2020 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 21–24, 2020, ESPOO, FINLAND

DETERMINED AUDIO SOURCE SEPARATION WITH MULTICHANNEL STAR GENERATIVE ADVERSARIAL NETWORK

Li Li[1], *Hirokazu Kameoka*[2], *Shoji Makino*[1]

[1] University of Tsukuba, Japan
[2] NTT Communication Science Laboratories, NTT Corporation, Japan

ABSTRACT

This paper proposes a multichannel source separation approach, which uses a star generative adversarial network (StarGAN) to model power spectrograms of sources. Various studies have shown the significant contributions of a precise source model to the performance improvement in audio source separation, which indicates the importance of developing a better source model. In this paper, we explore the potential of StarGAN for modeling source spectrograms and investigate the effectiveness of the StarGAN source model in determined multichannel source separation by incorporating it into a frequency-domain independent component analysis (ICA) framework. The experimental results reveal that the proposed StarGAN-based method outperformed conventional methods that use non-negative matrix factorization (NMF) or a variational autoencoder (VAE) for source spectrogram modeling.

***Index Terms*—** Multichannel audio signal processing, determined source separation, star generative adversarial network (StarGAN), spectrogram modeling, deep generative model

1. INTRODUCTION

The aim of blind source separation (BSS) [1] is to separate individual source signals from microphone array inputs without prior information about the sources and the mixing methodology. The frequency-domain BSS approach is usually preferred since it allows a fast implementation based on the instantaneous mixture assumption and provides the flexibility of utilizing various models for the time-frequency representations of source signals, e.g., spectrograms.

For example, independent vector analysis (IVA) [2, 3] models power spectrograms of sources as a single flat-shaped spectral basis scaled by time-varying amplitudes based on the assumption that the magnitudes of the frequency components originating from the same source tend to vary coherently over time. Multichannel extensions of non-negative matrix factorization (NMF), e.g., multichannel NMF (MNMF) [4, 5] and independent low-rank matrix analysis (ILRMA) [6, 7], incorporate the NMF concept into the source model to capture spectral structures of sources. Although these methods work reasonably well in most cases, they can fail to separate sources that do not satisfy the low-rank assumption. Motivated by the strong power of deep generative models, including variational autoencoders (VAEs) [8] and generative adversarial networks (GANs) [9], to learn data distributions, some attempts have recently been made to apply these models to speech enhancement and source separation tasks [10–15]. The multichannel variational autoencoder (MVAE) [12] is one such method, where a conditional VAE (CVAE) [8] is trained using the spectrograms of clean speech samples along with the corresponding speaker identity as an auxiliary label input so that the trained decoder distribution can be used as a universal generative model of source signals. At the separation phase, the trained decoder is applied to estimate the spectrograms of sources in a mixture. MVAE and its generalized version, GMVAE [16], have been demonstrated to significantly outperform ILRMA and MNMF, which confirms the effectiveness of incorporating a more precise source model in improving the source separation performance.

Compared to VAE, which explicitly assumes the prior distribution about the data, e.g., a Gaussian distribution, and learns data distribution by forcing an approximate posterior distributions to become consistent with the true one, GAN trains a generator network to deceive a real/fake discriminator network so that the generator distribution is optimized to fit the target distribution without explicit density estimation. This allows us to avoid the mismatch between the assumed and real distributions and the approximation error occurring in the posterior estimation. Thanks to this learning strategy, it is expected that GAN can learn a data distribution more accurately than VAE.

To take advantage of GAN, this paper proposes a determined multichannel source separation method that employs StarGAN [17] to learn the generative distribution of power spectrograms of sources. StarGAN is a GAN variant consisting of a generator, discriminator, and domain classifier. StarGAN, which was originally proposed for multi-domain translation, has recently been adapted for use in many-to-many voice conversion [18] and shown to perform remarkably. In addition, the following three benefits motivate us to

This work was supported by JSPS KAKENHI 18J20059 and 19H04131, and JST CREST JPMJCR19A3.

978-1-7281-6662-9/20/$31.00 ©2020 IEEE

adopt StarGAN in the proposed method. First, by using the domain classifier to measure how likely the generated spectrogram is to belong to the corresponding speaker, the model learns to avoid ignoring the class index input when generating spectrograms. This is in contrast to a regular conditional GAN, which is free to ignore the class index input when the networks have sufficient capacity. Second, it is expected that training a model in a conversion manner can promote the disentanglement between the latent representation and the class index, which makes the latent representation more meaningful. Third, the network composition involving a domain classifier makes it possible to apply the fast algorithm implemented in FastMVAE [19]. We evaluate the effectiveness of StarGAN in modeling source spectrograms by comparing the performance of the proposed StarGAN-based method, which we refer to as multichannel StarGAN (MSGAN), with ILRMA and MVAE in determined source separation.

2. DETERMINED MULTICHANNEL SOURCE SEPARATION

2.1. Problem formulation

Let us consider a determined situation where I source signals are captured by I microphones. Let $x_i(f,n)$ and $s_j(f,n)$ denote the short-time Fourier transform (STFT) coefficients of the signal observed at the ith microphone and the jth source signal, where f and n are the frequency and time indices, respectively. We denote the vectors containing $x_1(f,n),\dots,x_I(f,n)$ and $s_1(f,n),\dots,s_I(f,n)$ by

$$\boldsymbol{x}(f,n) = [x_1(f,n),\dots,x_I(f,n)]^{\mathsf{T}} \in \mathbb{C}^I, \tag{1}$$

$$\boldsymbol{s}(f,n) = [s_1(f,n),\dots,s_I(f,n)]^{\mathsf{T}} \in \mathbb{C}^I, \tag{2}$$

where $(\cdot)^{\mathsf{T}}$ denotes transpose. In a determined situation, the relationship between observed signals and source signals can be described as

$$\boldsymbol{s}(f,n) = \boldsymbol{W}^{\mathsf{H}}(f)\boldsymbol{x}(f,n), \tag{3}$$

$$\boldsymbol{W}(f) = [\boldsymbol{w}_1(f),\dots,\boldsymbol{w}_I(f)] \in \mathbb{C}^{I\times I}, \tag{4}$$

where $\boldsymbol{W}^{\mathsf{H}}(f)$ is called the separation matrix. $(\cdot)^{\mathsf{H}}$ denotes the Hermitian transpose.

We assume source signals follow the local Gaussian model (LGM) [20]. Namely, $s_j(f,n)$ independently follows a zero-mean complex proper Gaussian distribution with power spectral density $v_j(f,n) = \mathbb{E}[|s_j(f,n)|^2]$:

$$s_j(f,n) \sim \mathcal{N}_{\mathbb{C}}(s_j(f,n)|0, v_j(f,n)). \tag{5}$$

When $s_j(f,n)$ and $s_{j'}(f,n) (j \neq j')$ are independent, $\boldsymbol{s}(f,n)$ follows

$$\boldsymbol{s}(f,n) \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{s}(f,n)|\boldsymbol{0}, \boldsymbol{V}(f,n)), \tag{6}$$

where $\boldsymbol{V}(f,n) = \mathrm{diag}[v_1(f,n),\dots,v_I(f,n)]$. From (3) and (5), we can show that $\boldsymbol{x}(f,n)$ follows

$$\boldsymbol{x}(f,n) \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{x}(f,n)|\boldsymbol{0}, (\boldsymbol{W}^{\mathsf{H}}(f))^{-1}\boldsymbol{V}(f,n)\boldsymbol{W}(f)^{-1}). \tag{7}$$

Hence, the log-likelihood of the separation matrices $\mathcal{W} = \{\boldsymbol{W}(f)\}_f$ and source model parameters $\mathcal{V} = \{v_j(f,n)\}_{j,f,n}$ given the observed mixture signals $\mathcal{X} = \{\boldsymbol{x}(f,n)\}_{f,n}$ is expressed as

$$\log p(\mathcal{X}|\mathcal{W},\mathcal{V}) \stackrel{c}{=} 2N\sum_f \log|\det \boldsymbol{W}^{\mathsf{H}}(f)| - \sum_{f,n,j}\Big(\log v_j(f,n) + \frac{|\boldsymbol{w}_j^{\mathsf{H}}(f)\boldsymbol{x}(f,n)|^2}{v_j(f,n)}\Big), \tag{8}$$

where $=^c$ denotes equality up to constant terms.

2.2. Conventional methods with different source models

2.2.1. ILRMA

Constraints on $v_j(f,n)$ are usually imposed to eliminate the permutation ambiguity during the estimation of $\mathcal{W}$. The NMF model incorporated in ILRMA [7] is an example, which approximates $v_j(f,n)$ as a linear sum of spectral templates $b_{j,k}(f) \geq 0$, $k = 1,\dots,K_j$ scaled by time-varying magnitudes $h_{j,k}(n) \geq 0$, $k = 1,\dots,K_j$, namely, $v_j(f,n) = \sum_k^{K_j} b_{j,k}(f)h_{j,k}(n)$, to capture spectral structures of sources. The parameter estimation algorithm of ILRMA consists of iteratively updating $\mathcal{W}$ using the iterative projection (IP) method [21], and updating $\mathcal{B} = \{b_{j,k}(f)\}_{j,k,f}$ and $\mathcal{H} = \{h_{j,k}(n)\}_{j,k,n}$ using majorization-minimization (MM) algorithm-based update rules.

2.2.2. MVAE

In MVAE [12], a decoder distribution $p_\theta(\boldsymbol{S}|\boldsymbol{z},\boldsymbol{c},g)$ is trained jointly with an encoder distribution $q_\phi(\boldsymbol{z}|\boldsymbol{S},\boldsymbol{c})$ so that the encoder distribtuion $q_\phi(\boldsymbol{z}|\boldsymbol{S},\boldsymbol{c})$ becomes consistent with the true posterior $p_\theta(\boldsymbol{z}|\boldsymbol{S},\boldsymbol{c},g) \propto p_\theta(\boldsymbol{S}|\boldsymbol{z},\boldsymbol{c},g)p(\boldsymbol{z})p(\boldsymbol{c})$, where ϕ and θ are parameters of the encoder and decoder, respectively. Here, $\boldsymbol{S}$ and $\boldsymbol{c}$ respectively denote complex spectrograms and class labels indicating to which class the spectrogram $\boldsymbol{S}$ belongs. For example, if speaker identity is considered as the class category, $\boldsymbol{c}$ will be associated with a different speaker, which is represented as a one-hot vector consisting of M elements, which is filled with 1 at the index of a certain speaker and with 0 everywhere else. The trained decoder distribution, specifically defined as

$$p_\theta(\boldsymbol{S}|\boldsymbol{z},\boldsymbol{c},g) = \prod_{f,n}\mathcal{N}_{\mathbb{C}}(s(f,n)|0, v(f,n)), \tag{9}$$

$$v(f,n) = g\cdot\sigma_\theta^2(f,n;\boldsymbol{z},\boldsymbol{c}), \tag{10}$$

can then be applied as a spectrogram generator, which takes latent variable $\boldsymbol{z}$, speaker ID $\boldsymbol{c}$, and global scale parameter g

as inputs, and outputs the distribution parameter $v(f,n)$. This is called the CVAE source model. At the separation phase, $p_\theta(\boldsymbol{S}_j|\boldsymbol{z}_j,\boldsymbol{c}_j,g_j)$ is then used as the generative model of the complex spectrogram of the source j in a mixture. A stationary point of the log-likelihood (8) that we want to maximize is searched by iteratively updating (A) the separation matrices $\mathcal{W}$ using the IP method [21]:

$$\boldsymbol{w}_j \leftarrow (\boldsymbol{W}^{\mathsf{H}}(f)\boldsymbol{\Sigma}_j(f))^{-1}\boldsymbol{e}_j, \tag{11}$$

$$\boldsymbol{w}_j \leftarrow \frac{\boldsymbol{w}_j(f)}{\boldsymbol{w}_j^{\mathsf{H}}(f)\boldsymbol{\Sigma}_j(f)\boldsymbol{w}_j(f)}, \tag{12}$$

where $\boldsymbol{\Sigma}_j(f) = \frac{1}{N}\sum_n \boldsymbol{x}(f,n)\boldsymbol{x}^{\mathsf{H}}(f,n)/v_j(f,n)$ and $\boldsymbol{e}_j$ denotes the j-th column of an $I \times I$ identity matrix; (B) the CVAE source model parameters $\Psi = \{\boldsymbol{z}_j,\boldsymbol{c}_j\}_j$ with gredient descent (backpropagation); (C) the global scale parameter $\mathcal{G} = \{g_j\}_j$ with the following update rule:

$$g_j \leftarrow \frac{1}{FN}\sum_{f,n}\frac{|\boldsymbol{w}_j^{\mathsf{H}}(f)\boldsymbol{x}(f,n)|^2}{\sigma_\theta^2(f,n;\boldsymbol{z}_j,\boldsymbol{c}_j)}. \tag{13}$$

3. MSGAN WITH STARGAN SOURCE MODEL

Compared to the linear NMF model, the nonlinear CVAE source model not only increases the representation power but also makes it possible to capture the temporal structures of sources thanks to its well-designed network architectures for sequential modeling. MVAE has been shown to exceed ILRMA in multi-speaker separation tasks [12], which implies the effectiveness of the source model with stronger representation power in the LGM-based BSS framework. One promising approach to achieve further improvement includes GAN, where the generator distribution is optimized to fit the real data distribution by playing a minimax game between a generator and a discriminator. Thanks to the training strategy, GAN is expected to learn data distributions much closer to the real one than those that VAE can learn without suffering the mismatch between the assumed and real data distributions and the learning error coming from the approximation. This motivates us to exploit GAN to model power spectrograms. In this section, we first introduce a source model that is trained with StarGAN in a voice conversion fashion in Subsec. 3.1. We then describe network architectures and summarize the algorithm of MSGAN in Subsec. 3.2 and Subsec. 3.3, respectively.

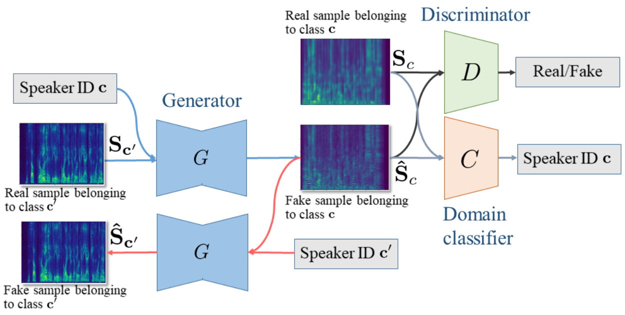

Fig. 1. Concept of StarGAN training.

3.1. Objective functions of StarGAN

Let G be a generator that takes a power spectrogram $\boldsymbol{S}$ and a target speaker ID $\boldsymbol{c}$ as the inputs and generates a power spectrogram $\hat{\boldsymbol{S}} = G(\boldsymbol{S},\boldsymbol{c})$. Note that the generated $\hat{\boldsymbol{S}}$ corresponds to $\sigma_\theta(f,n;\boldsymbol{z},\boldsymbol{c})$ in (10). One of the goals of StarGAN is to make $\hat{\boldsymbol{S}}$ as realistic as real spectrograms belonging to the speaker $\boldsymbol{c}$. To realize this, a real/fake discriminator D as with regular GAN and a domain classifier C are used. D is employed to produce a probability $D(\hat{\boldsymbol{S}})$ to measure how likely the generated $\hat{\boldsymbol{S}}$ is a real spectrogram, whereas C is employed to produce class probabilities $p_C(\boldsymbol{c}|\hat{\boldsymbol{S}})$ of $\hat{\boldsymbol{S}}$.

First, we define an adversarial loss using the Wasserstein GAN with gradient penalty (WGAN-GP) [22], which can stabilize the training procedure of GAN:

$$\begin{aligned}\mathcal{L}_{\text{adv}}^{D}(D) = {} & \mathbb{E}_{\boldsymbol{c}\sim p(\boldsymbol{c}),\, \boldsymbol{S}\sim p(\boldsymbol{S})}[D(G(\boldsymbol{S},\boldsymbol{c}))] - \mathbb{E}_{\boldsymbol{S}\sim p(\boldsymbol{S})}[D(\boldsymbol{S})] \\ & + \lambda_{\text{grad}}\mathbb{E}_{\tilde{\boldsymbol{S}}\sim p(\tilde{\boldsymbol{S}})}[(||\nabla_{\tilde{\boldsymbol{S}}}D(\tilde{\boldsymbol{S}})||_2 - 1)^2], \end{aligned} \tag{14}$$

$$\mathcal{L}_{\text{adv}}^{G}(G) = -\mathbb{E}_{\boldsymbol{S}\sim p(\boldsymbol{S}),\, \boldsymbol{c}\sim p(\boldsymbol{c})}[D(G(\boldsymbol{S},\boldsymbol{c}))]. \tag{15}$$

Here, $\mathbb{E}[\cdot]$ denotes sample mean, $||\cdot||_2$ denotes L_2 norm, and λ_{grad} is a non-negative weight parameter. $\tilde{\boldsymbol{S}}$ denotes data sampled uniformly along straight lines between pairs of points sampled from the real spectrogram distribution $p(\boldsymbol{S})$ and the generator distribution $p_G(\hat{\boldsymbol{S}})$. $\mathcal{L}_{\text{adv}}^{D}(D)$ takes a small value when D correctly classifies $G(\boldsymbol{S},\boldsymbol{c})$ and $\boldsymbol{S}$ as fake and real spectrograms, whereas $\mathcal{L}_{\text{adv}}^{G}(G)$ takes a small value when G successfully deceives D so that $G(\boldsymbol{S},\boldsymbol{c})$ is misclassified as a real spectrogram by D. Next, we consider domain classification losses for classifier C and generator G, which are defined as

$$\begin{aligned}\mathcal{L}_{\text{cls}}^{C}(C) &= -\mathbb{E}_{\boldsymbol{c}\sim p(\boldsymbol{c}),\, \boldsymbol{S}\sim p(\boldsymbol{S}|\boldsymbol{c})}[\log p_C(\boldsymbol{c}|\boldsymbol{S})], \\ \mathcal{L}_{\text{cls}}^{G}(G) &= -\mathbb{E}_{\boldsymbol{c}\sim p(\boldsymbol{c}),\, \boldsymbol{S}\sim p(\boldsymbol{S})}[\log p_C(\boldsymbol{c}|G(\boldsymbol{S},\boldsymbol{c}))]. \end{aligned} \tag{16}$$

Both $\mathcal{L}_{\text{cls}}^{C}(C)$ and $\mathcal{L}_{\text{cls}}^{G}(G)$ take small values when C correctly classifies $\boldsymbol{S} \sim p(\boldsymbol{S}|\boldsymbol{c})$ and $G(\boldsymbol{S},\boldsymbol{c})$ as belonging to speaker $\boldsymbol{c}$. Training G, D, and C using only the above losses does not guarantee that G will preserve the linguistic information of the input spectrogram. To encourage $G(\boldsymbol{S},\boldsymbol{c})$ to be a bijection, a cycle consistency loss is also employed for training, which is expressed as

$$\begin{aligned}&\mathcal{L}_{\text{cyc}}(G) \\ &= \mathbb{E}_{\boldsymbol{c}'\sim p(\boldsymbol{c}),\, \boldsymbol{S}\sim p(\boldsymbol{S}|\boldsymbol{c}'),\, \boldsymbol{c}\sim p(\boldsymbol{c})}[||G(G(\boldsymbol{S},\boldsymbol{c}),\boldsymbol{c}') - \boldsymbol{S}||_1],\end{aligned} \tag{17}$$

where $||\cdot||_1$ denotes L_1 norm. We also consider an identity mapping loss

$$\mathcal{L}_{\text{id}}(G) = \mathbb{E}_{\boldsymbol{c}\sim p(\boldsymbol{c}),\, \boldsymbol{S}\sim p(\boldsymbol{S}|\boldsymbol{c})}[||G(\boldsymbol{S},\boldsymbol{c}) - \boldsymbol{S}||_1] \tag{18}$$

to ensure that an input spectrogram into G will remain un-

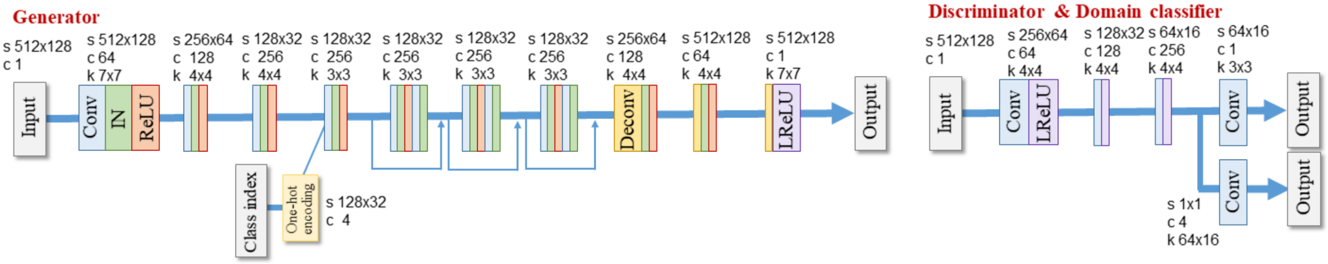

Fig. 2. Network architectures of the generator, discriminator, and domain classifier. “s”, “c”, and “k” denote data size, channel number, and kernel size, respectively. “Conv”, “Deconv”, “IN”, and “LReLU” denote 2-dimensional convolution and deconvolution, instance normalization, and Lecky ReLU, respectively. Class index is concatenated along channel dimension.

changed when the input already belongs to the target speaker.

To summarize, the full objectives of StarGAN to be minimized with respect to G, D, and C are given as

$$\begin{aligned}\mathcal{I}_G(G) =& \mathcal{L}^G_{\text{adv}}(G) + \lambda_{\text{cls}}\mathcal{L}^G_{\text{cls}}(G)\\ &+ \lambda_{\text{cyc}}\mathcal{L}_{\text{cyc}}(G) + \lambda_{\text{id}}\mathcal{L}_{\text{id}}(G), \qquad (19)\\ \mathcal{I}_D(D) =& \mathcal{L}^D_{\text{adv}}(D), \qquad (20)\\ \mathcal{I}_C(C) =& \mathcal{L}^C_{\text{cls}}(C), \qquad (21)\end{aligned}$$

respectively, where $\lambda_{\text{cls}} \geq 0$, $\lambda_{\text{cyc}} \geq 0$, $\lambda_{\text{id}} \geq 0$ are regularization parameters weighing the importance of the domain classification loss, the cycle consistency loss, and the identity mapping loss relative to the adversarial losses. Fig. 1 shows the concept of a StarGAN.

3.2. Network architectures

For network architectures, we consider networks constructed using fully convolutional layers to cope with signals having arbitrary lengths, and capture time dependencies. Specifically, we use 2-dimensional convolutional neural networks (CNNs) to design the architectures for all the networks. The generator consists of two parts, similar to an encoder-decoder structure. The first part aims to extract the low-dimensional latent representation $\boldsymbol{z}$ of the input spectrogram, whereas the second part takes class index $\boldsymbol{c}$ as an auxiliary input and performs spectrogram conversion. We used the second part as the source model at the separation phase, where the latent variable $\boldsymbol{z}$ and the auxiliary input $\boldsymbol{c}$ are treated as the model parameters to be estimated. We leverage the idea of PatchGANs [23] to devise a real/fake discriminator D, the output of which is a sequence of probabilities that measures how likely each segment of the input is to be real. This forces the generator to generate more local details. Otherwise, it will fail to deceive the discriminator. The domain classifier C is designed to share the low-level features with the discriminator. More architecture details are shown in Fig. 2.

3.3. Algorithm of MSGAN for source separation

The proposed MSGAN method for determined multichannel source separation is summarized as follows.

1. Train G, D, and C using (19), (20), and (21).
2. Initialize $\mathcal{W}$ using a BSS method, e.g., ILRMA.
3. Iterate the following steps for each j:
   (a) Compute $\boldsymbol{S}_j = \boldsymbol{w}^{\mathsf{H}}_j(f)\boldsymbol{x}(f,n)$.
   (b) Update $\Psi_j = \{\boldsymbol{z}_j, \boldsymbol{c}_j\}$ using backpropagation to maximize (8).
   (c) Update g_j using (13).
   (d) Update $\boldsymbol{w}_j(0), \ldots, \boldsymbol{w}_j(F)$ using (11), (12).

Note that the global scaling parameter g_j has to be considered as with MVAE to eliminate the scale mismatch between the normalized training data and test data.

4. EXPERIMENTAL EVALUATIONS

To evaluate the effectiveness of the proposed StarGAN source model, we conducted experiments designed to compare the multi-speaker separation performance of MSGAN with ILRMA [7] and MVAE[1] [12].

4.1. Experimental conditions

We excerpted speech utterances from two male speakers (‘SM1’, ‘SM2’) and two female speakers (‘SF1’, ‘SF2’) from the Voice Conversion Challenge (VCC) 2018 dataset [24]. The audio files for each speaker were about 7 minutes long and manually segmented into 116 short sentences, where 81 and 35 sentences (about 5 and 2 minutes long, respectively) were used as training and test sets, respectively. We used two-channel mixture signals of two sources as the test data, which were synthesized using the simulated room impulse responses (RIRs) generated using the image method and real RIRs (ANE and E2A) excerpted from the RWCP Sound Scene Database in Real Acoustic Environments [25]. The configuration of the room was the same as the one in [12]. The reverberation times (RT_{60}) of the simulated RIRs were 78 and 351 ms, and those of real RIRs were 173 and 225 ms. We generated test data comprising 4 speaker pairs and 10 sentences for each pair, each of which was about 4 to 7 seconds long. All the speech signals were resampled at 16 kHz. To decrease memory usage during the network training,

[1] Code: https://github.com/lili-0805/MVAE

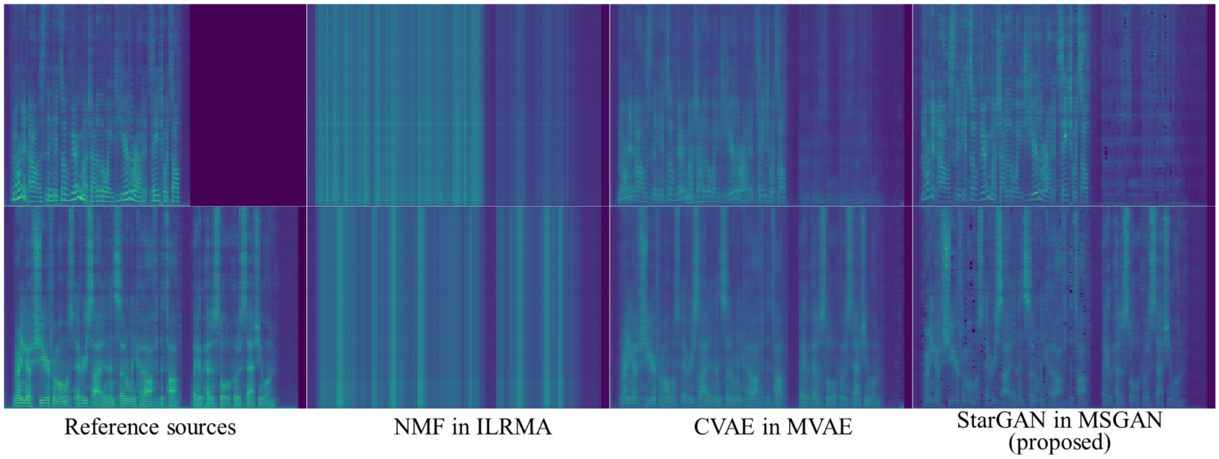

Fig. 3. Example of different source models obtained under “ANE” condition.

we computed STFT using a Hamming window with a length of 64 ms and a shift of 32 ms, whereas the spectrograms were computed with a window length of 128 ms and window shift of 64 ms at the separation phase.

ILRMA was run for 60 iterations. Both MVAE and MSGAN were run for 30 iterations after the initialization. To initialize $\mathcal{W}$ for MVAE and MSGAN, we used ILRMA run for 30 iterations. The basis number of ILRMA was set at 1. Adam [26] was used to train networks and estimate the parameter Ψ in the algorithms. The source-to-distortion ratio (SDR), source-to-interferences ratio (SIR), and sources-to-artifacts ratio (SAR) [27] were calculated for source separation performance. Perceptual evaluation of speech quality (PESQ) [2] [28] and short-time objective intelligibility (STOI) [3] [29] were also conducted to evaluate the speech quality and intelligibility.

4.2. Results

Table 1 shows SDR, SIR, SAR, PESQ, and STOI scores obtained by ILRMA, MVAE, and the proposed MSGAN. All the results were averaged over the 40 test signals under each reverberant condition. The results reveal that MSGAN outperformed ILRMA in terms of all the criteria. Meanwhile, it achieved slight improvement in MVAE in terms of SDR and comparative results in terms of other criteria. Comparing the results under each reverberant condition, we find that MSGAN performed better in low reverberant situations and the performance degraded with relatively heavy reverberation. Fig. 3 depicts an example of the power spectrograms estimated by different methods. Compared to the spectrogram estimated by ILRMA where the basis number was 1, the CVAE and StarGAN source models used in the MVAE and MSGAN captured spectro-temporal structures of sources more precisely. Moreover, we found that the StarGAN could represent more details of harmonics than CVAE, while it might lead to more distortions locally.

Table 1. SDR, SIR, SAR [dB], PESQ, and STOI achieved by ILRMA, MVAE, and MSGAN under various reverberant conditions.

Reverberant conditions	ILRMA	MVAE	MSGAN
$RT_{60} = 78$ ms	21.22	22.69	24.08
$RT_{60} = 351$ ms	5.93	7.63	6.09
ANE ($RT_{60} = 173$ ms)	19.61	19.44	20.92
E2A ($RT_{60} = 225$ ms)	6.05	6.76	6.36
Average SDR	**13.20**	**14.13**	**14.36**
$RT_{60} = 78$ ms	27.32	27.38	28.85
$RT_{60} = 351$ ms	12.45	14.95	12.34
ANE ($RT_{60} = 173$ ms)	25.16	23.73	25.69
E2A ($RT_{60} = 225$ ms)	13.43	15.28	13.98
Average SIR	**19.59**	**20.33**	**20.21**
$RT_{60} = 78$ ms	23.25	26.31	27.21
$RT_{60} = 351$ ms	7.81	8.97	8.11
ANE ($RT_{60} = 173$ ms)	21.97	23.41	24.08
E2A ($RT_{60} = 225$ ms)	7.61	7.94	7.95
Average SAR	**15.16**	**16.66**	**16.84**
$RT_{60} = 78$ ms	3.38	3.40	3.50
$RT_{60} = 351$ ms	1.96	2.05	1.96
ANE ($RT_{60} = 173$ ms)	3.11	3.18	3.19
E2A ($RT_{60} = 225$ ms)	2.32	2.36	2.31
Average PESQ	**2.69**	**2.75**	**2.74**
$RT_{60} = 78$ ms	0.9472	0.9375	0.9480
$RT_{60} = 351$ ms	0.8059	0.8221	0.8074
ANE ($RT_{60} = 173$ ms)	0.9065	0.9047	0.9047
E2A ($RT_{60} = 225$ ms)	0.7635	0.7666	0.7585
Average STOI	**0.8558**	**0.8577**	**0.8547**

5. CONCLUSIONS

This paper proposed a determined multichannel source separation method, which incorporates a source model trained

[2] Code: https://github.com/vBaiCai/python-pesq

[3] Code: https://github.com/mpariente/pystoi

using StarGAN into the LGM-based BSS framework. We investigated the effectiveness of the StarGAN source model in source separation and compared it with the NMF model adopted in ILRMA and the CVAE source model employed in the MVAE. The results showed that the proposed method outperformed ILRMA in terms of all the criteria and exceeded MVAE in terms of SDR, which confirmed the effectiveness of StarGAN in modeling spectrograms.

6. REFERENCES

[1] S. Makino, "Blind source separation of convolutive mixtures of speech," in *Adaptive signal processing*, pp. 195–225, 2003.

[2] T. Kim, T. Eltoft and T.-W. Lee, "Independent vector analysis: An extension of ICA to multivariate components," in *Proc. ICA*, pp. 165–172, 2006.

[3] A. Hiroe, "Solution of permutation problem in frequency domain ICA using multivariate probability density functions," in *Proc. ICA*, pp. 601-608, 2006.

[4] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. ASLP*, vol. 18, no. 3, pp. 550–563, 2010.

[5] H. Sawada, H. Kameoka, S. Araki and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE Trans. ASLP*, vol. 21, no. 5, pp. 971–982, 2013.

[6] H. Kameoka, T. Yoshioka, M. Hamamura, J. Le. Roux and K. Kashino, "Statistical model of speech signals based on composite autoregressive system with application to blind source separation," in *Proc. LVA/ICA*, pp. 245–253, 2010.

[7] D. Kitamura, N. Ono, H. Sawada, H. Kameoka and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Trans. ASLP*, vol. 24, no. 9, pp. 1622–1637, 2016.

[8] D. P. Kingma, S. Mohamed, D. J. Rezende and M. Welling, "Semi-supervised learning with deep generative models," in *Adv. Neural Information Processing Systems (NIPS)*, pp. 3581–3589, 2014.

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. NIPS*, pp. 2672–2680, 2014.

[10] Y. Bando, M. Mimura, K. Itoyama, K. Yoshii and T. Kawahara, "Statistical speech enhancement based on probabilistic integration of variational autoencoder and non-negative matrix factorization," in *Proc. ICASSP*, pp. 716–720, 2018.

[11] S. Leglaive, L. Girin and R. Horaud, "A variance modeling framework based on variational autoencoders for speech enhancement," in *Proc. MLSP*, 2018.

[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] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech enhancement generative adversarial network," *eprint arXiv: 1703.09452*, March, 2017.

[14] X. Hao, X. Su, Z. Wang, H. Zhang, and Batushiren, "UNet-GAN: A Robust Speech Enhancement Approach in Time Domain for Extremely Low Signal-to-Noise Ratio Condition," in *Proc. Interspeech*, pp. 1786–1790, 2019.

[15] J. Lin, S. Niu, Z. Wei, X. Lan, A. Jvan Wijngaarden, M. C Smith, and K. Wang, "Speech enhancement using forked generative adversarial networks with spectral subtraction," in *Proc. Interspeech*, pp. 3163–3167, 2019.

[16] S. Seki, H. Kameoka, L. Li, T. Toda and K. Takeda, "Generalized multichannel variational autoencoder for underdetermined source separation," in *Proc. EUSIPCO*, pp. 1973–1977. 2019.

[17] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. CVPR*, pp. 8789–8797, 2018.

[18] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "StarGAN-VC: Non-parallel many-tomany voice conversion using star generative adversarial networks," in *Proc. SLT*, pp. 266–273, 2018.

[19] 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 *Proc. ICASSP*, pp. 546–550, 2019.

[20] C. Févotte and J-F Cardoso, "Maximum likelihood approach for blind audio source separation using time-frequency Gaussian source models," in *Proc. WASPAA*, pp. 78–81, 2005.

[21] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. WASPAA*, pp. 189–192, 2011.

[22] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C Courville, "Improved training of Wasserstein GANs," in *Proc. NIPS*, pp. 5767–5777, 2017.

[23] P. Isola, J.-Y. Zhu, T. Zhou, and A. A Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. CVPR*, pp. 1125–1134, 2017.

[24] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F, Villavicencio, T. Kinnunen and Z. Ling, "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods," *eprint arXiv: 1804.04262*, Apr. 2018.

[25] S. Nakamura, K. Hiyane, F. Asano and T. Endo, "Sound scene data collection in real acoustical environments," *J. Acoust. Soc. Jpn. (E)*, vol. 20, no. 3, pp. 225–231, 1999.

[26] D. Kingma, J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.

[27] E. Vincent, R. Gribonval and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. ASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.

[28] A. W. Rix, J. G. Beerends, M. P. Hollier and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*, Cat. No. 01CH37221, vol. 2, pp. 749–752, 2001.

[29] C. H. Taal, R. C. Hendriks, R. Heusdens and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. ICASSP*, pp. 4214–4217, 2010.