

DOI:10.13232/j.cnki.jnju.2021.04.016

指定输出通道排序的半监督盲源分离算法

顾昭仪, 卢 晶*

(近代声学教育部重点实验室, 南京大学声学研究所, 南京, 210093)

摘 要:在频域操作的联合盲源分离算法可以有效解决频点间的内部排序问题,然而对于输出通道的排序,即全局排序,现有的基于频域的联合盲源分离算法仍无法有效确定.使用基于变分自编码器的声源模型,通过预先指定从注册语料中获得的各话者编码向量的排列顺序来调控分离后输出通道的顺序.该方法使用带有实例归一化和自适应实例归一化层的变分自编码器来确保这一排序方式的有效性.此外,为了减少频域上联合盲源分离算法可能出现的块排序问题,提出使用人为构造的两类含噪信号对变分自编码器中的解码器网络单独进行降噪训练的方案.利用实际录制的房间冲激响应的仿真结果表明,该方案可以在保证算法分离性能的同时,有效地按照预期的输出顺序调控输出通道.

关键词:联合盲源分离,全局排序,变分自编码器,实例归一化

中图分类号:O429

文献标志码:A

Semi-supervised blind speech separation with designated channel order

Gu Zhaoyi, Lu Jing*

(Key Laboratory of Modern Acoustics, Ministry of Education, Institute of Acoustics of Nanjing University,
Nanjing, 210093, China)

Abstract: Frequency domain joint blind source separation (FD-JBSS) has been demonstrated as an effective method to deal with the internal permutation problem. However, the ordering of the output channels cannot be efficiently arranged or recognized using current FD-JBSS algorithms, resulting in an unsettled global permutation problem. In this paper, this problem is addressed simultaneously in the separation process by pre-assigning the order of speaker embeddings extracted from enrollment utterances with a variational autoencoder(VAE)-based source model. Two normalization strategies, i.e. instance normalization (IN) and adaptive instance normalization (AdaIN) are utilized in the VAE architecture to enforce an arbitrary channel permutation. To mitigate the possible block permutation problem of FD-JBSS and further improving the separation performance, we propose a denoising training stage solely to the decoder network using two kinds of artificially constructed noisy signals. Separation performance and accuracy of output channel arrangement of the proposed method are evaluated using measured room impulse responses (RIRs) using both seen and unseen speakers.

Key words: joint blind source separation, global permutation, variational autoencoder, instance normalization

盲源分离(Blind Source Separation, BSS)的目标是从传声器接收到的混合信号中恢复各声源的信号,该过程作为自动语音识别(Automatic Speech Recognition, ASR)的前端模块,在智能家

居、车载系统以及线上线下会议系统中有很高的应用价值.频域盲源分离算法因其较低的计算负担以及稳定的分离效果,近年来得到了广泛的研究^[1-3].其中,基于独立向量分析(Independent

基金项目:国家自然科学基金(11874219)

收稿日期:2021-03-12

* 通讯联系人, E-mail: lujing@nju.edu.cn

Vector Analysis, IVA)^[4-5]的联合盲源分离(Joint Blind Source Separation, JBSS)算法通过对每个声源所有频点信号联合建模的方式刻画频点间高阶统计信息,成功解决了独立成分分析(Independent Component Analysis, ICA)^[6]算法的内部排序问题^[7-8],是一种稳定高效的频域分离算法.通过寻找更合适的声源模型,如学生 t 分布^[9-10]、混合高斯分布^[11],使IVA算法的分离性能得到进一步提升.引入非负矩阵分解^[12](Non-Negative Matrix Factorization, NMF)的独立低秩矩阵分析(Independent Low-Rank Matrix Analysis, ILRMA)算法^[13-14]是IVA算法目前广为认可的一类最优方案,ILRMA提出的基于局部高斯分布的声源模型以及基于迭代投影^[15](Iterative Projection, IP)的参数优化方式为应用更加灵活的声源模型提供了一种通用的算法框架,通过选取各声源的空间相关矩阵为满秩的矩阵可以导出一系列适用于复杂场景的频域分离算法^[16-18].虽然上述算法解决了频点间的排序错误,一个尚待解决的问题是算法分离后输出通道之间的排序问题,也称为全局排序问题.频域的联合盲源分离算法在分离过程中对每个声源的处理都是等同的,因此无法决定输出通道和源信号间的对应关系.这一问题使此类盲源分离算法在实际应用中受到了许多限制.

为了解决全局排序的问题,一种有监督的IVA算法^[19-20]通过引入引导信号的方式对目标信号进行提取,并且利用引导信号和目标信号之间的相关性,使目标信号在某一特定通道被输出.然而,该引导信号的选择方式并不直观,所以该方法或是受限于特定的声源方位范围^[19],或是仅适用于单目标话者的场景^[20].另一类方法假设目标声源方位已知,通过对IVA算法中目标声源对应的分离向量施加空间约束的方式引导该算法在指定的通道上产生指定方向源信号的输出^[21-23].该类方法需要同时知道各个声源的方位信息以及麦克风阵列的配置信息来获得声源的导向向量,因此仅适用于有限的应用场景.

对于智能语音交互系统,说话人注册语音是更容易获得的信息,可以用来设计输出通道顺序可控的盲源分离算法.其中,基于端到端方案

的话者提取网络,如SpeakerBeam^[24-25],能利用注册语音或多通道的阵列空间信息从混合信号特征谱中估计特定声源信号的时频掩模(Time-frequency Mask),以实现目标源的提取^[26-27].然而,这种端到端的训练方式一方面限制了测试时能够提取的话者数量,另一方面也面临泛化性能的挑战.近年来,一种将数据驱动方案和频域分离算法框架相结合的半监督盲源分离算法(Multichannel Variational Autoencoder, MVAE)得到了广泛的研究和关注^[28].该算法使用条件变分自编码器^[29](Conditional Variational Autoencoder, CVAE)这一深度生成模型为声源信号建模,充分利用了神经网络在大数据支撑下的学习能力,可以为基于局部高斯模型的盲源分离算法提供更准确的声源分布信息,提升了算法的分离性能;此外,在模型中引入的基于独热(one-hot)编码的话者嵌入向量能使该算法在理论上有判别输出通道顺序的能力.后续的工作在MVAE的基础上,通过训练一个额外的分类器来提高算法对输出通道顺序判别的准确率^[30].另外,通过引入话者分类器和音素分类器,语料的音素信息对提升分离效果和提升输出通道判别准确性的作用也被研究和证实^[31].

由于上述半监督的盲源分离算法采用了独热编码的方式,其判别通道顺序的能力局限于训练时已见的话者,并且判别的准确性也有很大的提升空间.受话者迁移领域相关研究的启发^[32],本文引入基于实例归一化^[33](Instance Normalization, IN)和自适应实例归一化^[34](Adaptive Instance Normalization, AdaIN)的变分自编码器(Variational Autoencoder, VAE)作为声源模型,提出一种可以指定输出通道排列顺序的半监督盲源分离算法.应用IN和AdaIN策略,该算法使用的变分自编码器模型将语音信号中的话者信息和内容信息有效解耦,从而达到根据指定的话者编码向量的排列顺序来指导分离通道输出的效果.为了减少分离信号产生的块排序错误^[35],还提出专门针对解码器网络训练的降噪方案.此外,应用联合训练的话者编码器,该算法也具备泛化到训练时未见话者数据集上的能力.在仿真中,使用实测的房间冲激响应数据集对所提算法以及

MVAE算法的分离和通道排序能力进行了详细的对比评测.

1 多通道半监督盲源分离算法

1.1 问题陈述 考虑 J 个声源和 I 个传声器同时存在的混合场景. 各个声源发出的信号 $s_j(t)$ 经过一系列房间冲激响应 $a_{ij}(t)$ ($i=1, 2, \dots, I; j=1, 2, \dots, J$) 被第 i 个传声器采集的过程在时域上可以用如下数学形式表示:

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

其中, $x_i(t)$ 表示第 i 个传声器接收到的 J 个声源的混合信号, $a_{ij}(t)$ 表示第 j 个声源到第 i 个传声器的房间冲激响应, L 是房间冲激响应的长度. 对上述时域信号做短时傅里叶变换 (Short-Time Fourier Transform, STFT), 并假设 L 小于短时傅里叶变换的分析窗窗长, 上述混合模型可以转换为时频域上的线性瞬时混合形式, 即:

$$x(f, n) = A(f) s(f, n) \quad (2)$$

其中, $f=1, 2, \dots, F, n=1, 2, \dots, N$ 分别表示频点和帧数的索引, F, N 是总频点数和帧数.

$$x(f, n) = [x_1(f, n), x_2(f, n), \dots, x_I(f, n)]^T \quad (3)$$

$$s(f, n) = [s_1(f, n), s_2(f, n), \dots, s_J(f, n)]^T \quad (4)$$

$x(f, n)$ 和 $s(f, n)$ 分别是 I 维和 J 维的向量, 表示第 f, n 个时频点传声器阵列接收到的信号以及声源发出的信号, $[\cdot]^T$ 表示向量的转置. $A(f)$ 是频点 f 处的混合矩阵, 并且有:

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

其中, $a_{ij}(f)$ 是房间冲激响应 $a_{ij}(t)$ 经过傅里叶变换后在频点 f 处的取值.

考虑正定的情景, 即传声器数量 I 和声源数量 J 相等的情况, 那么当 $A(f)$ 是可逆矩阵的时候, 对声源信号 $s(f, n)$ 的估计 $y(f, n)$ 可以通过如下分离系统获得:

$$y(f, n) = W(f) x(f, n) \quad (6)$$

$$y(f, n) = [y_1(f, n), y_2(f, n), \dots, y_I(f, n)]^T \quad (7)$$

$$W(f) = [w_1(f), w_2(f), \dots, w_I(f)]^H \quad (8)$$

其中, $W(f)$ 是频点 f 处 $I \times I$ 维的分离矩阵, 并有 $W(f) = A^{-1}(f) \cdot (\cdot)^H$ 和 $(\cdot)^{-1}$ 分别表示共轭转置操作及求逆操作. 盲源分离算法的目标是从传声器接收到的混合信号 $x(f, n)$ 中得到对分离矩阵 $W(f)$ 的估计, 从而根据式 (7) 获得分离后的信号 y .

1.2 信号模型 理论上 I 个声源之间相互独立, 并且每个声源满足如下所示的局部高斯模型^[28]:

$$p_{S_i}(s_i) = \prod_{f,n} p(s_i(f, n) | 0, v_i(f, n)) \quad (9)$$

$$N_c(z | 0, \sigma^2) = \frac{1}{\pi \sigma^2} e^{-\frac{|z|^2}{\sigma^2}} \quad (10)$$

其中, 第 i 个声源每个时频点的信号 $s_i(f, n)$ 独立地服从均值为 0、方差为 $v_i(f, n)$ 的圆对称复高斯分布 $N_c, S_i = \{s_i(f, n)\}_{f,n}$ 表示第 i 个声源所有时频点处信号的集合.

根据式 (7) 和式 (9), 易得混合信号 $x(f, n)$ 服从如下所示的多变量圆对称复高斯分布:

$$p_{x(f,n)}(x(f, n)) = N_c\left(x(f, n) | 0, (W(f))^{-1} V(f, n) (W^H(f, n))^{-1}\right) \quad (11)$$

其中, $V(f, n)$ 是第 i 个对角元为 $v_i(f, n)$ 的对角矩阵. 因此, 给定观测信号 x , 基于最大似然的分离算法代价函数有如下形式:

$$J(W, V) = - \sum_{f,n,i} \left[\lg v_i(f, n) + \frac{|y_i(f, n)|^2}{v_i(f, n)} \right] + \sum_f \lg |\det W(f)|^{2N} \quad (12)$$

其中, $|\cdot|$ 和 $\det(\cdot)$ 分别用来表示取模运算和矩阵行列式; $W = \{W(f)\}_f, V = \{v_i(f, n)\}_{f,n,i}$ 是待优化参数. 本文用 $\{u(i_1, \dots, i_n)\}_{i_1, \dots, i_n}$ 表示下标 i_1, \dots, i_n 在所有取值下变量 $u(i_1, \dots, i_n)$ 的集合.

1.3 多通道变分自编码器算法 考虑语音信号结构的复杂性, MVAE 算法使用基于深度神经网络的 CVAE 作为语音信号的生成模型. CVAE 联

合训练一个编码器分布 $q_\phi(\mathbf{z}|\bar{\mathbf{S}}, \mathbf{c})$ 和一个解码器分布 $p_\theta(\bar{\mathbf{S}}|\mathbf{z}, \mathbf{c})$, 其中 $\bar{\mathbf{S}}$ 表示纯净语音的复数域时频谱信号, \mathbf{z} 是模型的隐变量, \mathbf{c} 是训练时已知的条件变量, θ 和 ϕ 分别表示该网络解码器和编码器的待训练参数. 在 MVAE 算法的训练过程中, \mathbf{c} 设置为表示话者身份信息的独热编码向量, 解码器输出的分布形式和式(9)一致, 可以表示为:

$$p_\theta(\bar{\mathbf{S}}|\mathbf{z}, \mathbf{c}) = \prod_{f,n} N_c(\bar{s}(f,n)|0, g\sigma_\theta^2(f,n;\mathbf{z}, \mathbf{c})) \quad (13)$$

其中, g 是和网络参数同时训练的全局幅度调整因子, $\{\sigma_\theta^2(f,n;\mathbf{z}, \mathbf{c})\}_{f,n}$ 是解码器输出. 在该生成模型下, MVAE 分离过程中假设的各声源模型也满足式(13)所示的形式, 即:

$$v_i(f,n) = g_i\sigma_{\theta_i}^2(f,n;\mathbf{z}_i, \mathbf{c}_i) \quad (14)$$

其中, $\{g_i\}_i, \{\mathbf{z}_i\}_i, \{\mathbf{c}_i\}_i$ 为分离算法中的待估计参数集合. 将式(14)代入式(12), 即可得到 MVAE 的代价函数. 在分离阶段, $\{\mathbf{z}_i\}_i, \{\mathbf{c}_i\}_i$ 通过反向传播进行优化, $\{g_i\}_i$ 通过对代价函数求导获得最优解, 对分离矩阵 \mathbf{W} 的更新采用 IP 算法^[15]. 上述参数优化过程交替迭代进行直至算法达到收敛.

2 确定输出通道排序的半监督分离算法

MVAE 算法在声源模型中引入表示话者身份信息独热编码向量, 使该算法在对话者编码向量 \mathbf{c} 的更新过程中具有一定的判别输出通道顺序的能力. 然而在网络训练的过程中, 由于该算法未对编、解码器的结构或代价函数做出任何约束, 导致 CVAE 容易出现如下的退化问题^[30], 即 $q_\phi(\mathbf{z}|\bar{\mathbf{S}}, \mathbf{c}) = q_\phi(\mathbf{z}|\bar{\mathbf{S}})$, 并且 $p_\theta(\bar{\mathbf{S}}|\mathbf{z}, \mathbf{c}) = p_\theta(\bar{\mathbf{S}}|\mathbf{z})$. 此时, CVAE 的解码器模型倾向于忽略话者信息, 仅从隐变量 \mathbf{z} 中学习目标语音的条件概率密度分布. 在分离阶段, 这一退化问题使反向传播时对分离通道的话者信息 \mathbf{c} 的估计不再准确, 因而难以获得令人满意的通道顺序判别效果. 此外, 基于独热编码的话者编码方式使该通道判别方案无法用于训练时未见的话者, 限制了该算法解决全局排序问题的普适性.

受近年来基于变分自编码器的话者迁移方案的启发^[32], 本文在第 1 节描述的分离算法的框架下, 利用基于实例归一化^[33]以及自适应实例归一化^[34]的变分自编码器模型作为声源模型, 将表示话者信息的编码和表示内容信息的编码解耦, 针对各声源注册信号已知的场景提出一种可以确定输出通道间排列顺序的半监督分离算法.

2.1 变分自编码器 该变分自编码器采用 Chou et al^[32]的结构, 由内容编码器 E_c 、话者编码器 E_s 以及解码器 D 三部分构成, 其网络框架如图 1 所示. 其中话者编码器用于输出每段训练数据的话者编码向量 \mathbf{z}_s , $\mathbf{z}_s = E_{s,\psi}(\bar{\mathbf{S}})$; 内容编码器用于估计模型中隐藏变量 \mathbf{z}_c 的后验概率密度分布 $q_\phi(\mathbf{z}_c|\bar{\mathbf{S}})$; 解码器则用于生成目标语音信号的条件概率密度分布 $p_\theta(\bar{\mathbf{S}}|\mathbf{z}_c, \mathbf{z}_s)$. θ, ϕ, ψ 分别表示解码器、内容编码器以及话者编码器的网络参数.

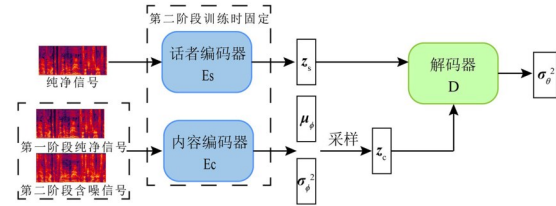


图 1 VAE 网络框架图

Fig. 1 The overview of VAE network

假设 $\bar{\mathbf{S}} \in \mathbb{C}^{F \times N}$ 表示一段纯净语音训练样本, 其中 F, N 分别代表该样本的总频点数和帧数, 该生成模型旨在通过最大化训练数据的似然函数来训练解码器刻画的目标信号模型 $p_\theta(\bar{\mathbf{S}}|\mathbf{z}_c, \mathbf{z}_s)$, 训练的目标函数如下:

$$J(\theta, \phi, \psi) = \mathbb{E}_{\bar{\mathbf{S}} \sim p_{\text{clean}}(\bar{\mathbf{S}})} \left[\lg \int p_\theta(\bar{\mathbf{S}}|\mathbf{z}_c, \mathbf{z}_s) p(\mathbf{z}_c) d\mathbf{z}_c \right] \quad (15)$$

其中, p_{clean} 表示纯净信号的数据集分布; $p(\mathbf{z}_c)$ 选为 $N_c(\mathbf{z}_c|0, \mathbf{I})$, 表示隐藏变量的先验分布, \mathbf{I} 是单位阵; \mathbb{E} 是期望运算符. 由于无法准确得到 \mathbf{z}_c 的真实后验概率密度分布 $p(\mathbf{z}_c|\bar{\mathbf{S}})$, VAE 采用变分机制, 通过优化式(15)的变分下界来间接优化该目标函数. 利用 Jensen 不等式可以导出该下界函数为^[36]:

$$Q(\theta, \phi, \psi) \triangleq \mathbb{E}_{\tilde{S} \sim p_{\text{clean}}(S)} \left\{ \lambda_{\text{rec}} \mathbb{E}_{z_c \sim q_\phi(z_c | \tilde{S})} [\lg p_\theta(\tilde{S} | z_c, z_s)] - \lambda_{\text{KL}} \text{KL}[q_\phi(z_c | \tilde{S}) \| p(z_c)] \right\} \quad (16)$$

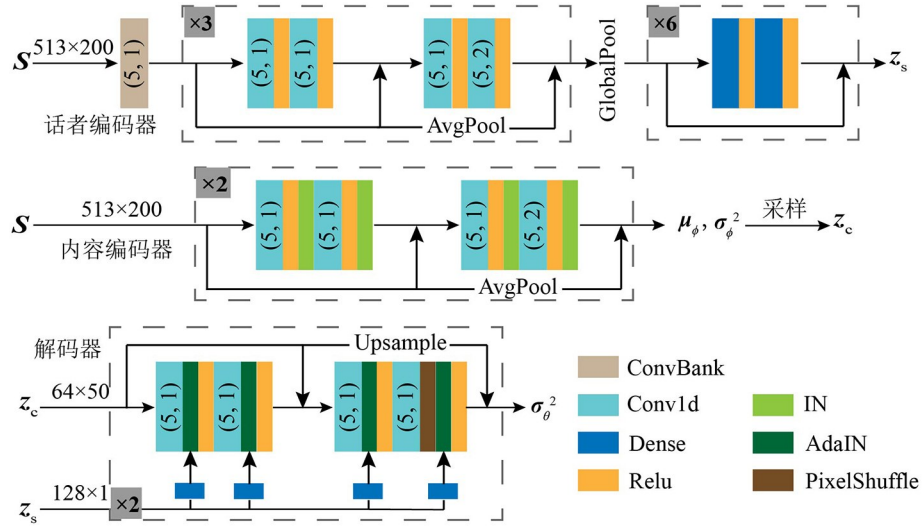
其中, $q_\phi(z_c | \tilde{S})$ 为内容编码器刻画出的 z_c 的后验概率密度分布. 式(16)右侧第一项反映解码器的重构能力, 第二项 $\text{KL}[p \| q]$ 表示两个概率分布 p 和 q 之间的 Kullback-Leibler(KL) 散度, λ_{rec} 和 λ_{KL} 分别用于调整信号条件概率密度的期望和 KL 散度在目标函数中的权重. 对于普通 VAE, $\lambda_{\text{rec}} = \lambda_{\text{KL}} = 1$. 当内容编码器生成的后验概率密度分布 $q_\phi(z_c | \tilde{S})$ 和真实的后验概率密度分布 $p(z_c | \tilde{S})$ 一致时, $Q(\theta, \phi, \psi)$ 与 $J(\theta, \phi, \psi)$ 完全相同. 对式(16)涉及的各项分布选取如下形式:

$$q_\phi(z_c | \tilde{S}) = \prod_d N_c(z_c(d) | \mu_\phi(d | \tilde{S}), \sigma_\phi^2(d | \tilde{S})) \quad (17)$$

$$p_\theta(\tilde{S} | z_c, z_s) = \prod_{f,n} N_c(\tilde{s}(f,n) | 0, \sigma_\theta^2(f,n; z_c, z_s)) \quad (18)$$

其中, d 表示隐藏变量 z_c 中元素的索引, $\{\mu_\phi(d | \tilde{S})\}_d, \{\sigma_\phi^2(d | \tilde{S})\}_d$ 以及 $\{\sigma_\theta^2(f,n; z_c, z_s)\}_d$ 分别由内容编码器以及解码器输出. 实际网络训练时的目标为最大化式(16)所示的变分下界.

图2展示了本文所用的网络架构图. 为了让解码器网络的输出能够尽可能地同时利用话者编码向量 z_s 以及内容编码向量 z_c 的信息, 该网络对内容编码器隐藏层的输出应用实例归一化以逐层消除内容编码向量 z_c 中包含的全局信息. 由于部分话者信息, 如基频、谐频等可以被视作全局信息, 因此该归一化方法能够有效减少 z_c 中话者相关的成分, 达到将 z_c 和 z_s 解耦的目的. 与此同时, 对解码器应用的自适应实例归一化方法可以迫使解码器网络仅依据话者编码向量恢复信号的全局信息, 进一步避免解码器网络在生成信号的条件概率密度分布时发生退化.



Conv1d 及 ConvBank 括号里的元素分别表示卷积核和步长大小, 箭头上方的元素表示输入特征图的长×宽

图2 VAE 网络架构图

Fig. 2 The architecture of the VAE network

2.2 针对解码器的降噪训练 仿真中发现直接应用上述训练的变分自编码器作为声源模型会使语音分离的结果出现大量的块排序错误, 影响算法的分离性能. 为此在对 VAE 进行上述第一阶段训练的基础上, 提出第二阶段降噪训练的方案. 和式(13)类似, 考虑分离算法仅用了解码器

的输出, 该降噪训练的目标在于提升解码器网络输出纯净语音信号分布的鲁棒性. 因此对话者编码器和内容编码器网络的训练在第一阶段由纯净的训练数据完成, 降噪训练时仅调整解码器的网络参数, 从而针对性地提升解码器的降噪能力. 考虑分离过程中经常出现的块排序噪声以及算法

中间迭代步骤产生的含噪信号,本文设计了两种加噪方式,具体的数据增强方式在 3.1 进行了详细的说明.

根据上述分析,第二阶段训练的目标函数可以写成如下的形式:

$$Q_{II}(\theta; \phi, \psi) \triangleq \mathbb{E}_{\bar{s} \sim p_{\text{clean}}(s)} \left\{ \mathbb{E}_{z_c \sim q_{\phi}(z_c | f(s))} \left[\lg p_{\theta}(\bar{s} | z_c, E_{s, \psi}(\bar{s})) \right] \right\} \quad (19)$$

其中,话者编码向量通过纯净的信号获得, z_c 从含噪信号 $f(\bar{s})$ 经过内容编码器得到的后验分布中采样获得, $f(\cdot)$ 表示对纯净信号加噪的方式. 经过第二阶段训练,解码器网络能够在一定程度上抑制特定含噪信号的产生,减轻分离时的块排序问题.

2.3 分离算法流程 应用解码器的输出并引入全局幅度调整因子 g_i , 各声源满足的局部高斯模型可以重述为如下的形式:

$$p_{s_i}(S_i) = \prod_{f, n} N_c(s_i(f, n) | 0, g_i \sigma_{\theta}^2(f, n; z_c^i, z_s^i)) \quad (20)$$

其中, z_c^i, z_s^i 分别表示第 i 个输出通道信号的内容及话者编码向量. 令 $Senroll_i$ 表示第 i 个话者的注册语音, 其话者编码向量 z_s^i 可由 $z_s^i = E_{s, \psi}(Senroll_i)$ 得到. 该算法通过指定 $\{z_s^i\}_i$ 的通道排序控制对应通道的信号输出.

利用 IP 算法和求导准则^[15], 易得分离矩阵和全局幅度因子的更新公式如下:

$$V_i(f) = \frac{1}{N} \sum_n \frac{x(f, n) x^H(f, n)}{g_i \sigma_{\theta}^2(f, n; z_c^i, z_s^i)} \quad (21)$$

$$w_i(f) = (W(f) V_i(f))^{-1} e_i \quad (22)$$

$$w_i(f) = \frac{w_i(f)}{\sqrt{w_i^H(f) V_i(f) w_i(f)}} \quad (23)$$

$$g_i = \frac{1}{FN} \sum_{f, n} \frac{|y_i(f, n)|^2}{\sigma_{\theta}^2(f, n; z_c^i, z_s^i)} \quad (24)$$

下面给出该算法的处理流程.

算法 指定输出通道排序的半监督分离算法

1. 利用式(16)所示的目标函数对变分自编码器的网络参数 θ, ϕ 和 ψ 进行第一阶段训练.

2. 固定内容编码器和话者编码器的网络参数 ϕ 和 ψ , 利用式(19)所示的目标函数对解码器的网络参数 θ 进行第二阶段训练.

3. 将各话者的注册信号输入话者编码器网络得到 $\{z_s^i\}_i$, 根据期望的通道输出顺序调整 i 的排列.

4. 对下列步骤进行循环迭代直至算法收敛:

(1) 通过反向传播更新 $\{z_c^i\}_i$;

(2) 根据式(24)更新全局幅度调整因子 $\{g_i\}_i$;

(3) 根据式(21)至式(23)更新所有频点的分离矩阵集合 W ;

5. 应用最小失真准则^[5]并根据式(7)得到分离后的信号.

3 仿 真

3.1 训练配置 VAE 网络第一阶段的训练使用 Librispeech^[37] 开源数据集 train-clean-100 和 train-clean-360 两部分中 100 位话者的纯净语料, 其中选用的男、女性话者的数量分别为 45 和 55. 训练数据和验证数据的总时长分别为 27 h 和 2.7 h, 分别包含 30000 段和 3000 段时长为 3.2 s 的训练样本.

在第二阶段, 通过数据增强的方式单独对解码器网络进行降噪训练, 增强的数据包括: (1) 带块排序干扰的含噪数据; (2) 带语音干扰的混合含噪数据. 含噪信号中目标信号和干扰信号均选自上述 100 位话者的语料集, 共生成 44 h 的训练数据, 其中 (1)(2) 两种含噪语音信号和纯净语音信号的比例为 2:2:1. 两种数据增强方式如下:

(1) 从目标信号 2 kHz 以上的频段中随机选取 1~3 段子频带, 将这些子频带的信号用某一干扰信号的对应频段信号代替, 得到人为构造的含噪信号.

(2) 随机选取一段干扰信号, 在时频域将其幅度和目标信号对应时频点的幅值加权相加得到人为构造的混合含噪信号, 即:

$$f(\bar{s}(f)) = \alpha(f) |\tilde{s}(f)| + (1 - \alpha(f)) |\bar{s}(f)|$$

其中, \tilde{s} 为干扰信号; $\alpha(f) \in (0, 1)$ 表示频点 f 处于干扰信号的权重. 该权重从随机生成的混合学生 t 分布曲线上采样获得, 并且使目标信号能量在 2 kHz 以下占主导地位.

上述两种增强方式将 2 kHz 以下的信号成分作为基准信号, 驱使网络依据该低频信息恢复其他频段的目标信号.

在训练时,首先对纯净训练数据进行静音段剪裁以及时域上的均值幅值归一化,之后将信号经过STFT变换后的幅度对数谱作为网络输入特征.信号的采样率为16 kHz,STFT的窗长和帧移分别为64 ms和16 ms,分析窗为汉宁窗.训练使用Adam优化器^[38],学习率设置为 $1e^{-4}$.第一阶段训练时设置权重参数 $\lambda_{\text{rec}}=10$, $\lambda_{\text{KL}}=1$,第二阶段训练时应用早停法^[39]使网络在验证集上的效果最优.

3.2 测试配置 本文使用来自MIRD^[40]的实录房间冲激响应(Room Impulse Response, RIR)数据集构造测试所用的混合信号,该数据集包含多种录制场景和传声器-声源参数配置.仿真中,选取的房间混响时间 RT_{60} 为0.160 s以及0.360 s,传声器的间距为8 cm,声源和传声器之间的距离为1 m.仿真考虑两声源两通道的场景,其中两个传声器选为MIRD数据集中的4,5两通道,两声源的波达方向(Direction of Arrival, DOA)间距为 $90^\circ, 105^\circ, 110^\circ$ 中随机选择.

该仿真考察了所提算法在训练时已见以及未见话者的测试数据集上的处理效果.对于已见话者的场景,声源信号来自训练时所用的100位话者的其他语料数据;对于未见话者的场景,声源信号来自Librispeech test-clean和dev-clean两个子数据集中的语料.对于每一种测试数据集,在两种混响时间下,针对男性-男性(Male-Male)话者、女性-女性(Female-Female)话者以及男性-女性(Male-Female)话者三种混合场景分别生成40段不同的混合信号,每段混合信号来自不同的话者对,混合信号的初始信扰比从-5,0,5 dB中均匀选择.对于训练时已见及未见话者的场景,测试时涵盖的男性以及女性话者的数量均为40.

3.3 结果与讨论 仿真针对本文提出的基于实例归一化的多通道变分自编码器算法(Instance Normalization based MVAE, IN-MVAE),对其在语音分离以及输出通道排序两个方面的效果进行了详细的测试和评估.对于语音分离效果,采用信号失真比提升量(Signal-to-Distortion Ratio Improvement, SDR_i)、信号干扰比提升量(Signal-to-Interference Ratio Improvement, SIR_i)以及短时客观可懂度(Short-Time Objective Intelligibility,

$STOI$)进行评测.其中, SDR_i 和 SIR_i 由BSS-EVAL工具箱、 $STOI$ 由pystoi工具箱分别计算得到.对于输出通道排序效果,考察所有测试样本下输出通道的排列顺序符合期望排列顺序的准确率.对于仿真中每一段样本,测试其所有可能的期望排序下算法的分离和排序性能.

首先对训练时已见话者的测试数据进行考察,采用的基线方法为1.3中介绍的MVAE算法.为了保证对比的严谨,MVAE算法网络训练使用的数据和IN-MVAE算法第一阶段使用的训练数据相同.值得注意的是,在MVAE算法的官方实践中^[28],其分离矩阵的初始化使用ILRMA算法^[14]迭代30次之后的结果,而对于本文提出的IN-MVAE算法,为了避免初始化带来的通道偏置影响,采用单位阵和反单位阵交替对各个频点分离矩阵进行初始化的方式.为此在对比分离性能时,同时考察了经ILRMA初始化的MVAE算法(记为MVAE-I)和用单位阵初始化的MVAE算法(记为MVAE-II)的效果.图3和图4分别展示了两种混响场景下各算法的分离性能的评测结果,其中每一种评价指标的得分为各组测试数据两通道结果的平均.可以看出,相比同样使用单位阵初始化的MVAE-II算法,IN-MVAE算法在各项评价指标上都有很大的提升,并在绝大多数场景下与利用ILRMA算法初始化的MVAE-I算法 SDR_i 和 SIR_i 的评分相差小于1 dB, $STOI$ 的得分相当.通过观察分离结果还可以发现,MVAE-II效果不佳的原因是出现了大量频域间块排序错误,而IN-MVAE算法由于对解码器网络进行了降噪训练,在很大程度上规避了该问题的发生,因而在评价指标上的得分更高.图4的结果表明,随着混响时间的增加,IN-MVAE算法在和基线方法的对比中仍然维持着稳定的分离性能.

表1展示了IN-MVAE算法与MVAE-I算法在通道排序准确度上的性能对比,其中MVAE-I算法根据算法收敛后得到的通道判别结果对输出通道的顺序进行重排.结果显示,由于MVAE-I算法几乎无法获得有效的判别结果,其排序准确度在绝大多数测试场景下不超过50%.与之相比,IN-MVAE算法在已见话者测试数据集上有着稳定和准确的通道排序能力.对于异性别话者

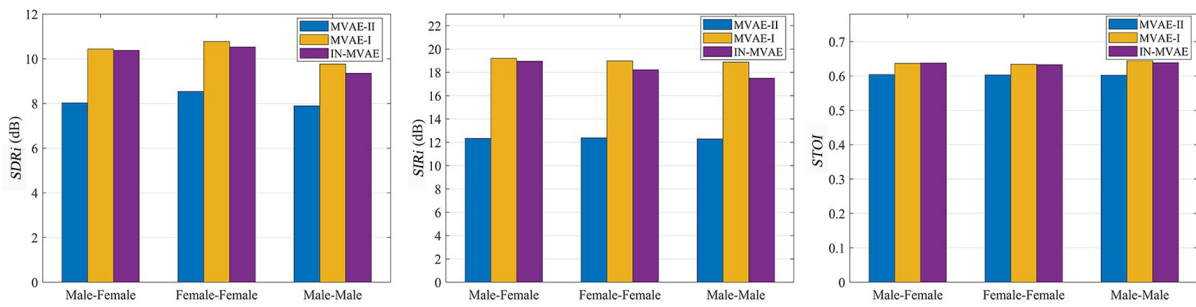
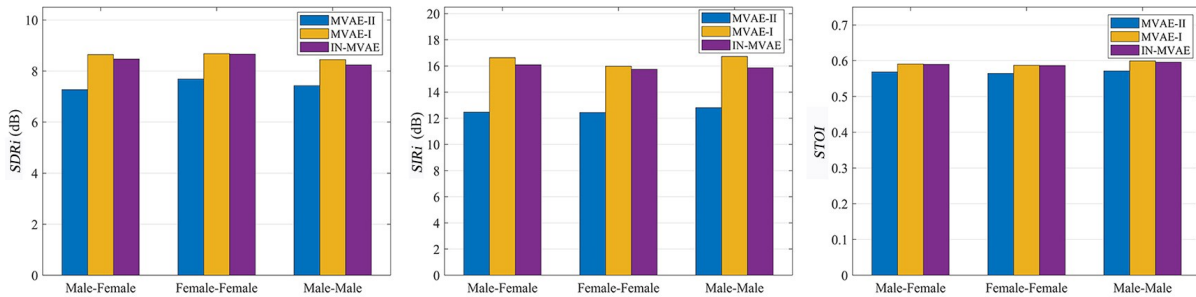
图3 RT_{60} 为0.160 s时各算法在已见话者测试数据上的平均 SDR_i , SIR_i 以及 $STOI$ 得分Fig.3 Averaged SDR_i , SIR_i and $STOI$ with seen speakers when $RT_{60}=0.160$ s图4 RT_{60} 为0.360 s时各算法在已见话者测试数据上的平均 SDR_i , SIR_i 以及 $STOI$ 得分Fig.4 Averaged SDR_i , SIR_i and $STOI$ with seen speakers when $RT_{60}=0.360$ s

表1 MVAE-I算法和IN-MVAE算法在已见话者的测试数据集上对输出通道排序的准确率

Table 1 Channel arrangement accuracy for MVAE-I and IN-MVAE with seen speakers

RT_{60} (s)	测试场景	MVAE-I算法 准确率 (%)	IN-MVAE算法 准确率 (%)
0.160	男性-女性	42.50	100.00
0.160	女性-女性	38.75	97.50
0.160	男性-男性	62.50	100.00
0.360	男性-女性	42.50	100.00
0.360	女性-女性	50.00	95.00
0.360	男性-男性	45.00	96.25

组合的情况,在两种测试的混响时间下,其排序准确度均达到100%;对于同性别话者,其在所有测试场景下的排序准确度均不小于95%。

其次考察IN-MVAE算法在训练时未见话者的混合信号上的分离和排序性能。考虑到MVAE算法受独热编码的限制,无法对训练时未见的话者进行通道判别,因此在排序评测时考察了IN-MVAE算法只经过第一阶段训练(记为IN-MVAE-I)以及同时经过一、二阶段训练(记为IN-

MVAE-II)的性能对比。图5和图6展示了混响时间分别为0.160 s和0.360 s时MVAE-I算法和上述两种IN-MVAE算法的分离效果。

可以看出,对于异性别话者场景,IN-MVAE-II的分离性能与MVAE-I算法相当;对于同性别话者场景,IN-MVAE-II与MVAE-I算法的 SDR_i 和 SIR_i 指标的差距比已见话者的测试有所扩大,但是对于大多数测试情景,该评分差距仍在1 dB之内。相比IN-MVAE-II算法,未经降噪训练的IN-MVAE-I算法在分离性能上有明显的弱化。二者结果的对比证明,通过数据增强的方式对解码器网络进行降噪训练的有效性,该训练提升了IN-MVAE-II算法的分离性能。

表2展示了经过和没有经过降噪训练的IN-MVAE算法在训练时未见话者的测试数据上的对输出通道进行排序的准确率。对于所有场景,在增加降噪训练后,输出通道的排序准确度都有一定的提升。结合表1的结果可知,IN-MVAE算法对异性别话者输出通道的排序能力有很高的鲁棒性,而对于同性别话者,排序准确度则会有一定的下降。这可能是由于解码器网络从同性别话者

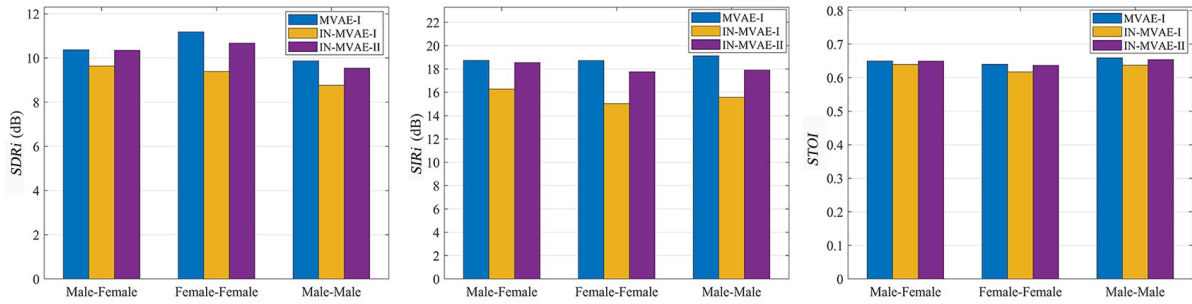
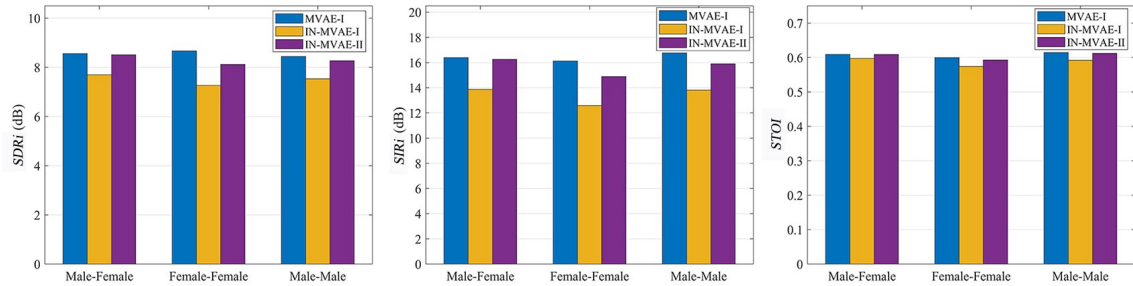
图5 RT_{60} 为 0.160 s 时各算法在未见话者测试数据上的平均 SDR_i , SIR_i 以及 $STOI$ 得分Fig.5 Averaged SDR_i , SIR_i and $STOI$ with unseen speakers when $RT_{60}=0.160$ s图6 RT_{60} 为 0.360 s 时各算法在未见话者测试数据上的平均 SDR_i , SIR_i 以及 $STOI$ 得分Fig.6 Averaged SDR_i , SIR_i and $STOI$ with unseen speakers when $RT_{60}=0.360$ s

表2 IN-MVAE-I算法和IN-MVAE-II算法在未见话者的测试数据集上对输出通道排序的准确率

Table 2 Channel arrangement accuracy for IN - MVAE-I and IN-MVAE-II with unseen speakers

RT_{60} (s)	测试场景	IN-MVAE-I算法 准确率 (%)	IN-MVAE-II算法 准确率 (%)
0.160	男性-女性	91.25	100.00
0.160	女性-女性	76.25	81.25
0.160	男性-男性	83.75	86.25
0.360	男性-女性	93.75	100.00
0.360	女性-女性	73.75	81.25
0.360	男性-男性	82.50	83.75

的话者编码向量中得到的全局信息较为相似,因而排序错误的信号不会对算法的代价函数产生较大的惩罚,导致算法收敛到了局部最优. 对于未见话者的场景,受限的话者编码器的泛化能力,同性别话者间的通道排序准确度会进一步下降. 此外,仿真结果还显示算法在女性-女性测试场景下的排序稳定性不如男性-男性话者的组合.

图7展示了在0.36 s的混响场景以及未见话者的测试数据下,话者注册语音的长度对IN-

MVAE算法排序准确率的影响. 从5 s长的注册语音开始,每次增加5 s的注册数据进行测试,直至注册语音的长度为30 s. 结果显示在注册语音不小于15 s的情况下,IN-MVAE算法在所有场景下均能获得超过80%的输出通道排序准确率.

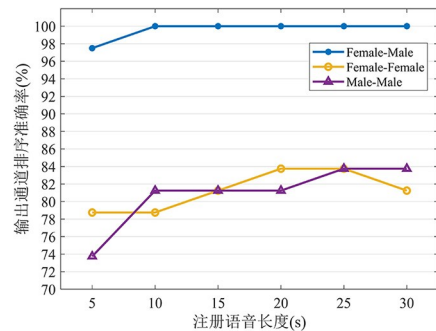


图7 不同的注册语音长度对IN-MVAE算法输出通道排序的准确率的影响

Fig. 7 Channel arrangement accuracy of IN - MVAE with different enrollment utterance lengths

4 结 论

针对频域盲源分离算法的全局排序问题,提出一种能够指定输出通道排列顺序的半监督盲源

分离算法. 算法以 MVAE 的算法框架为基础, 通过引入基于实例归一化和自适应实例归一化的变分自编码器作为声源模型, 解决原始的 MVAE 算法中可能发生的模型退化问题; 为了抑制分离结果中块排序错误的产生, 该方法利用人为构造的包含块排序错误和干扰信号的两种含噪数据对解码器网络参数进行第二阶段的降噪训练, 提升了算法的分离性能和对输出通道排序的稳定性. 最后, 实录房间冲激响应数据的仿真结果验证了该算法的分离性能以及其在训练时已见和未见话者数据集上对分离后输出通道排序的有效性.

参考文献

- [1] Rahbar K, Reilly J P. A frequency domain method for blind source separation of convolutive audio mixtures. *IEEE Transactions on Speech and Audio Processing*, 2005, 13(5): 832—844.
- [2] Mitianoudis N, Davies M E. Audio source separation of convolutive mixtures. *IEEE Transactions on Speech and Audio Processing*, 2003, 11(5): 489—497.
- [3] Nion D, Mokios K N, Sidiropoulos N D, et al. Batch and adaptive PARAFAC-based blind separation of convolutive speech mixtures. *IEEE Transactions on Audio, Speech, and Language Processing*, 2010, 18(6): 1193—1207.
- [4] Kim T, Eltoft T, Lee T W. Independent vector analysis: an extension of ICA to multivariate components//The 6th International Conference on Independent Component Analysis and Signal Separation. Springer Berlin Heidelberg, 2006: 165—172.
- [5] Lee I, Kim T, Lee T W. Independent vector analysis for convolutive blind speech separation//Makino S, Sawada H, Lee T W. Blind speech separation. Springer Berlin Heidelberg, 2007: 169—192.
- [6] Hyvärinen A, Oja E. Independent component analysis: algorithms and applications. *Neural Networks*, 2000, 13(4—5): 411—430.
- [7] Kang F, Yang F R, Yang J. A low-complexity permutation alignment method for frequency-domain blind source separation. *Speech Communication*, 2019(115): 88—94.
- [8] Sawada H, Mukai R, Araki S, et al. A robust and precise method for solving the permutation problem of frequency-domain blind source separation. *IEEE Transactions on Speech and Audio Processing*, 2004, 12(5): 530—538.
- [9] Liang Y, Chen G, Naqvi S M R, et al. Independent vector analysis with multivariate student's t -distribution source prior for speech separation. *Electronics Letters*, 2013, 49(16): 1035—1036.
- [10] Kitamura D, Mogami S, Mitsui Y, et al. Generalized independent low-rank matrix analysis using heavy-tailed distributions for blind source separation. *EURASIP Journal on Advances in Signal Processing*, 2018: 28.
- [11] Gu Z Y, Lu J, Chen K. Speech separation using independent vector analysis with an amplitude variable Gaussian mixture model//The 20th Annual Conference of the International Speech Communication Association. Graz, Austria: ISCA, 2019: 1358—1362.
- [12] Lee D D, Seung H S. Learning the parts of objects by non-negative matrix factorization. *Nature*, 1999, 401(6755): 788—791.
- [13] Sawada H, Ono N, Kameoka H, et al. A review of blind source separation methods: Two converging routes to ILRMA originating from ICA and NMF. *APSIPA Transactions on Signal and Information Processing*, 2019(8): e12.
- [14] Kitamura D, Ono N, Sawada H, et al. Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2016, 24(9): 1626—1641.
- [15] Ono N. Stable and fast update rules for independent vector analysis based on auxiliary function technique//The 12th IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. New Paltz, NY, USA: IEEE, 2011: 189—192.
- [16] Sekiguchi K, Nugraha A A, Bando Y, et al. Fast multichannel source separation based on jointly diagonalizable spatial covariance matrices//The 27th European Signal Processing Conference. A Coruna, Spain: IEEE, 2019: 1—5.
- [17] Kubo Y, Takamune N, Kitamura D, et al. Efficient full-rank spatial covariance estimation using independent low-rank matrix analysis for blind source

- separation//The 27th European Signal Processing Conference. A Coruna, Spain: IEEE, 2019: 1–5.
- [18] Sekiguchi K, Bando Y, Nugraha A A, et al. Fast multichannel nonnegative matrix factorization with directivity - aware jointly - diagonalizable spatial covariance matrices for blind source separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2020(28): 2610–2625.
- [19] Janský J, Málek J, Čmejla J, et al. Adaptive blind audio source extraction supervised by dominant speaker identification using X - vectors//2020 IEEE International Conference on Acoustics, Speech and Signal Processing. Barcelona, Spain: IEEE, 2020: 676–680.
- [20] Nesta F, Koldovský Z. Supervised independent vector analysis through pilot dependent components//The 42th IEEE International Conference on Acoustics, Speech and Signal Processing. New Orleans, LA, USA: IEEE, 2017: 536–540.
- [21] Mitsui Y, Takamune N, Kitamura D, et al. Vectorwise coordinate descent algorithm for spatially regularized independent low - rank matrix analysis//2018 IEEE International Conference on Acoustics, Speech and Signal Processing. Calgary, Canada: IEEE, 2018: 746–750.
- [22] Li L, Koishida K. Geometrically constrained independent vector analysis for directional speech enhancement//2020 IEEE International Conference on Acoustics, Speech and Signal Processing. Barcelona, Spain: IEEE, 2020: 846–850.
- [23] Brendel A, Haubner T, Kellermann W. A unified probabilistic view on spatially informed source separation and extraction based on independent vector analysis. *IEEE Transactions on Signal Processing*, 2020(68): 3545–3558.
- [24] Žmolíková K, Delcroix M, Kinoshita K, et al. SpeakerBeam: Speaker aware neural network for target speaker extraction in speech mixtures. *IEEE Journal of Selected Topics in Signal Processing*, 2019, 13(4): 800–814.
- [25] Li G J, Liang S, Nie S, et al. Direction-aware speaker beam for multi-channel speaker extraction//The 20th Annual Conference of the International Speech Communication Association. Graz, Austria: ISCA, 2019: 2713–2717.
- [26] Delcroix M, Ochiai T, Zmolikova K, et al. Improving speaker discrimination of target speech extraction with time - domain speakerbeam//2020 IEEE International Conference on Acoustics, Speech and Signal Processing. Barcelona, Spain: IEEE, 2020: 691–695.
- [27] Wang Q, Muckenhirn H, Wilson K, et al. VoiceFilter: Targeted voice separation by speaker - conditioned spectrogram masking. 2019, arXiv: 1810.04826.
- [28] Kameoka H, Li L, Inoue S, et al. Supervised determined source separation with multichannel variational autoencoder. *Neural Computation*, 2019, 31(9): 1891–1914.
- [29] Kingma D P, Rezende D J, Mohamed S, et al. Semi-supervised learning with deep generative models//Proceedings of the 27th International Conference on Neural Information Processing Systems. Cambridge, MA, USA: MIT Press, 2014: 3581–3589.
- [30] Li L, Kameoka H, Makino S. Fast MVAE: joint separation and classification of mixed sources based on multichannel variational autoencoder with auxiliary classifier//2019 IEEE International Conference on Acoustics, Speech and Signal Processing. Brighton, UK: IEEE, 2019: 546–550.
- [31] Du Y C, Sekiguchi K, Bando Y, et al. Semi-supervised multichannel speech separation based on a phone- and speaker-aware deep generative model of speech spectrograms//2020 28th European Signal Processing Conference. Amsterdam, Netherlands: IEEE, 2021: 870–874.
- [32] Chou J C, Yeh C C, Lee H Y. One-shot voice conversion by separating speaker and content representations with instance normalization//The 20th Annual Conference of the International Speech Communication Association. Graz, Austria: ISCA, 2019: 664–668.
- [33] Ulyanov D, Vedaldi A, Lempitsky V. Instance normalization: the missing ingredient for fast stylization. 2017, arXiv: 1607.08022v3.
- [34] Huang X, Belongie S. Arbitrary style transfer in real-time with adaptive instance normalization//2017 International Conference on Computer Vision. Venice, Italy: IEEE, 2017: 1510–1519.

- [35] Itahashi T, Matsuoka K. Stability of independent vector analysis. *Signal Processing*, 2012, 92(8): 1809—1820.
- [36] Kingma D P, Welling M. Auto-encoding variational bayes. 2013, arXiv:1312.6114.
- [37] Panayotov V, Chen G G, Povey D, et al. Librispeech: an ASR corpus based on public domain audio books//2015 IEEE International Conference on Acoustics, Speech and Signal Processing. South Brisbane, Australia: IEEE, 2015: 5206—5210.
- [38] Kingma D P, Ba J. Adam: a method for stochastic optimization. 2014, arXiv:1412.6980v1.
- [39] Yao Y, Rosasco L, Caponnetto A. On early stopping in gradient descent learning. *Constructive Approximation*, 2007, 26(2): 289—315.
- [40] Hadad E, Heese F, Vary P, et al. Multichannel audio database in various acoustic environments//The 14th International Workshop on Acoustic Signal Enhancement. Juan-les-Pins, France: IEEE, 2014: 313—317.
- [41] Vincent E, Gribonval R, Fevotte C. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 2006, 14(4): 1462—1469.
- [42] Taal C H, Hendriks R C, Heusdens R, et al. A short-time objective intelligibility measure for time-frequency weighted noisy speech//2010 IEEE International Conference on Acoustics, Speech and Signal Processing. Dallas, TX, USA: IEEE, 2010: 4214—4217.

(责任编辑 杨可盛)