ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 39,No. 5,Sep. 2024,pp. 1044－1061
DOI:10. 16337/j. 1004-9037. 2024. 05. 002

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

多说话人分离与目标说话人提取的研究现状与展望

鲍长春，杨　雪

(北京工业大学信息科学技术学院语音与音频信息处理研究所,北京 100124)

摘　要:语音分离作为语音信号处理领域的前沿技术,具有重要的研究价值和广阔的应用前景。通常,麦克风拾取的信号包含有多个说话人的语音、噪声和混响。为了提升用户的听觉体验以及后端设备的处理性能,需要对混合信号进行语音分离。语音分离起源于著名的鸡尾酒会问题,旨在从混合信号中分离出说话人的语音信号。近年来,研究人员提出了大量的语音分离方法,显著提升了分离性能。本文对这些语音分离方法进行了系统的归纳和总结。首先,根据目标说话人的辅助信息利用与否,将语音分离方法分为两大类,即多说话人分离与目标说话人提取;其次,从传统到基于深度学习的角度,分别对多说话人分离和目标说话人提取两类方法进行详细介绍;最后,讨论了当前语音分离领域面临的一些挑战,并对未来的研究方向进行展望。

关键词:语音分离;鸡尾酒会问题;多说话人分离;目标说话人提取;深度学习

中图分类号:TN912.3; TP183　　**文献标志码**:A

Research Situation and Prospects of Multi-speaker Separation and Target Speaker Extraction

BAO Changchun, YANG Xue

(Institute of Speech and Audio Information Processing, School of Information Science and Technology, Beijing University of Technology, Beijing 100124, China)

Abstract: As a cutting-edge technology in speech signal processing, speech separation has significant research value and broad application prospects. Typically, the signal captured by the microphones contains speech signals from multiple speakers, noise and reverberation. To improve the user experience and the performance of backend devices, it is necessary to perform speech separation. Speech separation originated from the well-known cocktail party problem. It aims to separate the speech signals from the mixed signal. In recent years, researchers have proposed a large number of speech separation methods, which have significantly improved separation performance. This paper systematically reviews and summarizes these methods. First, based on whether the auxiliary information of the target speaker is leveraged, speech separation is divided into two categories, i. e., multi-speaker separation and target speaker extraction. Second, these methods are introduced in detail, following the progression from conventional approaches to deep learning-based techniques. Finally, the existing challenges in speech separation are discussed and prospective research in the future are highlighted.

Key words: speech separation; cocktail party problem; multi-speaker separation; target speaker extraction; deep learning

---

基金项目:国家自然科学基金(61831019)。
收稿日期:2024-06-27;**修订日期**:2024-08-15

引　言

语音是人们交流沟通最重要、最自然的方式之一。近年来,随着语音信号处理技术的日益成熟和新一代信息技术的飞速发展,语音的应用范围与应用前景也变得愈加广阔,多种语音产品现已涉足实时通信、个人助手、智能安防和电商零售等不同领域。在不同的应用场景中,麦克风拾取到的信号通常包含有多个说话人的语音、噪声与混响,而用户或设备往往只需要其中的一个或多个语音信号,因此,需要进行语音分离[1-5],即从混合信号中分离出一个或多个说话人的语音信号。

语音分离的研究已有半个多世纪的历史,早在1953年,英国的认知科学家Colin Cherry就提出了著名的"鸡尾酒会问题"(Cocktail party problem, CPP)[6],即在参加鸡尾酒会时,即便所处的声学场景中充斥着交谈声、背景音乐以及餐具碰撞声等,人们依旧能够听清并理解当前所关注的说话人讲述的内容,并且可以将注意力在不同说话人间进行切换。虽然人们可以轻而易举地处理鸡尾酒会问题,但长时间处于嘈杂环境中极易使人产生听觉疲劳。此外,现有技术尚无法有效地处理鸡尾酒会问题,即无法选择性地听取不同说话人的语音,为此需要对混合信号进行语音分离,从而为后续处理语音信号提供可能。

作为语音信号处理中的一项前端技术,语音分离具有极为重要的研究意义以及十分广阔的应用前景。在语音识别领域,随着智能家居、便携式设备和车载语音控制系统的涌现,人机语音交互变得日益常态化,为人们的日常生活与工作带来了许多便利。然而,实际应用场景中,多个说话人语音的重叠、噪声与混响的存在将导致智能设备的识别性能显著降低,严重影响用户的使用体验,而利用语音分离技术,可以大大提高复杂声学场景中智能语音设备的识别率,达到提升系统性能的目的,使其能够更好地为人们提供服务。在通信领域,为提高传输速率,往往需要从语音信号中提取参数,并对所提参数进行编码传输,然而多个说话人语音的重叠以及噪音与混响的存在可能导致参数无法准确提取,最终影响编解码语音的质量。经过语音分离处理后,语音信号的质量以及可懂度将大大提高,从而可以有效减轻用户的听觉疲劳。此外,对于听力障碍的患者而言,语音分离显得尤为重要,只有依靠助听设备对环境噪声和干扰声源的抑制,听力障碍患者才能获得较好的听感,从而与他人进行较为顺畅的沟通交流,减弱患者认知能力和言语能力下降等不利影响。

本文对现有语音分离方法进行了归纳总结。首先,本文对语音分离的基本概念进行阐述,并根据目标说话人的辅助信息利用与否,将其分为多说话人分离和目标说话人提取两类;其次,从传统到深度学习的角度,分别对这两类方法进行详细介绍;最后,本文对语音分离的现有挑战进行讨论,并对未来的研究方向进行展望。

1　语音分离的基本概念

实际场景中,麦克风拾取到的信号通常包含有多个说话人的语音信号、噪声与混响。以单个麦克风为例,其拾取的混合信号可表示为

$$\boldsymbol{y}=\begin{cases}\sum_{i=1}^{S}\boldsymbol{a}_i*\boldsymbol{x}_i+\boldsymbol{n}=\sum_{i=1}^{S}\boldsymbol{z}_i+\boldsymbol{n} & \nexists \boldsymbol{A}_{\text{tgt}}\\ \boldsymbol{a}_{\text{tgt}}*\boldsymbol{x}_{\text{tgt}}+\sum_{l=1}^{S-1}\boldsymbol{a}_l*\boldsymbol{x}_l+\boldsymbol{n}=\boldsymbol{z}_{\text{tgt}}+\sum_{l=1}^{S-1}\boldsymbol{z}_l+\boldsymbol{n} & \exists \boldsymbol{A}_{\text{tgt}}\end{cases} \tag{1}$$

式中:$\boldsymbol{y}$为麦克风拾取的混合信号;$\boldsymbol{a}_i$为第i个说话人与麦克风之间的房间冲击响应;$\boldsymbol{x}_i$为第i个说话人的纯净语音信号;$\boldsymbol{n}$为加性噪声;$\boldsymbol{z}_i$为麦克风拾取的第i个说话人的语音信号;S为混合语音中的说话人数量;符号"*"代表卷积运算;$\boldsymbol{A}_{\text{tgt}}$为目标说话人的辅助信息;符号"∄"表示该辅助信息不存在;$\boldsymbol{a}_{\text{tgt}}$为目标说话人与麦克风之间的房间冲击响应;$\boldsymbol{x}_{\text{tgt}}$为目标说话人的纯净语音信号;$\boldsymbol{a}_l$为第$l$个非目标说话人与麦克

风之间的房间冲击响应；$\boldsymbol{x}_l$为第l个非目标说话人的纯净语音信号；$\boldsymbol{z}_{\text{tgt}}$为麦克风拾取的目标说话人的语音信号；$\boldsymbol{z}_l$为麦克风拾取的第$l$个非目标说话人的语音信号；符号“∃”表示目标说话人的辅助信息存在。

语音分离旨在从混合信号中分离出说话人的语音信号。根据目标说话人的辅助信息利用与否，语音分离可以分为两类，即多说话人分离（Multi-speaker separation，MSS）与目标说话人提取（Target speaker extraction，TSE）。图1给出了两类语音分离的示意图，其中，图1(a)为多说话人分离，它无法利用目标说话人的信息，因此需要将多个说话人的语音信号从混合语音信号中分离开来。多说话人分离的数学模型可表示为

$$\left\{\hat{\boldsymbol{x}}_i \middle| i=1,2,\cdots,S\right\}=\{\hat{\boldsymbol{x}}_1,\hat{\boldsymbol{x}}_2,\cdots,\hat{\boldsymbol{x}}_S\}=M_{\text{MSS}}(\boldsymbol{y})=M_{\text{MSS}}\left(\sum_{i=1}^{S}\boldsymbol{z}_i+\boldsymbol{n}\right) \tag{2}$$

式中：$\hat{\boldsymbol{x}}_i$为分离得到的第i个说话人的语音信号；$\{\hat{\boldsymbol{x}}_i|i=1,2,\cdots,S\}$为分离得到的$S$个说话人语音信号的集合；$M_{\text{MSS}}(\cdot)$表示多说话人分离对应的映射关系，即将混合信号$\boldsymbol{y}$映射为多个说话人的语音信号$\{\hat{\boldsymbol{x}}_i|i=1,2,\cdots,S\}$。多说话人语音分离的典型应用场景包括多说话人语音识别和说话人日志等。

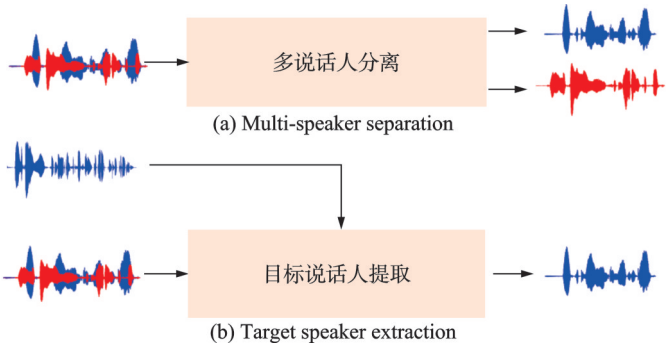

图1 两类语音分离示意图

Fig.1 Schematic diagrams of two kinds of speech separation

图1(b)为目标说话人提取，与多说话人分离不同，目标说话人提取可以利用目标说话人的辅助信息，因此只需将目标说话人的语音信号从混合语音信号中分离出来。目标说话人提取的数学模型可表示为

$$\hat{\boldsymbol{x}}_{\text{tgt}}=M_{\text{TSE}}\left(\boldsymbol{y}\middle|\boldsymbol{A}_{\text{tgt}}\right)=M_{\text{TSE}}\left(\left(\boldsymbol{z}_{\text{tgt}}+\sum_{l=1}^{S-1}\boldsymbol{z}_l+\boldsymbol{n}\right)\middle|\boldsymbol{A}_{\text{tgt}}\right) \tag{3}$$

式中：$\hat{\boldsymbol{x}}_{\text{tgt}}$为提取到的目标说话人语音信号；$M_{\text{TSE}}(\cdot)$表示目标说话人提取对应的映射关系，即利用目标说话人的辅助信息$\boldsymbol{A}_{\text{tgt}}$，将混合信号$\boldsymbol{y}$映射为目标说话人的语音信号$\hat{\boldsymbol{x}}_{\text{tgt}}$。目标说话人提取的典型应用场景包括目标说话人语音识别、个性化用户体验和说话人情感分析等。

2 多说话人分离

自“鸡尾酒会问题”提出以来，多说话人分离的研究已有半个多世纪的历史，研究人员提出了许多不同类型的方法。根据是否使用深度神经网络，多说话人分离方法可以分为传统多说话人分离方法与基于深度学习的多说话人分离方法两类。

2.1 传统多说话人分离方法

根据分离原理不同，传统多说话人分离方法主要有计算听觉场景分析、隐马尔可夫模型、非负矩阵分解和盲源分离。在这些方法中，通常假定说话人数量是已知的。

2.1.1 基于计算听觉场景分析的方法

计算听觉场景分析（Computational acoustic scene analysis，CASA）[7]的核心思想是利用计算的方式模拟人类听觉感知的过程，即模拟听觉系统对混合信号的分割与重组。计算听觉场景分析同样可分为分割阶段与重组阶段，其基本框图如图2所示。分割阶段包括听觉外围模块、特征提取模块与中层表示模块，而重组阶段对应场景组织模块。在分割阶段，听觉外围模块用于模拟基底膜和听神经系统的工

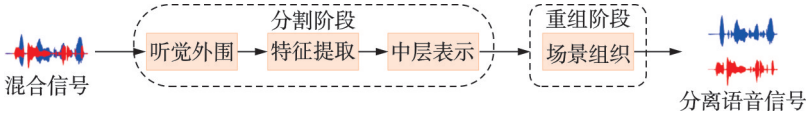

图2 计算听觉场景分析的基本框图

Fig.2 Diagram of computational acoustic scene analysis

作机理,利用耳蜗模型实现了从一维混合信号到二维时频谱的变换;特征提取模块从二维时频谱中提取基频、起始与终止等线索;中层表示模块利用提取到的线索将混合信号分解为由时频点组成的片段,每一片段对应着某个声学事件在听觉场景中的局部描述,不同的分割方式(如周期性、平滑性等)可以联合起来对时频谱进行分割[8]。在重组阶段,场景组织模块将来自同一声源的片段重组到一起形成听觉流,听觉流对应着声学事件在听觉场景中的整体描述,组织过程可分为同时组织与顺序组织。其中,同时组织将处于同一时刻但不同频率范围的片段组织到听觉流中,而顺序组织将片段按照时间先后顺序依次组织到听觉流中[9]。

在处理混合信号时,计算听觉场景分析模拟了人类听觉系统的处理机制,通过提取基于听觉感知的特征,使其在分离过程中展现出良好的直观性与可解释性,且该方法无需依赖大量标注数据进行训练。然而,计算听觉场景分析通常需要进行复杂的时频分析和特征提取,并包含多个独立的处理模块,这使得系统难以进行整体优化。此外,该方法主要依赖于低层次的听觉感知特征,难以有效利用信号中的高层次语义信息进行分离,从而在某些复杂场景中表现不足。

2.1.2 基于隐马尔可夫模型的方法

阶乘隐马尔可夫模型(Factorial hidden Markov model, FHMM)[10]是一种应用于多说话人分离的经典模型,它是具有多个马尔可夫链的隐马尔可夫模型。隐马尔可夫模型的观测值与状态变量间不存在简单的对应关系,而是通过概率相联系。因此,隐马尔可夫模型中包含两个随机过程:第1个是马尔可夫链,用以描述状态间的转移;第2个是各状态的输出,用以描述状态与观测值间的统计对应关系。在隐马尔可夫模型中,状态间的转移是隐式的,对于某一观测序列,只能以概率的形式推断其所属的状态,即状态将以随机的方式对外产生观测值。图3给出了隐马尔可夫模型的示意图,其中,图3(a)为基本隐马尔可夫模型,图中h^k为k时刻的状态变量,y^k为k时刻的观测值。相比于该基本的隐马尔可夫模型,图3(b)所示的阶乘隐马尔可夫模型利用了更为复杂的状态变量提升其表征能力,即模型的观测值与多个状态变量相联系。若将阶乘隐马尔可夫模型用于多说话人分离,则多个说话人可分别用多个隐马尔可夫模型进行建模,不同说话人对应的状态变量将同时影响观测值[11]。

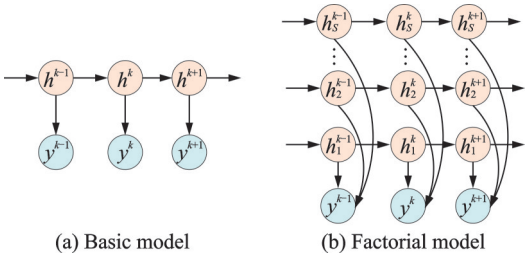

图3 隐马尔可夫模型

Fig.3 Hidden Markov model

在处理混合信号时,阶乘隐马尔可夫模型能够对多个说话人进行同时建模。作为一种概率模型,阶乘隐马尔可夫模型为分离过程提供了清晰的统计解释。然而,由于阶乘隐马尔可夫模型需要同时处理多个隐马尔可夫模型,且每个隐马尔可夫模型都包含多个状态与状态转移概率,这使得其计算复杂度较高。此外,阶乘隐马尔可夫模型的分离性能高度依赖于对模型参数的准确估计,且该模型难以利用高层次的语义信息。

2.1.3 基于非负矩阵分解的方法

非负矩阵分解(Non-negative matrix factorization, NMF)[12]的基本思想是将由混合信号得到的混合

矩阵 V 分解为两个非负矩阵 W 与 H 相乘的形式，即

$$V=WH \quad W\geqslant 0, H\geqslant 0 \tag{4}$$

式中：W 称为基矩阵；H 称为系数矩阵（也称为激活矩阵）。由于基矩阵 W 与系数矩阵 H 均为非负矩阵（矩阵的各元素均取非负值），故矩阵 V 也为非负矩阵（又称为非负语谱）。因此，非负矩阵分解的对象通常选为幅度谱或功率谱。此外，非负矩阵分解中不存在减法运算，是基于部分的表示，即基矩阵中的向量是一些基向量，通过对这些基向量进行不同的线性叠加，即可得到不同的混合矩阵。非负矩阵分解通常需要利用混合信号求解基矩阵 W 与系数矩阵 H，故可写为

$$\min_{W,H} D(V|WH) \quad W\geqslant 0, H\geqslant 0 \tag{5}$$

式中 $D(\cdot|\cdot)$ 为后者相较于前者的失真程度。同时求解基矩阵 W 与系数矩阵 H 十分困难，通常利用迭代优化算法对两矩阵进行迭代求解。值得注意的是，在非负矩阵分解中，由于分解对象为非负矩阵，故通常不考虑相位信息。此外，使用不同的失真测度、约束条件与先验假设，可以得到不同的非负矩阵分解方法[13-14]。

图4给出了两种情形下利用NMF进行多说话人分离的原理框图。当说话人的纯净语音已知时，NMF可以利用这些纯净语音信号训练得到对应说话人的基矩阵 W_i，从而构成基矩阵 $W=[W_1, W_2, \cdots, W_S]$，即基矩阵是已知的，只需利用混合语音信号求解系数矩阵 H 即可，如图4(a)所示；当部分说话人的纯净语音已知时，非负矩阵分解则利用这些部分已知的纯净语音信号训练说话人基矩阵，从而得到部分基矩阵 $[W_1, W_2, \cdots, W_i]$，此时，需要利用混合语音信号求解基矩阵的未知部分 $[W_{i+1}, W_{i+2}, \cdots, W_S]$ 与系数矩阵 H，如图4(b)所示。

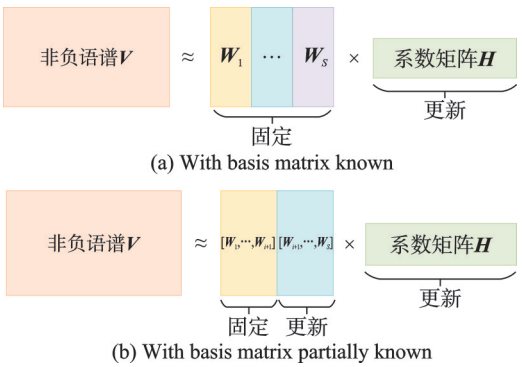

图4 两种情形下利用NMF进行多说话人分离的原理图

Fig.4 Two multi-speaker separation diagrams using NMF

在处理混合信号时，NMF方法将混合信号分解为非负的基矩阵和系数矩阵。该分解方式使得每个分量具有实际的物理意义，如基矩阵可以表示不同说话人的频谱特征，而系数矩阵则表示不同说话人在不同时刻的激活程度。此外，非负矩阵分解作为一种无监督学习方法，不需要用预先标注的数据进行训练，且其迭代更新算法易于实现。然而，非负矩阵分解的结果高度依赖于初始基矩阵和系数矩阵的选择。如果初始值选择不佳，算法可能会陷入局部最优解，从而无法找到全局最优解，进而影响分离的效果。此外，NMF主要依赖于低级别的频谱特征，难以捕捉到语音信号中的复杂语义信息，这限制了它在复杂场景中的应用。

2.1.4 基于盲源分离方法

盲源分离方法（Blind source separation, BSS）是一类十分重要的多说话人分离方法。其中，“盲”指混合过程与各个声源均是未知的，即只能根据声源信号的统计特性，利用观测到的混合信号来分离出不同的声源信号。独立成分分析（Independent component analysis, ICA）[15-16]是最早出现的一种盲源分离方法，其基本假设条件是多说话人间具有独立性。早期的独立成分分析考虑麦克风数目与说话人数目一致的情况，多说话人的语音是瞬时混合的且混合信号中不包含噪声，考虑任意时刻 k，混合信号可表示为

$$\begin{cases} y_1(k)=b_{11}x_1(k)+b_{12}x_2(k)+\cdots+b_{1S}x_S(k) \\ \vdots \\ y_j(k)=b_{j1}x_1(k)+b_{j2}x_2(k)+\cdots+b_{jS}x_S(k) \\ \vdots \\ y_S(k)=b_{S1}x_1(k)+b_{S2}x_2(k)+\cdots+b_{SS}x_S(k) \end{cases} \tag{6}$$

式中:$y_j(k)$为第j个麦克风在k时刻拾取的混合信号;b_{ji}为混合系数;$x_i(k)$为第i个说话人在k时刻的语音信号。式(6)的矩阵形式为

$$\boldsymbol{y}^k=\boldsymbol{B}\boldsymbol{x}^k \tag{7}$$

式中:$\boldsymbol{y}^k=[y_1(k),y_2(k),\cdots,y_S(k)]^{\mathrm{T}}$为$k$时刻混合信号构成的矢量;$\boldsymbol{B}$为混合矩阵;$\boldsymbol{x}^k=[x_1(k),x_2(k),\cdots,x_S(k)]^{\mathrm{T}}$为$k$时刻多说话人语音信号构成的矢量,上标"T"代表矩阵的转置运算。

独立成分分析利用$\boldsymbol{x}^k$各分量间的统计独立性,从混合信号矢量$\boldsymbol{y}^k$中求解混合矩阵$\boldsymbol{C}$,以分离多说话人的语音信号,其基本原理如图5所示,数学表达式可写为

$$\hat{\boldsymbol{x}}^k=\boldsymbol{C}\boldsymbol{y}^k \tag{8}$$

式中$\hat{\boldsymbol{x}}^k$为分离的多说话人语音信号构成的矢量。独立成分分析具有两个不确定性:一是分离的语音信号排列顺序具有不确定性;二是分离的语音信号尺度具有不确定性。

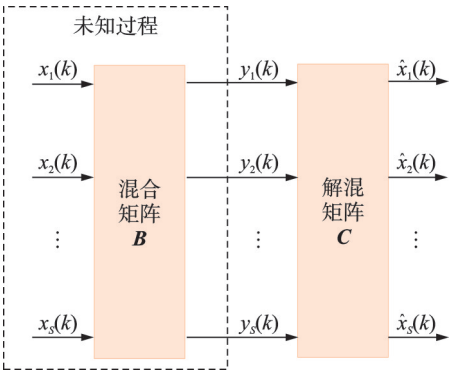

图5 ICA的原理框图

Fig.5 Diagram of the ICA

常用的ICA方法主要基于最大化非高斯性准则、极大似然准则和最小化互信息准则。在基于最大化非高斯性准则的ICA中,需假定各独立成分(即不同说话人语音信号)服从非高斯分布,该方法以中心极限定理为依据,认为多个独立成分的和(即混合语音信号)比各独立成分更接近于高斯分布;若对混合语音矢量$\boldsymbol{y}^k$进行变换得到$\boldsymbol{c}^{\mathrm{T}}\boldsymbol{y}^k$,则使得$\boldsymbol{c}^{\mathrm{T}}\boldsymbol{y}^k$的非高斯性最大的$\boldsymbol{c}$即为解混矩阵$\boldsymbol{C}$中的列向量,对应的$\boldsymbol{c}^{\mathrm{T}}\boldsymbol{y}^k$即为某一独立成分的估计,其中,非高斯性的度量方式有高阶累积量[17]和负熵[18]等。在基于极大似然准则的ICA中,需写出似然函数表达式,但独立成分的概率密度函数通常是未知的,通常需采用近似的概率密度函数进行求解[19]。基于最小化互信息准则的ICA[20]则从信息论的角度求解盲源分离问题,此方法中各成分的独立性假设并不严格成立,而互信息可看作是各成分间相关性的测度,使互信息最小化等价于使各成分间的相关性最小化。此外,式(6)中的信号模型并未考虑噪声和混响等因素,故其适用范围有限。为将独立成分分析应用于更加复杂的声学场景中,研究人员提出了诸多改进方法,如频域独立成分分析[21]。

独立矢量分析(Independent vector analysis, IVA)[22]是ICA的扩展。多个麦克风拾取到的混合信号在时频域可以表示为

$$\boldsymbol{y}^{(k,f)}=\boldsymbol{B}^{(f)}\boldsymbol{x}^{(k,f)} \tag{9}$$

式中:$\boldsymbol{y}^{(k,f)}=[y_1{}^{(k,f)},y_2{}^{(k,f)},\cdots,y_S{}^{(k,f)}]^{\mathrm{T}}$是由各个麦克风在第$k$帧、第$f$个频点处的混合信号构成的矢量;$\boldsymbol{B}^{(f)}$为第$f$个频点处的混合矩阵;$\boldsymbol{x}^{(k,f)}=[x_1{}^{(k,f)},x_2{}^{(k,f)},\cdots,x_S{}^{(k,f)}]^{\mathrm{T}}$是由不同说话人在第$k$帧、第$f$个频点处的语音信号构成的矢量。此外,将第$i$个说话人在第$k$帧的不同频点处的取值构成的矢量,记为$\boldsymbol{x}_i{}^{(k)}=[x_i{}^{(k,1)},x_i{}^{(k,2)},\cdots,x_i{}^{(k,F)}]^{\mathrm{T}}$,其中,$F$为频点数。在IVA中,通常假定矢量$\boldsymbol{x}^{(k,f)}$的各分量间相互独立,但矢量$\boldsymbol{x}_i{}^{(k)}$的各分量间是非独立的。相比于ICA,IVA可以利用不同频点间的关系,因此实现了更好的分离性能。

在处理混合信号时,盲源分离方法无需先验知识,即可在不知道混合过程和声源信号的情况下进

行分离，并能提供明确的统计解释。然而，盲源分离方法通常需要使用多个麦克风，且对初始参数估计、噪声和混响较为敏感。此外，该方法难以利用高层次的语义信息进行分离。

2.2 基于深度学习的多说话人分离方法

近几年，深度学习技术被广泛应用于多说话人分离领域。基于深度学习的多说话人分离方法取得了优异的分离性能，极大地推动了该领域的发展。与传统多说话人分离方法相比，基于深度学习的多说话人分离方法充分利用了神经网络强大的建模能力，不仅使得模型能够对低层次的特征进行建模，还可以学习到信号中包含的高层次语义信息。此外，通过使用数据增扩技术，该类方法可以扩大训练集，从而使模型能够适应各种复杂的分离场景，如不同噪声类型、不同信噪比和不同混响强度等。然而，基于深度学习的多说话人分离方法往往需要大量的数据进行训练，且模型的计算复杂度较高，在模型部署与实际应用时，通常需在性能与资源消耗之间进行折中考虑。

现有基于深度学习的多说话人分离方法大多使用有监督学习的方式训练模型，即需要为模型提供多个说话人的纯净语音作为标签，为实现模型的有效训练，需解决标签排列问题，即模型的多个输出与不同说话人语音如何对应的问题，如图6所示。

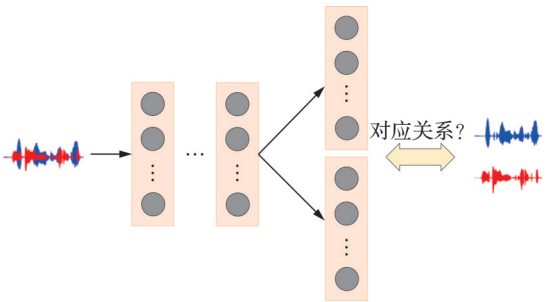

图6 标签排列问题示意图

Fig.6 Schematic diagram of label permutation problem

此外，基于深度学习的多说话人分离方法还需考虑说话人数目未知问题，即混合信号中包含的说话人数目往往是未知的，从而难以确定模型的输出通道数目。下面将分别对标签排列问题的解决方式、说话人数目已知情况下与说话人数目未知情况下的多说话人分离方法进行介绍。

2.2.1 标签排列问题的解决方式

目前，标签排列问题的解决方式大致可分为3类，分别为深度聚类(Deep clustering, DPCL)方式[23-24]、排列不变训练(Permutation invariant training, PIT)方式[25-26]与基于位置训练(Location-based training, LBT)方式[27-28]。

DPCL方式源自于计算听觉场景分析，该方式对不同说话人的时频掩蔽进行估计。具体而言，DPCL方式利用深度神经网络将每一个时频点映射为高维空间的一个向量，且归属于同一说话人的时频点对应的高维向量相似，归属于不同说话人的时频点对应的高维向量相异，通过对这些学习到的高维向量进行聚类(相似度计算)，得到对应不同说话人的时频掩蔽。

相应地，为便于解释，图7给出了解决标签排列问题的PIT和LBT方式，其中PIT方式如图7(a)所示，其在计算深度神经网络损失函数时解决了网络输出与不同说话人语音的对应问题。具体而言，PIT方式将网络的输出与不同说话人的语音进行配对，分别计算不同对应方案的损失函数，若有S个不同说话人，则需计算$S!$次损失函数，并选取其中的最小值作为最终的损失函数。由于PIT方式本质上是一种训练技巧，并不依赖于模型的输入、输出以及目标函数的具体形式，因此它可以应用于任何一种网络结构，具有很强的通用性。

LBT方式如图7(b)所示，其只适用于多通道语音信号，这是因为多通道语音信号中包含有不同说话人的空间位置信息。具体而言，LBT是将不同说话人的语音按说话人的位置(角度、距离)进行排序，并通过训练让网络学习这种排序方式。当说话人角度差异较大时，仅使用角度排序方式便可获得较好的分离效果，而当说话人角度差异较小时，则需要联合考虑角度与距离的影响，以达到较好的分离性能。

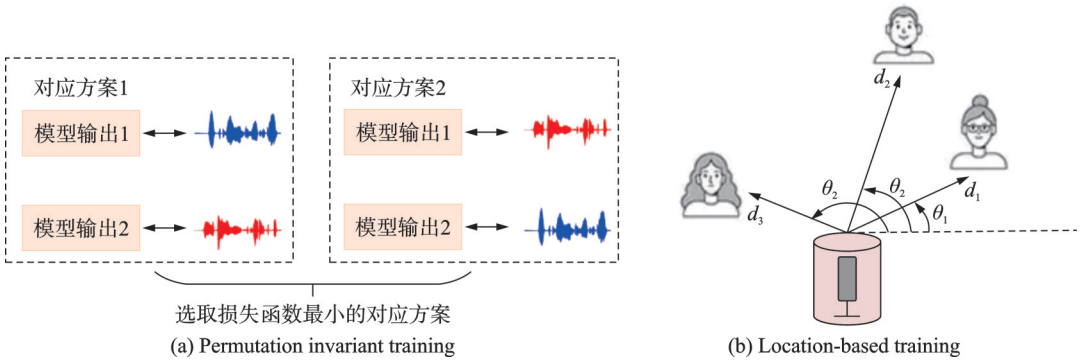

图7 基于PIT和LBT的标签排列问题解决方式示意图

Fig.7 Diagrams of solutions for the label permutation problem based on PIT and LBT

2.2.2 说话人数目已知情况下的多说话人分离方法

通常,说话人数目已知情况下的多说话人分离采用掩蔽估计或谱映射的方式得到多说话人的语音信号,下面以掩蔽估计为例进行说明,其基本框架如图8所示。

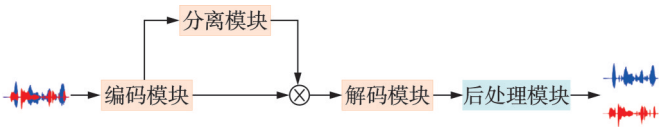

图8 说话人数目已知情况下的多说话人分离基本框图

Fig.8 Diagram of multi-speaker separation with a known number of speakers

通过改变编/解码模块、分离模块、后处理模块与网络输入,可得到不同的分离方法。针对编/解码模块,现有方法可以利用短时傅里叶变换/短时傅里叶逆变换直接得到具有结构性的时频谱[29],也可以利用单层或深层的卷积/转置卷积结构学习得到更适合于分离任务的变换域[30-32];此外,参数化滤波器组同样可以应用到编/解码模块以提高模型的可解释性[33-34]。对于分离模块,在时域方法中使用双路径策略可以有效地对特征块内与特征块间的关联性进行建模[35-36],而在时频域方法中,该策略可以有效地对帧间与频点间的关联性建模[37-39];此外,注意力机制[40-41]、多尺度特征[42-43]、超分辨率技术[44]及状态空间模型[45]等也可应用于分离模块以提升分离性能。针对后处理模块,在一些现有方法中使用了粗-细粒度两阶段结构[46-47]或多次迭代结构[48],使用这些结构可以进一步降低分离语音中的失真。对于网络输入,除单通道混合信号外,还可使用多通道混合信号作为模型输入,以利用多通道信号中包含的空间信息[49-50];当使用多通道混合信号作为输入时,现有方法通常将深度学习与波束形成技术结合起来,以减少由神经网络引起的非线性失真[51]。除此以外,现有方法还设计了多种损失函数,如对比损失函数[52]等。

2.2.3 说话人数目未知情况下的多说话人分离方法

考虑说话人数目未知情况下的多说话人分离时,需要同时解决说话人数目估计和多说话人分离这两个子任务。因此,相较于说话人数目已知情况,说话人数目未知情况下的多说话人分离显得更加复杂和困难。现有方法大多采用以下3种模式之一解决问题,即并行模式、串行模式以及说话人条件链模式。

采用并行模式的方法通常需要事先假定模型的最大输出通道数目$N_{\max}$。在这类方法中,分离模块同时分离出$N_{\max}$个语音信号,并通过额外的判别机制对说话人数目进行直接估计或依次确定每个输出通道的有效性,即该类方法中分离模块主要用于实现多说话人的分离,而判别机制则用于说话人数目的估计[53-54],如图9所示。

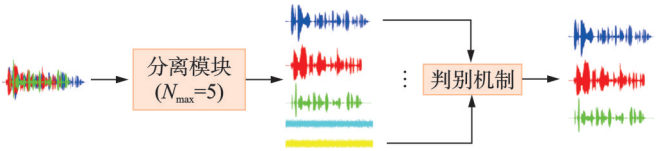

图9 说话人数目未知情况下采用并行模式的多说话人分离方法示意图

Fig.9 Schematic diagram of multi-speaker separation using parallel mode with an unknown number of speakers

采用串行模式的方法的核心思想是通过迭代的方式,依次输出不同说话人的语音信号[55-56],如图10所示。混合信号经过分离模块分别输出某一说话人的语音信号与残留信号,此时,需利用额外的判别机制对残留信号中是否存在语音进行判别,若存在语音,则进行下一次分离,若不存在语音,则停止迭代。

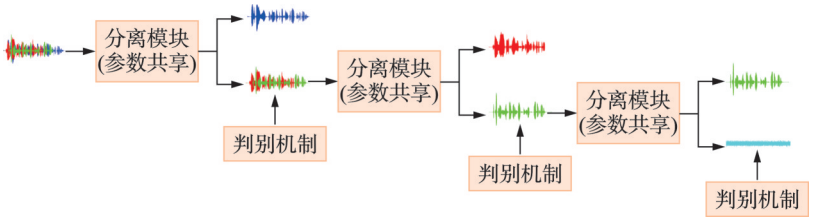

图10 说话人数目未知情况下采用串行模式的多说话人分离方法示意图

Fig.10 Schematic diagram of multi-speaker separation using serial mode with an unknown number of speakers

采用说话人条件链模式的方法通常包含一个说话人特征生成模块,如图11所示。该模块以混合信号作为输入,依次输出表征不同说话人的高维向量,输出的高维向量个数即为对说话人数目的估计,随后以这些多说话人的高维向量作为指导,将其与混合信号一起送入分离模块,从而分离出多说话人的语音信号[57-58]。

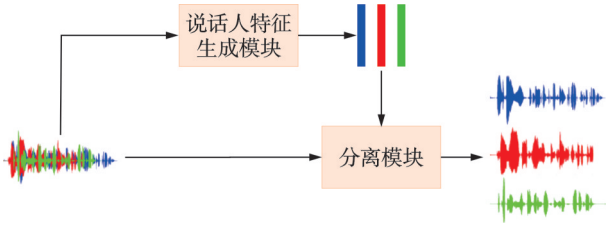

图11 说话人数目未知情况下采用说话人条件链模式的多说话人分离方法示意图

Fig.11 Schematic diagram of multi-speaker separation using speaker-conditional chain mode with an unknown number of speakers

在说话人数目未知情况下,上述3种模式均能有效地估计说话人数目并分离出不同说话人的语音。采用并行模式的方法能够同时输出多个说话人的语音信号,但当实际说话人数目超过预先设定的最大输出通道数目时,其分离效果将有所降低。采用串行模型的方法通过迭代的方式分离出不同说话人的语音,但其推理时间较长,且多次迭代还将导致误差累积。采用说话人条件链模式的方法依赖于对不同说话人高维向量的估计,且需要实现这些高维向量与混合信号的有效融合。

3 目标说话人提取

与多说话人分离不同,目标说话人提取利用目标说话人的辅助信息作为指导,仅仅从混合信号中分离出目标说话人的语音信号。根据目标说话人辅助信息类型的不同,现有的目标说话人提取方法大致可分为3类。第1类是基于空间信息的目标说话人提取,这类方法利用目标说话人的空间方位信息作为指导,常见的空间信息包括声源到达方向、目标说话人与麦克风间的距离以及目标说话人所在区域等。第2类是基于视觉信息的目标说话人提取,这类方法通过融合视觉信息来辅助目标说话人语音的提取,常用的视觉信息包括视频流、面部图像和姿态信息等。第3类是基于音频信息的目标说话人提

取,这类方法往往以目标说话人的音频特征作为指导,常用的音频信息包括注册语音、目标说话人话音活动性等。此外,与多说话人分离类似,根据是否使用深度神经网络,目标说话人提取方法可以分为传统目标说话人提取方法与基于深度学习的目标说话人提取方法两类。

3.1 传统目标说话人提取方法

常见的传统目标说话人提取方法包括波束形成方法与盲信号提取方法,这些方法通常需要使用多个麦克风以拾取多通道混合信号,并基于混合信号的统计特性提取目标说话人语音信号。与传统多说话人分离方法类似,这些传统目标说话人提取方法往往难以利用信号中高层次的语义信息,故而在复杂声学场景中可能表现不足。

3.1.1 波束形成方法

波束形成方法又可称为空间滤波方法,是一种基于空间信息的目标说话人提取方法。该方法的基本思想是利用多通道混合信号中包含的空间信息构造空间滤波器,使其在目标方向上形成主瓣,从而提取目标方向的语音信号,并抑制来自非目标方向的信号[59-60]。此外,常用的波束形成方法有固定波束形成方法和自适应波束形成方法,其中,固定波束形器的系数是固定的,而自适应波束形成器的系数随着信号统计特性的变化自适应地改变。用于评测波束形成器性能的常用指标包括白噪声增益、指向性因子与波束图等。白噪声增益反映了波束形成器在白噪声环境下的信噪比增益。在麦克风阵列中,各麦克风带有的自噪声以及麦克风间的不匹配性等因素都可看作是白噪声,因此,白噪声增益可以用来衡量波束形成器的鲁棒性。指向性因子则反映了波束形成器在各向同性噪声环境下的信噪比增益,其中,各向同性噪声往往是由多个均匀分布在麦克风阵列周围的噪声源产生的,也可能是强混环境中由各个方向反射信号构成的,因此,指向性因子可以用来衡量波束形成器对来自非目标方向信号的抑制能力。波束图反映了波束形成器对来自不同方向信号的增益程度,如图12所示。在图12中,目标信号方向为0°与180°,因此,该方向上的信号经过波束形成器滤波后可以无失真地保留下来,而来自非目标方向的信号将有不同程度的衰减。

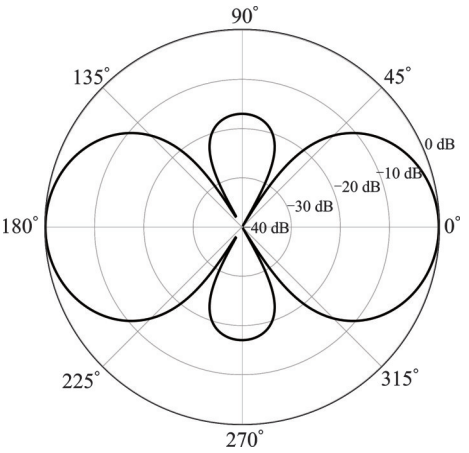

图12 波束图示例

Fig.12 An example of beam pattern

在固定波束形成器中,最为直接的波束形成器是延迟求和波束形成器。它是将各麦克风拾取的信号进行适当的延迟并叠加,其中延迟取决于目标声源与麦克风阵列的位置关系,其目的是使得来自目标方向的语音信号在时间上同步,可以使得白噪声增益最大化。超指向性波束形成器对应的优化问题是在满足目标方向无失真约束条件下最大化指向性因子,因此,该波束形成器对非目标方向的信号有较强的抑制能力,但其缺点在于白噪声增益低。差分波束形成是一类特殊的固定波束形成器,其同样可以用线性麦克风阵列实现,但需要麦克风间距较小,利用通过零点约束以及近似条件得到具有频率一致性的波束形成器。

在自适应波束形成器中,应用最为广泛的是最小方差无失真波束形成器,其对应的优化问题为在满足目标方向无失真约束条件下最小化残留噪声的方差,理论上该波束形成器可根据噪声的统计特性自适应地调制系数,从而具有较为理想的降噪能力,但实际应用中由于参数估计不准确等问题,其降噪能力受到限制;线性约束无失真波束形成器则考虑多个多说话人的情况,添加了额外的约束条件;此外,常用的自适应波束形成器还包括最大增益波束形成器和折中波束形成器等。

在处理混合信号时,波束形成方法可以提取目标方向的说话人语音并抑制非目标方向的说话人语音,该方法无需依赖大量数据且通常具有较低的计算复杂度,其提取性能与抑制效果可通过波束图进行直接的观察与分析。然而,波束形成方法需要使用麦克风阵列,这意味着其性能很大程度上取决于麦克风阵列的设计与布局。此外,在强混响的复杂声学场景中,波束形成方法的性能会显著下降。

3.1.2 盲信号提取方法

盲信号提取方法(Blind signal extraction, BSE)是一种特殊形式的盲源分离方法,但与盲源分离方法不同,盲信号提取方法仅提取目标声源信号,而无需完全分离出所有声源信号。盲信号提取方法的核心思想是在仅有少量先验信息的情况下,利用信号的统计特性,从混合信号中提取出目标声源信号。通常,盲信号提取方法假设目标声源信号在某些统计特性上(如稀疏性、非高斯性、独立性等)与非目标声源信号不同,从而能够利用这些差异来实现信号的提取。常用的盲信号提取方法有独立成分提取方法[61]和独立矢量提取方法[62]等。

在处理混合信号时,盲信号提取方法能够在复杂环境中有效地提取出目标声源信号。由于只需提取目标声源信号,而无需分离出所有声源信号,因此盲信号提取的计算复杂度更低。然而,其提取性能依赖于目标声源信号与非目标声源信号的统计特性差异,故当二者特性相似时,提取效果可能并不理想。此外,盲信号提取方法要求对算法进行适当的初始化和参数调整,否则容易得到局部最优解而非全局最优解。与盲源分离方法相似,盲信号提取方法主要依赖于信号的统计特性,难以对信号中的高层次语义信息进行建模。

3.2 基于深度学习的目标说话人提取方法

与基于深度学习的多说话人分离方法类似,基于深度学习的目标说话人提取方法可以既对低层次的特征进行建模,又可以对信号中包含的高层次语义信息进行建模。基于深度学习的目标说话人提取方法具有以下两个优势:一是该方法无需事先知道混合信号中的说话人数目,即无需对说话人数目已知情况与说话人数目未知情况进行分别讨论;二是该方法仅需从混合信号中提取目标说话人的语音信号,故模型的输出通道数只有1个,从而避免了标签排列问题。由于目标说话人辅助信息的类型多种多样,为简单起见,本文将着重探讨利用目标说话人注册语音作为辅助信息的提取方法。

3.2.1 基于说话人嵌入的目标说话人提取方法

基于说话人嵌入的目标说话人提取方法(Embedding-based TSE)通常以表征目标说话人特征的说话人嵌入(Speaker embedding)[63-65]作为指导,以提取目标说话人的语音信号。通常,说话人嵌入是从注册语音中提取到的高维向量,如图13所示。目标说话人的注册语音作为多层网络的输入进行帧级别的特征学习,所得帧级别的特征进行聚合后再经过映射得到段级别的特征,从而得到说话人嵌入。

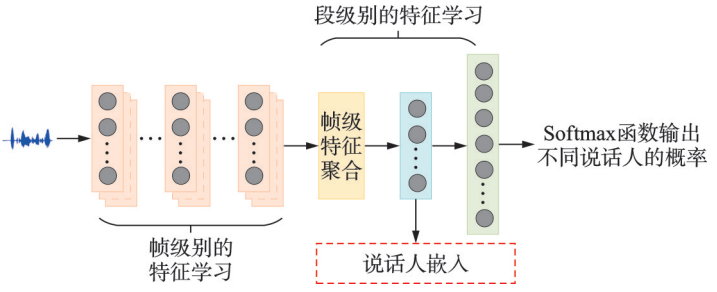

图13 提取说话人嵌入的基本框图

Fig.13 Diagram of the extraction of speaker embedding

基于说话人嵌入的目标说话人提取的基本框图如图14所示,现有的方法根据编/解码模块、提取模块和说话人嵌入提取模块的不同,可进行不同的分类,本文将现有方法分为时频域方法和时域方法。时频域方法通过短时傅里叶变换将麦克风拾取的混合信号从时域变换到时频域。在早期的目标说话人提取方法[66-67]中,首先对混合信号的幅度谱进行处理,通过多层网络提取混合信号的高维特征。与此同时,目标说话人的注册语音经过预

训练的说话人嵌入提取模块,得到相应的说话人嵌入。随后,将该说话人嵌入作为指导,与混合信号的高维特征拼接在一起送入到提取模块中,以估计得到目标说话人的掩蔽。最后,该掩蔽与混合信号的幅度谱进行元素乘运算并结合混合信号的相位谱,估计得到目标说话人的语音信号。然而,使用预训练的说话人嵌入提取模型得到的说话人嵌入对于目标说话人提取任务而言可能不是最优的,为缓解这一问题,可以将说话人嵌入提取模块与网络的其他部分进行联合训练[68-69]。此外,现有方法还引入了多种技术来提升提取性能,例如使用自适应特征融合[70]、注意力机制[71]、基于滤波器的掩蔽估计[72]与说话人表示损失[73]等。然而,时频域方法通常使用混合信号的相位谱来恢复目标说话人的语音信号,这可能导致提取结果不够理想。相比之下,时域方法往往直接处理波形,从而避免相位估计问题,因此逐渐成为主流方法。一些现有时域方法设计了多尺度模块以捕获不同时间分辨率下的结构[74]。此外,通过使用权重共享技术[75]、不同的融合技术[76-77]、多级结构[78]、说话人嵌入的迭代更新[79]、起始/终止信息[80]和合适的网络结构[81]等,可以进一步提升目标说话人提取的性能。

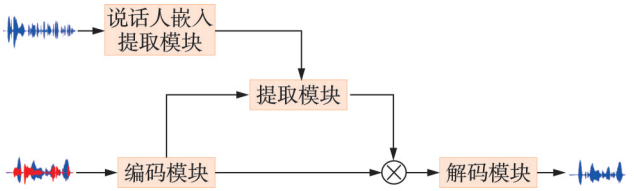

图14 基于说话人嵌入的目标说话人提取基本框图

Fig.14 Diagram of the embedding-based TSE method

3.2.2 基于上下文信息的目标说话人提取方法

现有的基于说话人嵌入的目标说话人提取方法已经实现了很高的提取性能。然而,说话人嵌入仅仅是一个高维向量,尽管它能够有效地对说话人进行表征,但它往往丢掉了注册语音中的内容信息。为了解决这一问题,研究人员提出了基于上下文信息的目标说话人提取方法(Contextual information-based TSE),这些方法不仅可以学习目标说话人的特征,还可以对注册语音中的内容信息进行建模。为有效利用注册语音包含的上下文信息,一些现有方法使用参数共享的网络对混合信号与注册语音进行处理,并利用注意力机制实现两组特征序列间的融合[82-83]。此外,现有方法还可以使用双向长短时记忆网络的状态变量来对注册语音包含的上下文信息进行建模[84]。然而,使用处理后的特征序列或状态变量可能会导致目标说话人信息的部分丢失。为解决这一问题,现有方法提出了直接在时频域中让注册语音与混合信号的实部与虚部分别进行交互,经过交互得到的一致时频表示能够更好地指导目标说话人的提取过程[85],图15给出了文献[85]的语谱图对比。

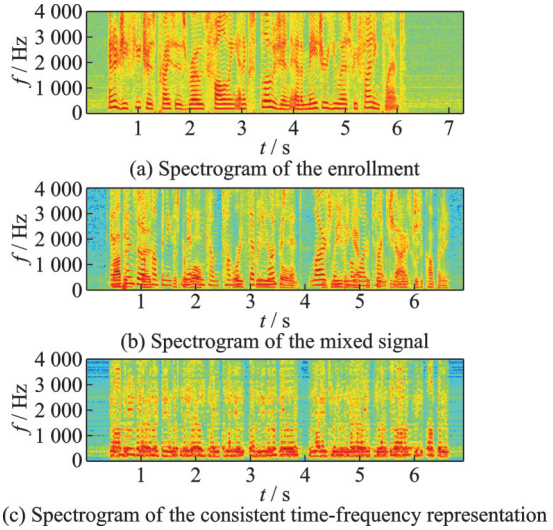

图15 语谱图对比

Fig.15 Comparison of the spectra

与图15(a)中的注册语音语谱相比,图15(c)中的一致时频表示语谱在语音信号的起始、终止、话音活动性与帧数方面存在显著差异。然而,目标说话人的一些特征,如谐波结构,仍在一定程度上得以保留。进一步比较图15(b)和图15(c)可以发现,它们展现出一致的谱模式,即具有相似的起始、终止、语音活动性和相同的帧数。因此,该方法中设计的直接交互机制可以有效地突出那些与混合信号帧高度相似的注册语音帧,从而有效利用这些帧中包含的说话人特征和内容信息。将一致时频表示用作后续提取过程的指导,有助于提升目标说话人提取的性能。

4 现存问题与研究展望

相比于传统的多说话人分离方法和目标说话人提取方法,基于深度学习的多说话人分离和目标说话人提取方法能够充分利用深度神经网络的强大建模能力,在多个方面取得了显著的性能提升。这种提升不仅体现在分离和提取的准确性上,也体现在模型对复杂声学环境的适应能力上。然而,尽管取得了这些进展,现有方法依然面临一些亟待解决的问题。下面将详细探讨基于深度学习的多说话人分离和目标说话人提取方法的共性问题和个性化问题。

基于深度学习的多说话人分离方法与基于深度学习的目标说话人提取方法的共性问题。(1)鲁棒性与泛化性问题。目前,许多基于深度学习的分离和提取模型在特定的训练集上表现出色,但在与训练集不匹配的测试集上,性能往往会显著下降。这种性能下降主要是由于模型在处理新环境、新噪声类型或未见过的说话人时,无法很好地泛化。在实际应用场景中,噪声类型与混响环境往往是多种多样的,而且模型可能会遇到在训练阶段从未见过的说话人。因此,模型必须具有较强的泛化能力才能适应这些变化,保证在不同条件下仍然能提供稳定和高效的性能。(2)实时性问题。为了实现高性能的分离和提取,模型通常需要对语音的特征序列进行复杂的建模,这往往依赖于未来帧的信息。然而,在实际应用中,实时系统对处理的时延通常有严格的要求。例如,在实时通信或实时语音交互应用中,延迟过大会影响用户体验。因此,模型设计中需要在性能和实时性之间进行折中,即在不显著降低分离和提取性能的情况下,尽量减少处理延迟。(3)计算复杂度问题。高性能的分离和提取模型通常具有较大的参数量和较高的计算复杂度,这使得它们难以在计算资源有限的终端设备上部署,如移动设备与嵌入式系统。因此,需要在模型设计中考虑到计算资源的限制,通过优化模型结构或引入轻量级的算法,在性能与计算复杂度之间进行折中。

尽管基于深度学习的多说话人分离方法取得了显著的进展,但其仍面临一些个性化挑战。(1)多说话人场景的扩展性问题。当前,许多基于深度学习的分离方法在处理两个说话人同时讲话的场景时,能够展现出卓越的性能。然而,当面对3个或更多说话人同时讲话时,模型的分离性能往往显著下降。这是因为随着说话人数目的增加,语音信号的复杂性急剧上升。多个说话人的语音信号会相互重叠,使得分离任务变得更加困难。(2)模型的可解释性问题。多说话人分离模型通常依赖于深度神经网络,利用多层网络结构提取深层次的特征从而实现分离。这种复杂的网络结构和深层次的特征表示导致了模型的“黑箱”性质,使得研究人员难以理解模型是如何在层层网络中区分并分离不同说话人的。在模型出现性能下降时,难以快速定位问题的根源并进行针对性的改进。(3)说话人数目未知问题。在实际场景中,说话人数目往往是不确定的,且可能是动态变化的。然而,现有的多说话人分离模型通常假定说话人数目是已知的。由于无法预先确定说话人数目,现有模型的处理能力和灵活性受到显著限制。综合来看,尽管基于深度学习的多说话人分离方法已取得出色的性能,但在处理多说话人场景、提升模型的可解释性以及应对动态变化的说话人数时,仍面临着挑战。未来的研究需要进一步探索和解决这些多说话人分离的个性化问题,以提升分离模型在实际应用中的鲁棒性和灵活性。

类似地,基于深度学习的目标说话人提取也存在一些个性化问题。(1)目标说话人不存在问题。在实际应用中,可能会遇到目标说话人并未发声的情况,即目标说话人的语音信号并不存在于混合信号中。现有的提取模型在这种情况下往往无法准确判断目标说话人是否存在,可能会误将噪声或非目标说话人的语音认为是目标说话人的语音。这种误判会导致模型错误地输出非静音信号,进而影响提取的准确性和实际应用效果。特别是在嘈杂或多说话人环境中,提取模型需要具备更强的判断能力,以避免将非目标说话人语音提取为目标说话人语音。这要求模型在训练过程中能够学习并适应多样化的环境和场景,从而提高其在目标说话人不存在时的处理能力和鲁棒性。(2)多模态信息利用问题。在

目标说话人提取任务中,除了使用注册语音外,结合目标说话人的其他辅助信息(如视频、文本等)可以显著提升提取性能。然而,不同模态的数据各自具有独特的特点,如何在一个统一的框架中高效地融合和处理这些多模态信息,是一个极具挑战性的问题。在实际应用中,获取和处理多模态信息可能带来额外的复杂性和成本。此外,并不是所有的应用场景都能够提供完整的多模态数据,如视频信息可能由于摄像头视角、遮挡或环境限制而不可用。因此,如何在多模态信息缺少的情况下,继续保持提取模型的高效性和准确性,是一个亟待解决的问题。综上所述,尽管基于深度学习的目标说话人提取方法已展现了出色的性能,但在处理目标说话人不存在的情况和有效利用多模态信息时,仍然面临着挑战,解决这些个性化问题对于提升提取模型在实际应用中的可靠性与稳定性至关重要。

5 结束语

语音分离作为语音信号处理领域的一个研究热点,近年来受到了广泛关注。研究人员提出了多种不同类型的语音分离方法。本文根据目标说话人的辅助信息利用与否,将语音分离方法分为两大类,即多说话人分离和目标说话人提取。进一步,根据是否采用深度学习技术,本文将多说话人分离细分为传统多说话人分离方法和基于深度学习的多说话人分离方法;类似地,将目标说话人提取分为传统目标说话人提取方法和基于深度学习的目标说话人提取方法。目标说话人提取分为传统目标说话人提取方法和基于深度学习的目标说话人提取方法。本文对这些方法进行了详细的介绍,并分析了各类方法的优缺点。其中,传统多说话人分离方法与传统目标说话人提取方法通常依赖于信号的统计特性进行建模,这些方法具有较好的可解释性,但难以利用高层次的语义信息。相反,基于深度学习的多说话人分离与目标说话人提取方法依赖于深度神经网络强大的建模能力,能够利用信号中高层次的语义信息,因此性能通常优于传统方法,但需要大量的数据进行模型训练,并且计算复杂度更高。最后,本文总结了基于深度学习的多说话人分离和目标说话人提取方法中存在的一些共性问题和个性化问题,并对未来的研究方向进行展望。

参考文献:

[1] WANG D, CHEN J. Supervised speech separation based on deep learning: An overview[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 26(10): 1702-1726.

[2] QIAN Y, WENG C, CHANG X, et al. Past review, current progress, and challenges ahead on the cocktail party problem[J]. Frontiers of Information Technology & Electronic Engineering, 2018, 19: 40-63.

[3] 刘文举, 聂帅, 梁山, 等. 基于深度学习语音分离技术的研究现状与进展[J]. 自动化学报, 2016, 42(6): 819-833.
LIU Wenju, NIE Shuai, LIANG Shan, et al. Deep learning based speech separation technology and its developments[J]. Acta Automatica Sinica, 2016, 42(6): 819-833.

[4] 黄雅婷, 石晶, 许家铭, 等. 鸡尾酒会问题与相关听觉模型的研究现状与展望[J]. 自动化学报, 2019, 45(2): 234-251.
HUANG Yating, SHI Jing, XU Jiaming, et al. Research advances and perspectives on the cocktail party problem and related auditory models[J]. Acta Automatica Sinica, 2019, 45(2): 234-251.

[5] ZMOLIKOVA K, DELCROIX M, OCHIAI T, et al. Neural target speech extraction: An overview[J]. IEEE Signal Processing Magazine, 2023, 40(3): 8-29.

[6] CHERRY C. Some experiments on the recognition of speech, with one and with two ears[J]. The Journal of the Acoustical Society of America, 1953, 25(5): 975-979.

[7] WANG D, BROWN G J. Computational auditory scene analysis: Principles, algorithms, and applications[M]. USA: Wiley-IEEE Press, 2006.

[8] HU G, WANG D. A tandem algorithm for pitch estimation and voiced speech segregation[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2010, 18(8): 2067-2079.

[9] HU G. Monaural speech organization and segregation[D]. Columbus: The Ohio State University, 2006.

[10] GHAHRAMANI Z, JORDAN M I. Factorial hidden Markov models[J]. Machine Learning, 1997, 29: 245-273.

[11] COOKE M, HERSHEY J R, RENNIE S J. Monaural speech separation and recognition challenge[J]. Computer Speech and Language, 2010, 24(1): 1-15.

[12] LEE D D, SEUNG H S. Learning the parts of objects with nonnegative matrix factorization[J]. Nature, 1999, 401: 788-791.

[13] FEVOTTE C, IDIER J. Algorithms for nonnegative matrix factorization with the β-divergence[J]. Neural Computation, 2011, 23(9): 2421-2456.

[14] SMARAGDIS P, FEVOTTE C, MYSORE G J, et al. Static and dynamic source separation using nonnegative factorizations: A unified view[J]. IEEE Signal Processing Magazine, 2014, 31(3): 66-75.

[15] COMON P. Independent component analysis, A new concept?[J]. Signal Processing, 1994, 36(3): 287-314.

[16] HYVARINEN A, KARHUNEN J, OJA E. Independent component analysis[M]. USA: John Wiley & Sons, 2001.

[17] HYVARINEN A, OJA E. A fast fixed-point algorithm for independent component analysis[J]. Neural Computation, 1997, 9(7): 1483-1492.

[18] HYVARINEN A. Fast and robust fixed-point algorithms for independent component analysis[J]. IEEE Transactions on Neural Networks, 1999, 10(3): 626-634.

[19] PHAM D T, GARAT P. Blind separation of mixture of independent sources through a quasi-maximum likelihood approach[J]. IEEE Transactions on Signal Processing, 1997, 45(7): 1712-1725.

[20] YANG H H, AMARI S. Adaptive online learning algorithms for blind separation: Maximum entropy and minimum mutual information[J]. Neural Computation, 1997, 9(7): 1457-1482.

[21] SMARAGDIS P. Blind separation of convolved mixtures in the frequency domain[J]. Neurocomputing, 1998, 22(1/2/3): 21-34.

[22] KIM T, ATTIAS H T, LEE S Y, et al. Blind source separation exploiting higher-order frequency dependencies[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2007, 15(1): 70-79.

[23] HERSHEY J R, CHEN Z, LE ROUX J, et al. Deep clustering: Discriminative embeddings for segmentation and separation [C]//Proceedings of 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Shanghai, China: IEEE, 2016: 31-35.

[24] ISIK Y, LE ROUX J, CHEN Z, et al. Single-channel multi-speaker separation using deep clustering[C]//Proceedings of Interspeech 2016. San Francisco, USE: ISCA, 2016: 545-549.

[25] YU D, KOL M, TAN Z H, et al. Permutation invariant training of deep models for speaker-independent multi-talker speech separation[C]//Proceedings of 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New Orleans, LA, USA: IEEE, 2017: 241-245.

[26] KOL M, YU D, TAN Z H, et al. Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2017, 25(10): 1901-1913.

[27] TAHERIAN H, TAN K, WANG D. Location-based training for multi-channel talker-independent speaker separation[C]// Proceedings of 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore: IEEE, 2022: 696-700.

[28] TAHERIAN H, TAN K, WANG D. Multi-channel talker-independent speaker separation through location-based training[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022, 30: 2791-2800.

[29] YANG L, LIU W, WANG W. TFPSNet: Time-frequency domain path scanning network for speech separation[C]// Proceedings of 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore: IEEE, 2022: 6842-6846.

[30] LUO Y, MESGARANI N. TaSNet: Time-domain audio separation network for real-time, single-channel speech separation[C]// Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, AB, Canada: IEEE, 2018: 696-700.

[31] LUO Y, MESGARANI N. Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 27(8): 1256-1266.

[32] KADIOGLU B, HORGAN M, LIU X, et al. An empirical study of Conv-TasNet[C]//Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain: IEEE, 2020: 7264-7268.

[33] DITTER D, GERKMANN T. A multi-phase Gammatone filterbank for speech separation via TasNet[C]//Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain: IEEE, 2020: 36-40.

[34] PARIENTE M, CORNELL S, DELEFORGE A, et al. Filterbank design for end-to-end speech separation[C]//Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain: IEEE, 2020: 6364-6368.

[35] LUO Y, CHEN Z, YOSHIOKA T. Dual-path RNN: Efficient long sequence modeling for time-domain single-channel speech separation[C]//Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain: IEEE, 2020: 46-50.

[36] CHEN J, MAO Q, LIU D. Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation[C]//Proceedings of Interspeech 2020. Shanghai, China: ISCA, 2020: 2642-2646.

[37] WANG Z Q, CORNELL S, CHOI S, et al. TF-GRIDNET: Making time-frequency domain models great again for monaural speaker separation[C]//Proceedings of 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Rhodes Island, Greece: IEEE, 2023: 1-5.

[38] WANG Z Q, CORNELL S, CHOI S, et al. TF-GridNet: Integrating full- and sub-band modeling for speech separation[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023, 31: 3221-3236.

[39] YANG X, BAO C, ZHANG X, et al. Monaural speech separation method based on recurrent attention with parallel branches[C]//Proceedings of Interspeech 2023. Dublin, Ireland: ISCA, 2023: 3794-3798.

[40] SUBAKAN C, RAVANELLI M, CORNELL S, et al. Attention is all you need in speech separation[C]//Proceedings of 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto, ON, Canada: IEEE, 2021: 21-25.

[41] LAM M W Y, WANG J, SU D, et al. Sandglasset: A light multi-granularity self-attentive network for time-domain speech separation[C]//Proceedings of 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto, ON, Canada: IEEE, 2021: 5759-5763.

[42] TZINIS E, WANG Z, SMARAGDIS P. Sudo RM-RF: Efficient networks for universal audio source separation[C]//Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain: IEEE, 2020: 1-5.

[43] HU X, LI K, ZHANG W, et al. Speech separation using an asynchronous fully recurrent convolutional neural network[C]//Proceedings of the 35th Annual Conference on Neural Information Processing Systems. Canada: iNSPIRE, 2021: 22509-22522.

[44] RIXEN J, RENZ M. SFSRNet: Super-resolution for single-channel audio source separation[C]//Proceedings of the 36th AAAI Conference on Artificial Intelligence. Vancouver, BC, Canada: AAAI, 2022: 11220-11228.

[45] CHEN C, YANG C H H, LI K, et al. A neural state-space modeling approach to efficient speech separation[C]//Proceedings of Interspeech 2023. Dublin, Ireland: ISCA, 2023: 3784-3788.

[46] YAO Z, PEI W, CHEN F, et al. Stepwise-refining speech separation network via fine-grained encoding in high-order latent domain[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022, 30: 378-393.

[47] YANG X, BAO C, CHEN X. Coarse-to-fine speech separation in the time-frequency domain[J]. Speech Communication, 2023, 155: 103003.

[48] LUTATI S, NACHMANI E, WOLF L. SepIt: Approaching a single channel speech separation bound[C]//Proceedings of Interspeech 2022. Incheon, Korea: ISCA, 2022: 5323-5327.

[49] WANG Z Q, LE ROUX J, HERSHEY J R. Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation[C]//Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, AB, Canada: IEEE, 2018: 1-5.

[50] WANG Z Q, WANG D. Combining spectral and spatial features for deep learning based blind speaker separation[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 27(2): 457-468.

[51] OCHIAI T, DELCROIX M, IKESHITA R, et al. Beam-TasNet: Time-domain audio separation network meets frequency-

domain beamformer[C]//Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain: IEEE, 2020: 6384-6388.

[52] ZHANG Z, CHEN C, CHEN H H, et al. Noise-aware speech separation with contrastive learning[C]//Proceedings of 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Seoul, Korea: IEEE, 2024: 1381-1385.

[53] LUO Y, MESGARANI N. Separating varying numbers of sources with auxiliary autoencoding loss[C]//Proceedings of Interspeech 2020. Shanghai, China: ISCA, 2020: 2622-2626.

[54] CHAZAN S E, WOLF L, NACHMANI E, et al. Single channel voice separation for unknown number of speakers under reverberant and noisy settings[C]//Proceedings of 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto, ON, Canada: IEEE, 2021: 3730-3734.

[55] KINOSHITA K, DRUDE L, DELCROIX M, et al. Listening to each speaker one by one with recurrent selective hearing networks[C]//Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, AB, Canada: IEEE, 2018: 5064-5068.

[56] TAKAHASHI N, PARTHASAARATHY S, GOSWAMI N, et al. Recursive speech separation for unknown number of speakers[C]// Proceedings of Interspeech 2019. Graz, Austria: ISCA, 2019: 1348-1352.

[57] SHI J, XU J, FUJITA Y, et al. Speaker-conditional chain model for speech separation and extraction[C]//Proceedings of Interspeech 2020. Shanghai, China: ISCA, 2020: 2707-2711.

[58] CHETUPALLI S R, HABETS E A P. Speaker counting and separation from single-channel noisy mixtures[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023, 31: 1681-1692.

[59] BENESTY J, CHEN J, HUANG Y. Microphone array signal processing[M]. Berlin, Germany: Springer-Verlag, 2008.

[60] BENESTY J, COHEN I, CHEN J. Fundamentals of signal enhancement and array signal processing[M]. USA: John Wiley & Sons, 2017.

[61] KOLDOVSKY Z, TICHAVSKY P. Gradient algorithms for complex non-Gaussian independent component/vector extraction, question of convergence[J]. IEEE Transactions on Signal Processing, 2019, 67(4): 1050-1064.

[62] AMOR N, CMEJLA J, KAUTSKY V, et al. Blind extraction of moving sources via independent component and vector analysis: Examples[C]//Proceedings of 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto, ON, Canada: IEEE, 2021: 3725-3729.

[63] SNYDER D, GARCIA-ROMERO D, SELL G, et al. X-Vectors: Robust DNN embeddings for speaker recognition[C]// Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, AB, Canada: IEEE, 2018: 5329-5333.

[64] WAN L, WANG Q, PAPIR A, et al. Generalized end-to-end loss for speaker verification[C]//Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, AB, Canada: IEEE, 2018: 4879-4883.

[65] DESPLANQUES B, THIENPONDT J, DEMUYNCK K. ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification[C]//Proceedings of Interspeech 2020. Shanghai, China: ISCA, 2020: 3830-3834.

[66] WANG Q, MUCKENHIRN H, WILSON K, et al. VoiceFilter: Targeted voice separation by speaker-conditioned spectrogram masking[C]//Proceedings of Interspeech 2019. Graz, Austria: ISCA, 2019: 2728-2732.

[67] LI W, ZHANG P, YAN Y. Target speaker recovery and recognition network with average x-vector and global training[C]// Proceedings of Interspeech 2019. Graz, Austria: ISCA, 2019: 3233-3237.

[68] DELCROIX M, ZMOLIKOVA K, KINOSHITA K, et al. Single channel target speaker extraction and recognition with speaker beam[C]//Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, AB, Canada: IEEE, 2018: 5554-5558.

[69] XU C, RAO W, CHNG E S, et al. Optimization of speaker extraction neural network with magnitude and temporal spectrum approximation loss[C]//Proceedings of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, UK: IEEE, 2019: 6990-6994.

[70] ZMOLIKOVA K, DELCROIX M, KINOSHITA K, et al. SpeakerBeam: Speaker aware neural network for target speaker extraction in speech mixtures[J]. IEEE Journal of Selected Topics in Signal Processing, 2019, 13(4): 800-814.

[71] LI T, LIN Q, BAO Y, et al. Atss-Net: Target speaker separation via attention-based neural network[C]//Proceedings of Interspeech 2020. Shanghai, China: ISCA, 2020: 1411-1415.

[72] HE S, LI H, ZHANG X. Speakerfilter: Deep learning-based target speaker extraction using anchor speech[C]//Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain: IEEE, 2020: 376-380.

[73] MUN S, CHOE S, HUH J, et al. The sound of my voice: Speaker representation loss for target voice separation[C]//Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain: IEEE, 2020: 7289-7293.

[74] XU C, RAO W, CHNG E S, et al. SpEx: Multi-scale time domain speaker extraction network[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 28: 1370-1384.

[75] GE M, XU C, WANG L, et al. SpEx+: A complete time domain speaker extraction network[C]//Proceedings of Interspeech 2020. Shanghai, China: ISCA, 2020: 1406-1410.

[76] WANG W, XU C, GE M, et al. Neural speaker extraction with speaker-speech cross-attention network[C]//Proceedings of Interspeech 2021. Brno, Czechia: ISCA, 2021: 3535-3539.

[77] HAN J, RAO W, LONG Y, et al. Attention-based scaling adaptation for target speech extraction[C]//Proceedings of 2021 IEEE Workshop Automatic Speech Recognition and Understanding Workshop. Taipei, China: IEEE, 2021: 658-662.

[78] GE M, XU C, WANG L, et al. Multi-stage speaker extraction with utterance and frame-level reference signals[C]//Proceedings of 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto, ON, Canada: IEEE, 2021: 6109-6113.

[79] DENG C, MA S, SHA Y, et al. Robust speaker extraction network based on iterative refined adaptation[C]//Proceedings of Interspeech 2021. Brno, Czechia: ISCA, 2021: 3530-3534.

[80] HAO Y, XU J, ZHANG P, et al. Wase: Learning when to attend for speaker extraction in cocktail party environments[C]//Proceedings of 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto, ON, Canada: IEEE, 2021: 6104-6108.

[81] LIU K, DU Z, WAN X, et al. X-SEPFORMER: End-to-end speaker extraction network with explicit optimization on speaker confusion[C]//Proceedings of 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Rhodes Island, Greece: IEEE, 2023: 1-5.

[82] XIAO X, CHEN Z, YOSHIOKA T, et al. Single-channel speech extraction using speaker inventory and attention network [C]//Proceedings of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, UK: IEEE, 2019: 86-90.

[83] ZENG B, HONGBIN S, WAN Y, et al. SEF-Net: Speaker embedding free target speaker extraction network[C]//Proceedings of Interspeech 2023. Dublin, Ireland: ISCA, 2023: 3452-3456.

[84] YANG L, LIU W, TAN L, et al. Target speaker extraction with ultra-short reference speech by VE-VE framework[C]//Proceedings of 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Rhodes Island, Greece: IEEE, 2023: 1-5.

[85] YANG X, BAO C, ZHOU J, et al. Target speaker extraction by directly exploiting contextual information in the time-frequency domain[C]//Proceedings of 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Seoul, Korea: IEEE, 2024: 10476-10480.

作者简介:

鲍长春(1965-),**通信作者**,男,教授,研究方向:语音与音频信号处理,E-mail: baochch@bjut.edu.cn。

杨雪(1991-),女,博士研究生,研究方向:语音分离,E-mail:yangx11@emails.bjut.edu.cn。

(编辑:张黄群)