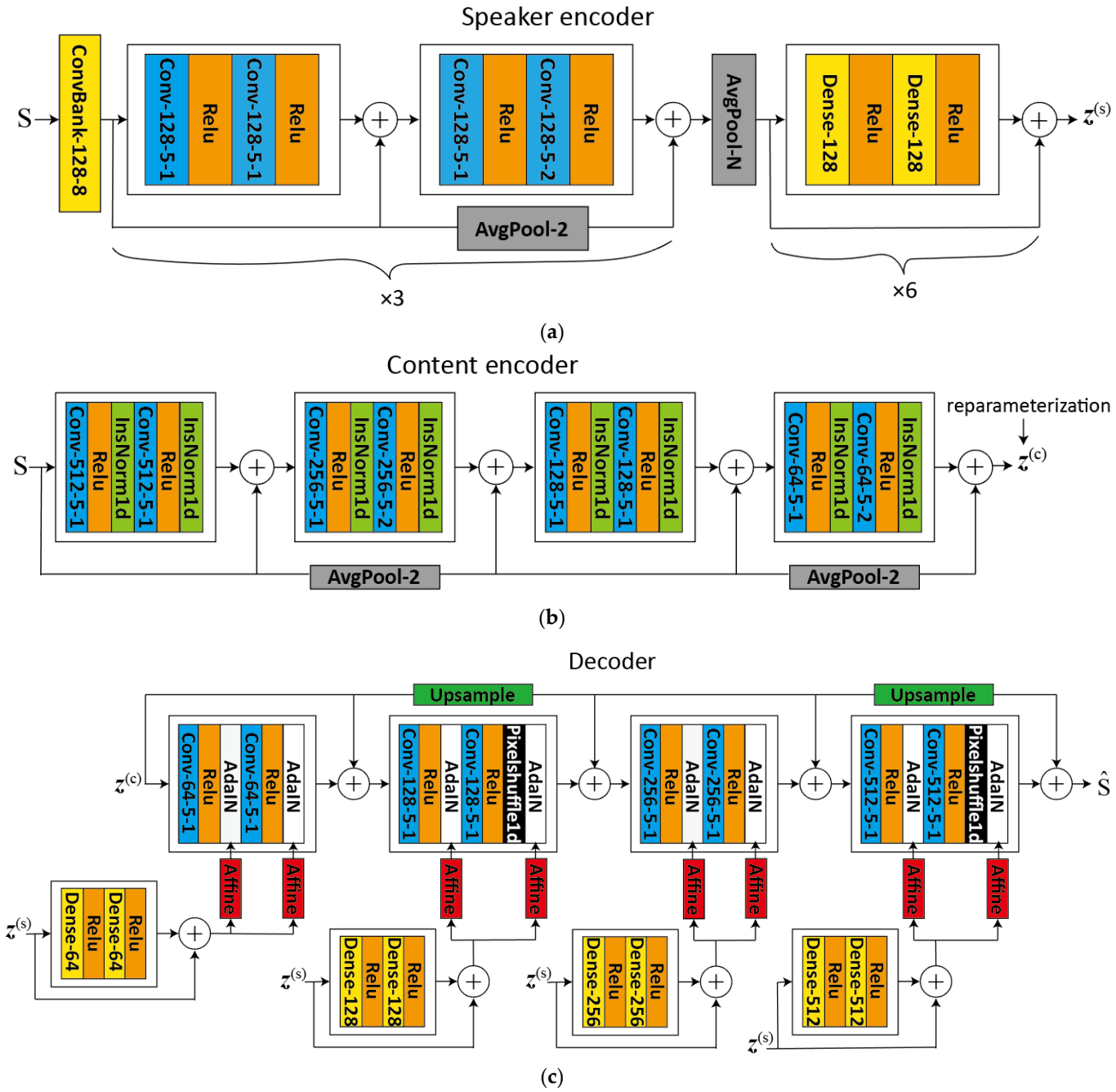

Figure 3. The network architecture: (**a**) Speaker encoder; (**b**) Content encoder; (**c**) Decoder. The channel number and the bank number are labeled in ConvBank layers, the channel number is labeled in each Dense layer, the channel number, kernel size, and stride are labeled in each convolution layer.

4.3. Output in Designated Speaker Order

Using the enrolled utterances of the target speaker(s), we can infer the speaker representations with the pretrained $\mathrm{E}^{(\mathrm{s})}$ and consider it as a sequence denoted by $\{\tilde{z}_j^{(s)}\}$. Note that we use $\{u_i\}$ to denote the sequence where the set of variable u_i is arranged in a specific order, distinguished from the notation $\{u_i\}_i$ defined in Section 2. It is straightforward to preassign $\{z_j^{(s)}\}$ by $\{\tilde{z}_j^{(s)}\}$ to determine the order of the outputs $\{y_j\}$. However, this scheme is sensitive to initialization and likely to be trapped in poor local optima when the source index preassigned to each network output v_j mismatches the source to which $\boldsymbol{w}_j$ corresponds [21]. To mitigate this problem, we initialized the demixing matrices (denoted by $\boldsymbol{W}^{\mathrm{ini}}$) using the

ILRMA run for 30 iterations in the same way as the original MVAE algorithm. The initial speaker representation denoted by $\{\hat{\boldsymbol{z}}_j^{(s)}\}$ is inferred from $\boldsymbol{y}^{\text{ini}}$ using $\boldsymbol{y}^{\text{ini}} = \boldsymbol{W}^{\text{ini}}\boldsymbol{x}$ and $\hat{\boldsymbol{z}}_j^{(s)} = \mathrm{E}^{(\mathrm{s})}(y_j^{ini})$ for each j. To avoid the mismatch between v_j and $\boldsymbol{w}_j$, we rearrange $\{\widetilde{\boldsymbol{z}}_j^{(s)}\}$ to align with the initial $\{\hat{\boldsymbol{z}}_j^{(s)}\}$. Then, the variable $\boldsymbol{z}^{(\mathrm{s})}$ is assigned by the rearranged $\{\widetilde{\boldsymbol{z}}_j^{(s)}\}$ and held fixed throughout the separation process while the other variables, $\boldsymbol{z}^{(\mathrm{c})}$, $\boldsymbol{W}$, and $\boldsymbol{g}$, are to be updated at each iteration. Compared with the original MVAE, less variables need to be optimized in the IN-MVAE. Moreover, the assignment of the speaker representation is an effective way of utilizing the prior information embedded in the enrolled utterances, which works on not only identification but also separation.

For the alignment of $\{\widetilde{\boldsymbol{z}}_j^{(s)}\}$ with $\{\hat{\boldsymbol{z}}_j^{(s)}\}$, we minimize the distance between the reference sequence $\{\hat{\boldsymbol{z}}_j^{(s)}\}$ and the sequence $\{\widetilde{\boldsymbol{z}}_j^{(s)}\}$ among all its possible permutations. Note that in Equation (29), $\boldsymbol{z}^{(\mathrm{s})}$ delivers the speaker information to the decoder in the form of $\boldsymbol{\beta}$, so the distance measure is defined as

$$L_\beta = \sum_{j=1}^{J}\sum_{c=1}^{C_\beta}\left(\widetilde{\beta}_{j,c} - \hat{\beta}_{j,c}\right)^2 \tag{30}$$

where $\{\widetilde{\beta}\}$ and $\{\hat{\beta}_j\}$ are the outputs of the Dense layers in the decoder with the inputs of $\{\widetilde{\boldsymbol{z}}_j^{(s)}\}$ and $\{\hat{\boldsymbol{z}}_j^{(s)}\}$, respectively, the subscript c represents the c-th channel, and the total channel number is C_β.

After separation, the IN-MVAE rearranges the final speaker representation $\{\hat{\boldsymbol{z}}_j^{(s)}\}$, where $\hat{\boldsymbol{z}}_j^{(s)} = \mathrm{E}^{(\mathrm{s})}(y_j)$. This corresponds to the rearrangement of the outputs $\{y_j\}$ to guarantee the designated order, in case that the final $\{\hat{\boldsymbol{z}}_j^{(s)}\}$ mismatches the preassigned $\{\widetilde{\boldsymbol{z}}_j^{(s)}\}$. Note that the distance measure of Equation (30) is symmetric; hence, we can use again the above permutation alignment scheme, but here the reference sequence is $\{\widetilde{\boldsymbol{z}}_j^{(s)}\}$ and the permutation of $\{\hat{\boldsymbol{z}}_j^{(s)}\}$ is to be optimized.

The processing flow of the IN-MVAE is summarized in Algorithm 1.

Algorithm 1: IN-MVAE

Train $\mathrm{E}^{(\mathrm{s})}$, $\mathrm{E}^{(\mathrm{c})}$, and D with (23)
Initialize g with all elements being 1 and $\boldsymbol{W}$ using the ILRMA run for 30 iterations
Calculate the initially separated signal $\boldsymbol{y}^{\text{ini}} = \boldsymbol{W}^{\text{ini}}\boldsymbol{x}$ and initialize $\boldsymbol{z}^{(\mathrm{c})}$ with content encoder outputs using $\boldsymbol{z}^{(\mathrm{c})} = \mathrm{E}^{(\mathrm{c})}(\boldsymbol{y}^{\text{ini}})$
Assign $\boldsymbol{z}^{(\mathrm{s})}$ by the rearranged $\{\widetilde{\boldsymbol{z}}_j^{(s)}\}$ using the proposed permutation alignment scheme
Repeat
Update $\boldsymbol{z}^{(\mathrm{c})}$ with (11) using backpropagation for a fixed number of iterations
Update $\boldsymbol{W}$ and $\boldsymbol{g}$ using (15)–(18)
until convergence
Calculate the final separated signal $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x}$
Rearrange $\{y_j\}$ in designated speaker order using the proposed permutation alignment scheme

5. Experiments

5.1. Simulation Environment

Utterances of 100 speakers (45 males and 55 females) are excerpted from the Librispeech database [27]. The training utterances are obtained from the train-clean-100 and train-clean-360 subsets and split into 30,000 segments of 3.2 s, and the total duration is 27 h. These segments are preprocessed by trimming out the silence and normalizing the volume, and then transformed into a logarithmic Mel spectrogram as the input features. Parameter settings are listed in Table 1. In the inference phase, the other utterances of these 100 speakers are used for speaker-closed evaluation and the utterances in the test-clean and

dev-clean subsets for speaker-open evaluation. All of them are truncated into segments of 5~30 s. Two randomly selected segments of different speakers convolved with the room impulse response (RIR) recordings are mixed and captured by two microphones, where the RIR recordings are excerpted from the MIRD database [28] and down-sampled to 16 kHz. Here, we choose the 4-th and 5-th channel recordings and the distance of 1 m from both sources to the array center. The subtending angles of sources are randomly chosen from {90°, 105°, 120°}. The code of all the experiments was implemented in Python using a PyTorch framework, which can be found in the Supplementary Materials. The results are evaluated by signal-to-distortion ratio (SDR) [29] and signal-to-interference ratio (SIR) [29] using the bss_eval_sources function of a mir_eval package, perceptual evaluation of speech quality (PESQ) [30] using the pesq function of a pypesq package, and short-time objective intelligibility (STOI) [31] using the stoi function of a pystoi package. Their improvement (ΔSDR, ΔSIR, ΔPESQ, ΔSTOI) with respect to the unprocessed signal is calculated and averaged over 80 different trials. Note that the performance metrics are all evaluated on the basis of the oracle permutation, i.e., the permutation of $\{y_j\}$ that maximizes the SDR, SIR, PESQ, and STOI scores. The permutation of $\{y_j\}$ when calculating these objective scores is not necessarily the same as the output permutation, since we tend to evaluate the separation performance and the sorting accuracy independently. The SIR of the unprocessed signal ranges from −5 dB to 5 dB. In all the experiments, the demixing matrices of the AuxIVA and ILRMA algorithms are initialized with identity matrices, and the MVAE and IN-MVAE algorithms are initialized using the ILRMA run for 30 iterations.

Table 1. Parameter settings.

Description	Parameter
Number of Mel basis	512
Sampling rate	16 kHz
STFT length	128 ms
Hop length	32 ms
Analysis window	Hanning
Optimizer	Adam
Learning rate	0.0001
Weight in loss	$\lambda_{\mathrm{rec}} = 10$, $\lambda_{\mathrm{kl}} = 1$

5.2. Performance Comparison with Existing Methods

In the first experiment, we compare the proposed method with three baseline approaches, i.e., AuxIVA [11], ILRMA [9] with the base number of two, and the MVAE with the same architecture as detailed in [15]. AuxIVA is a well-studied algorithm with wide popularity, ILRMA is the SOTA rule-based method without training, and the MVAE is the SOTA-hybrid method combining the merits of IVA and deep generative models. Since the baseline approaches are incapable of sorting the outputs, we cascade a speaker recognition (SR) system after the separation module. In the SR system, x-vectors are used for capturing speaker characteristics and we use the same network structure as described in [32]. The x-vectors of the separated signals are centered, and the dimensionality is reduced to 128 using linear discriminant analysis (LDA). After that, the embeddings are length normalized and a probabilistic linear discriminant analysis (PLDA) [33] backend is trained using the Bob toolbox [34] to score the similarity between the separated signal and the enrolled utterance. We identify the separated signal with the speaker whose enrolled utterance has the highest similarity with it and, finally, rearrange the separated signals in the designated speaker order. In this experiment, the middle reverberation time of 360 ms is selected and all the speech mixtures are the combinations of female and male speakers. The results are shown in Table 2. It can be seen that the IN-MVAE outperforms those baseline methods in terms of all the objective metrics and achieves 100% sorting accuracy. The cascaded x-vector system used for sorting the output of AuxIVA, ILRMA and MVAE cannot integrate the prior information of the enrolled utterances into the separation process. In contrast,

two complementary (content and speaker) encoders used in the IN-MVAE allow us to estimate the speaker identity simultaneously with source separation; thus, the prior speaker information can be also used for the guidance of separation. The frequency-domain separation algorithms need to align the frequency components of the same source. Although IVA is proposed to solve this problem by exploiting the inter-frequency dependence, it still suffers from the block permutation (BP) problems [35]. In contrast to the output (or global, or external) permutation ambiguity, which is the focus of this paper, the block permutation problem can be classified into the internal permutation problem, caused by the misalignment of the frequency blocks from different sources. The BP problems will severely deteriorate the separation performance and also limit the sorting accuracy, since the interfering source components from other speakers will mislead the inference of the speaker identity. Fortunately, in our proposed method, the preassigned $\{\tilde{\mathbf{z}}_j^{(s)}\}$ introduces useful global information that assists the correct distribution of the separated frequency slot to its corresponding speaker, thus effectively alleviating the block permutation (BP) problems. As shown in Table 2, all three baseline methods suffer from BP problems while the IN-MVAE shows zero BP problems in this test.

Table 2. The comparison of various algorithms under the reverberation time of 360 ms with the female and male combination.

Algorithm	ΔSIR (dB)	ΔSDR (dB)	ΔPESQ	ΔSTOI	Number of BP Problems	Sorting Accuracy
AuxIVA	10.19	5.77	0.50	0.09	16	0.88
ILRMA	14.51	7.28	0.60	0.11	16	0.94
MVAE	20.35	7.84	0.76	0.12	7	0.98
IN-MVAE	22.41	9.42	0.79	0.13	0	1

5.3. Performance Evaluation with Different Gender Combinations

We next test the performance of the IN-MVAE with different gender combinations, i.e., female and male (FM), female and female (FF), male and male (MM). The experimental results under the reverberation time of 360 ms are listed in Table 3. It can be found that the separation performances are all at a satisfactory level and the perfect sorting can be obtained no matter what the gender combination is. Note that the BP problem occurs once when the speakers are the combination of females.

Table 3. The performance of the IN-MVAE with different gender combinations under the reverberation time of 360 ms.

Combination Mode	ΔSIR (dB)	ΔSDR (dB)	ΔPESQ	ΔSTOI	Number of BP Problems	Sorting Accuracy
FM	22.41	9.42	0.79	0.13	0	1
FF	21.83	9.41	0.84	0.14	1	1
MM	21.76	9.03	0.77	0.14	0	1

5.4. Performance Evaluation under Different Reverberation Times

To further assess the performance of the proposed method in various reverberant conditions, we carry out experiments under three different reverberation times of 160 ms, 360 ms, and 610 ms, respectively, and the results are presented in Table 4. It can be seen that higher reverberations lead to deteriorated separation performance, but the IN-MVAE still achieves over 20 dB increments of SIR under the 610 ms reverberation time. The slight decline of the sorting accuracy under the 160 ms reverberation time is caused by the BP problems. Eradication of the BP problems is not a straightforward task and out of the focus of this paper.

Table 4. The performance of the IN-MVAE under different reverberation times with the female and male combination.

Reverberation Time (ms)	ΔSIR (dB)	ΔSDR (dB)	ΔPESQ	ΔSTOI	Number of BP Problems	Sorting Accuracy
160	24.78	10.83	1.30	0.16	2	0.98
360	22.41	9.42	0.79	0.13	0	1
610	20.20	8.17	0.56	0.11	0	1

5.5. Generalization Performance

All the above experiments are speaker-closed evaluations. In the final experiment, we test the generalization ability of the proposed method with utterances from unseen speakers during the training. The speaker-open results under the reverberation time of 360 ms with random gender combinations shown in Table 5 verify the good generalization capability of the IN-MVAE.

Table 5. The performance of the IN-MVAE in speaker-open conditions under the reverberation time of 360 ms with random gender combination.

ΔSIR (dB)	ΔSDR (dB)	ΔPESQ	ΔSTOI	Number of BP Problems	Sorting Accuracy
21.90	9.25	0.80	0.14	0	1

6. Discussion and Conclusions

We propose the unsupervised IN-MVAE method, a variant of the MVAE, that accomplishes source separation and speaker recognition in a unified framework. We adopt the convergence-guaranteed optimization algorithm of the MVAE and apply the instance normalization for feature disentanglement. A simple but reliable permutation alignment scheme is proposed for the output sorting. The proposed method can more effectively exploit the prior speaker information, which benefits not only the sorting of speakers but also the alleviation of the BP problems. Our experimental results reveal that the IN-MVAE can obtain a significantly higher sorting accuracy and good separation performance with robustness to different gender combinations and various reverberant conditions, achieving over a 2 dB advantage over the MVAE in terms of the SIR metric.

It should be noted that the speaker identification confidence would degrade with short utterances due to the lack of information, which forms an obstacle for the real-time implementation of the proposed method. Solving this problem is not an easy task and we will dedicate to it in a future study.

Supplementary Materials: A Python implementation of our algorithm can be downloaded at: https://github.com/LiaoLele/IN-MVAE (accessed on 24 November 2022).

Author Contributions: Conceptualization, L.L. and J.L.; methodology, L.L. and J.L.; software, L.L.; validation, L.L., G.C., H.R., K.C. and J.L.; investigation, L.L., G.C. and H.R.; resources, J.L. and K.C.; data curation, L.L.; writing—original draft preparation, L.L.; writing—review and editing, L.L., G.C., H.R., K.C. and J.L.; supervision, J.L.; project administration, J.L. and K.C.; funding acquisition, J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant numbers 11874219 and 12274221.

Data Availability Statement: The Librispeech database is available at: http://www.openslr.org/12/ (accessed on 24 November 2022) and the MIRD database is available at: https://www.eng.biu.ac.il/~gannot/RIR_DATABASE/ (accessed on 24 November 2022).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Makino, S.; Lee, T.-W.; Sawada, H. *Blind Speech Separation*; Springer: Dordrecht, The Netherlands, 2007; ISBN 9781402064791.
2. Hyvärinen, A.; Hoyer, P. Emergence of Phase- and Shift-Invariant Features by Decomposition of Natural Images into Independent Feature Subspaces. *Neural Comput.* **2000**, *12*, 1705–1720. [CrossRef] [PubMed]
3. Lee, I.; Hao, J.; Lee, T.-W. Adaptive Independent Vector Analysis for the Separation of Convoluted Mixtures Using EM Algorithm. In Proceedings of the 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, Las Vegas, NV, USA, 31 March–4 April 2008; pp. 145–148.
4. Hao, J.; Lee, I.; Lee, T.-W.; Sejnowski, T.J. Independent Vector Analysis for Source Separation Using a Mixture of Gaussians Prior. *Neural Comput.* **2010**, *22*, 1646–1673. [CrossRef] [PubMed]
5. Gu, Z.; Lu, J.; Chen, K. Speech Separation Using Independent Vector Analysis with an Amplitude Variable Gaussian Mixture Model. In Proceedings of the INTERSPEECH 2019, Graz, Austria, 15–19 September 2019; pp. 1358–1362.
6. Anderson, M.; Adali, T.; Li, X.-L. Joint Blind Source Separation with Multivariate Gaussian Model: Algorithms and Performance Analysis. *IEEE Trans. Signal Process.* **2012**, *60*, 1672–1683. [CrossRef]
7. Liang, Y.; Naqvi, S.M.; Wang, W.; Chambers, J.A. Frequency Domain Blind Source Separation Based on Independent Vector Analysis with a Multivariate Generalized Gaussian Source Prior. In *Blind Source Separation: Advances in Theory, Algorithms and Applications*; Naik, G.R., Wang, W., Eds.; Signals and Communication Technology; Springer: Berlin, Heidelberg, 2014; pp. 131–150. ISBN 9783642550164.
8. Khan, J.B.; Jan, T.; Khalil, R.A.; Altalbe, A. Hybrid Source Prior Based Independent Vector Analysis for Blind Separation of Speech Signals. *IEEE Access* **2020**, *8*, 132871–132881. [CrossRef]
9. Kitamura, D.; Ono, N.; Sawada, H.; Kameoka, H.; Saruwatari, H. Determined Blind Source Separation Unifying Independent Vector Analysis and Nonnegative Matrix Factorization. *IEEE ACM Trans. Audio Speech Lang. Process.* **2016**, *24*, 1626–1641. [CrossRef]
10. Sawada, H.; Ono, N.; Kameoka, H.; Kitamura, D.; Saruwatari, H. A Review of Blind Source Separation Methods: Two Converging Routes to ILRMA Originating from ICA and NMF. *APSIPA Trans. Signal Inf. Process.* **2019**, *8*, e12. [CrossRef]
11. Ono, N. Stable and Fast Update Rules for Independent Vector Analysis Based on Auxiliary Function Technique. In Proceedings of the 2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, 16–19 October 2011; pp. 189–192.
12. Lee, D.D.; Seung, H.S. Learning the Parts of Objects by Non-Negative Matrix Factorization. *Nature* **1999**, *401*, 788–791. [CrossRef] [PubMed]
13. Lee, D.; Seung, H.S. Algorithms for Non-Negative Matrix Factorization. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2001; Volume 13.
14. Xie, Y.; Xie, K.; Yang, J.; Xie, S. Underdetermined Blind Source Separation Combining Tensor Decomposition and Nonnegative Matrix Factorization. *Symmetry* **2018**, *10*, 521. [CrossRef]
15. Kameoka, H.; Li, L.; Inoue, S.; Makino, S. Supervised Determined Source Separation with Multichannel Variational Autoencoder. *Neural Comput.* **2019**, *31*, 1891–1914. [CrossRef] [PubMed]
16. Mogami, S.; Sumino, H.; Kitamura, D.; Takamune, N.; Takamichi, S.; Saruwatari, H.; Ono, N. Independent Deeply Learned Matrix Analysis for Multichannel Audio Source Separation. In Proceedings of the 2018 26th European Signal Processing Conference (EUSIPCO), Rome, Italy, 3–7 September 2018; pp. 1557–1561.
17. Makishima, N.; Mogami, S.; Takamune, N.; Kitamura, D.; Sumino, H.; Takamichi, S.; Saruwatari, H.; Ono, N. Independent Deeply Learned Matrix Analysis for Determined Audio Source Separation. *IEEE ACM Trans. Audio Speech Lang. Process.* **2019**, *27*, 1601–1615. [CrossRef]
18. Wang, D.; Chen, J. Supervised Speech Separation Based on Deep Learning: An Overview. *IEEE ACM Trans Audio Speech Lang Process.* **2018**, *26*, 1702–1726. [CrossRef]
19. Doersch, C. Tutorial on Variational Autoencoders. *arXiv* **2021**, arXiv:1606.05908.
20. Hyvärinen, A.; Karhunen, J.; Oja, E. *Independent Component Analysis*; Wiley: New York, USA, 2001; ISBN 9780471405405.
21. Seki, S.; Kameoka, H.; Li, L.; Toda, T.; Takeda, K. Underdetermined Source Separation Based on Generalized Multichannel Variational Autoencoder. *IEEE Access* **2019**, *7*, 168104–168115. [CrossRef]
22. Li, L.; Kameoka, H.; Makino, S. Fast MVAE: Joint Separation and Classification of Mixed Sources Based on Multichannel Variational Autoencoder with Auxiliary Classifier. In Proceedings of the ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 546–550.
23. Kameoka, H.; Kaneko, T.; Tanaka, K.; Hojo, N. ACVAE-VC: Non-Parallel Many-to-Many Voice Conversion with Auxiliary Classifier Variational Autoencoder. *arXiv* **2020**, arXiv:1808.05092. [CrossRef]
24. Ulyanov, D.; Vedaldi, A.; Lempitsky, V. Instance Normalization: The Missing Ingredient for Fast Stylization. *arXiv* **2017**, arXiv:1607.08022.
25. Wang, Q.; Zhang, Y.; Yin, S.; Wang, Y.; Wu, G. A Novel Underdetermined Blind Source Separation Method Based on OPTICS and Subspace Projection. *Symmetry* **2021**, *13*, 1677. [CrossRef]
26. Chou, J.; Yeh, C.; Lee, H. One-Shot Voice Conversion by Separating Speaker and Content Representations with Instance Normalization. *arXiv* **2019**, arXiv:1904.05742.

27. Panayotov, V.; Chen, G.; Povey, D.; Khudanpur, S. Librispeech: An ASR Corpus Based on Public Domain Audio Books. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, Australia, 19–24 April 2015; pp. 5206–5210.
28. Hadad, E.; Heese, F.; Vary, P.; Gannot, S. Multichannel Audio Database in Various Acoustic Environments. In Proceedings of the 2014 14th International Workshop on Acoustic Signal Enhancement (IWAENC), Juan-les-Pins, France, 8–11 September 2014; pp. 313–317.
29. Vincent, E.; Gribonval, R.; Fevotte, C. Performance Measurement in Blind Audio Source Separation. *IEEE Trans. Audio Speech Lang. Process.* **2006**, *14*, 1462–1469. [CrossRef]
30. Rix, A.W.; Beerends, J.G.; Hollier, M.P.; Hekstra, A.P. Perceptual Evaluation of Speech Quality (PESQ)-a New Method for Speech Quality Assessment of Telephone Networks and Codecs. In Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221), Salt Lake City, UT, USA, 7–11 May 2001; Volume 2, pp. 749–752.
31. Taal, C.H.; Hendriks, R.C.; Heusdens, R.; Jensen, J. A Short-Time Objective Intelligibility Measure for Time-Frequency Weighted Noisy Speech. In Proceedings of the 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, Dallas, TX, USA, 14–19 March 2010; pp. 4214–4217.
32. Snyder, D.; Garcia-Romero, D.; Sell, G.; Povey, D.; Khudanpur, S. X-Vectors: Robust Dnn Embeddings for Speaker Recognition. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; IEEE: New York, NY, USA, 2018; pp. 5329–5333.
33. Prince, S.J.; Elder, J.H. Probabilistic Linear Discriminant Analysis for Inferences about Identity. In Proceedings of the 2007 IEEE 11th International Conference on Computer Vision, Rio de Janeiro, Brazil, 26 December 2007; IEEE: New York, NY, USA, 2007; pp. 1–8.
34. Anjos, A.; El-Shafey, L.; Wallace, R.; Günther, M.; McCool, C.; Marcel, S. Bob: A Free Signal Processing and Machine Learning Toolbox for Researchers. In Proceedings of the 20th ACM International Conference on Multimedia, Nara, Japan, 29 October 2012; pp. 1449–1452.
35. Liang, Y.; Naqvi, S.M.; Chambers, J. Overcoming Block Permutation Problem in Frequency Domain Blind Source Separation When Using AuxIVA Algorithm. *Electron. Lett.* **2012**, *48*, 460–462. [CrossRef]