ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 41,No. 2,Mar. 2026,pp. 397－415
DOI:10. 16337/j. 1004-9037. 2026. 02. 007

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

# 基于预训练模型的目标音频处理研究进展

刘　琚[1]，马　豪[1]，李晓航[1]，李玉楷[1]，司　媛[1]，邢志坤[1]，王芷涵[1]，邵明杰[1,2]

(1. 山东大学信息科学与工程学院，青岛 266237；2. 中国科学院数学与系统科学研究院，北京 100190)

**摘　要：**目标音频处理旨在根据用户提供的线索从混合信号中恢复或识别特定目标声源，是人机交互、智慧办公及多媒体取证等领域的关键技术。本文对近年来作者团队基于预训练模型的目标音频处理研究进展进行了概述。首先，回顾了目标说话人语音识别、语音提取、目标音频提取及音源分离等方向的研究现状，介绍了Whisper、对比学习语言音频预训练(Contrastive language-audio pretraining，CLAP)等预训练模型及参数高效微调技术。针对目标音频提取和目标说话人识别任务综述了作者团队研究的基于对比学习的多模态查询目标音频提取方法、无需配对数据的语言查询目标音频提取方法、基于多任务学习的目标说话人语音提取方法，以及基于提示微调的目标说话人语音识别方法等。这些方法分别在多模态泛化、标注数据依赖、语义保持与参数效率等方面取得了显著进展。最后，对推理效率提升、多模态深度融合、开放域泛化及通用目标音频处理大模型的构建等未来研究方向进行了展望。

**关键词：**目标音频处理；预训练模型；参数高效微调；目标音频提取；目标说话人语音识别；对比学习

**中图分类号：**TP183　　**文献标志码：**A

**引用格式：**刘琚，马豪，李晓航，等. 基于预训练模型的目标音频处理研究进展[J]. 数据采集与处理，2026，41(2)：397-415. LIU Ju，MA Hao，LI Xiaohang，et al. Research progress in target audio processing methods based on pre-trained models[J]. Journal of Data Acquisition and Processing，2026，41(2)：397-415.

## 引　　言

在复杂的声学感知环境中，人类听觉系统具备极其敏锐的目标音频提取能力，即能够从背景噪声或多路并发的交谈声中，通过注意力机制锁定并提取出特定的目标音源，这通常被称为“鸡尾酒会效应”[1]。随着人工智能技术的发展，如何利用算法在数字信号处理中模拟这一过程已成为人机交互、智慧办公及多媒体取证等领域的重要研究和应用课题。20世纪90年代以来，人们持续开展了大量关于音源分离与目标音频提取的探索工作。从早期的独立成分分析(Independent component analysis，ICA)[2]到基于深度神经网络的置换不变训练(Permutation invariant training，PIT)[3]，研究者们逐步解决了同质音源的分离难题。

区别于旨在无差别剥离所有声源的传统盲音源分离(Blind source separation，BSS)[4]技术，目标音频处理(Target sound processing，TSP)技术则侧重于利用用户提供的多模态线索(如自然语言指令、参考音频或目标声纹等)作为先验引导，从混合信号中定向提取或识别出特定声源。这种范式的转变，使其研究重心从纯粹的信号解耦，扩展为多模态先验信息与底层声学表征的深度对齐与融合。尽管早期的分离算法在模型驱动下取得了一定进展，但在面对非稳态噪声及复杂非线性声场时，其泛化能力和处理精度仍存在明显不足。近年来，深度学习的进展为该领域带来了转机，通过大规模数据的特征学习，系统在特定任务下的目标音频提取性能得到了显著提升。然而，现有技术在处理海量参数带来的

**收稿日期：**2026-01-19；**修订日期：**2026-03-09

计算冗余,以及在缺乏标注数据的开放场景下实现高效迁移方面,依然面临严峻挑战。

具体来说,在目标说话人语音识别(Target speaker automatic speech recognition, TS-ASR)[5-8]领域,研究重点正在从繁琐的级联系统向端到端架构演进,旨在降低模块间误差累积的影响。然而,现有的研究成果在适配大规模预训练模型时,往往陷入全量微调成本过高与轻量化微调精度受损的权衡困境。此外,在目标音频提取(Target sound extraction, TSE)[9]任务中,现有方法普遍面临泛化能力差、依赖大规模人工精标数据等问题,这极大限制了系统向更广泛、无标注的互联网级数据进行扩展的能力,也使得模型在零样本泛化上的表现难以达到工业化应用的要求。本文重点描述作者团队近年在相关领域的研究进展。

为了突破上述技术瓶颈,充分利用预训练大模型的先验表征能力,作者团队围绕语义保持、参数效率及跨模态泛化等核心维度展开了深入研究,并取得了系列进展。首先,针对多说话人重叠识别的复杂性,作者团队提出了一种基于预训练Whisper模型[10]的轻量化提示微调框架[11]。该方法通过将目标声纹特征编码为可学习的提示向量,仅需微调极小比例的参数量,即可在保持模型原始语言处理能力的基础上,实现对目标音源的高精度识别。其次,为了兼顾提取音频的听感质量与语义可懂度,作者团队构建了基于最优传输条件流匹配(Optimal transport conditional flow matching, OT-CFM)[12-14]的生成式语音提取方法[15],利用预训练编码器提取的语义线索引导声学重建过程,有效解决了传统判别式方法易产生的音频失真问题。在进一步提升系统泛化性方面,作者团队引入了对比学习语言音频预训练(Contrastive language-audio pretraining, CLAP)[16-17]技术,设计了层次化编码器以支持文本与音频双模态的灵活查询[18]。通过引入低秩适配(Low-rank adaptation, LoRA)[19]微调策略,系统不仅能够识别特定的语音,还能对自然界中的各类环境声音进行普适化提取,且在零样本测试中表现出卓越的鲁棒性。针对标注数据稀缺的难题,作者团队提出了无配对数据训练范式,利用大语言模型(Large language model, LLM)生成伪标签,并借助跨模态对齐空间实现语义关联,从而摆脱了对高质量人工标注数据的过度依赖[20]。

## 1 目标音频处理问题描述

首先对目标音频处理任务进行形式化描述。设混合音频信号为$\boldsymbol{x}\in\mathbf{R}^N$,其由多个声源组合而成,可分解为目标声源$\boldsymbol{x}_t\in\mathbf{R}^N$以及其余干扰声源$\boldsymbol{v}\in\mathbf{R}^N$的叠加

$$\boldsymbol{x}=\boldsymbol{x}_t+\boldsymbol{v} \tag{1}$$

目标音频处理旨在依据用户提供的线索$\boldsymbol{c}$,从混合信号$\boldsymbol{x}$中恢复或识别出指定的目标声源$\boldsymbol{x}_t$。该过程可通过参数化神经网络模型$\mathcal{F}(\cdot)$实现,其数学表达为

$$\hat{\boldsymbol{x}}=\mathcal{F}(\boldsymbol{x},\boldsymbol{c};\theta) \tag{2}$$

式中:$\hat{\boldsymbol{x}}$表示模型对目标声源的处理结果,$\theta$为网络参数。根据具体任务不同,$\hat{\boldsymbol{x}}$的含义有所区别:在目标音频提取任务中,$\hat{\boldsymbol{x}}$是对目标声源$\boldsymbol{x}_t$的波形估计,而在目标说话人语音识别任务中,$\hat{\boldsymbol{x}}$对应识别出的文本词元序列。引导网络执行特定操作的条件嵌入$\boldsymbol{c}\in\mathbf{R}^{d_c}$通过多模态查询编码器获得,用于在模型中指定待处理的目标对象。

## 2 目标音频处理与预训练模型概述

### 2.1 音源分离方法概述

音源分离研究的早期阶段主要采用传统信号处理技术,例如ICA[2]与非负矩阵分解(Non-negative matrix factorization, NMF)[21]。这些模型驱动的方法对信号的先验知识依赖较强,在处理复杂声学场景时面临挑战。随着深度学习技术的发展,数据驱动的深度学习方法通过从大量音频数据中自动学习

特征表示,展现出更强的鲁棒性,并逐渐成为该领域的主流。Wang等[22]提出的基于理想二值掩码的方法是一项重要突破,该方法利用深度神经网络预测短时傅里叶变换(Short-time Fourier transform,STFT)频谱的掩码,实现了语音与噪声的分离,如图1所示[23]。

如图2所示[23],典型的基于深度学习的分离系统框架通常包含3个核心组成部分:编码器、掩码估计网络和解码器。

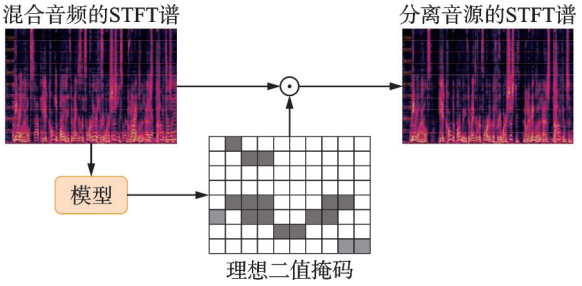


图1　理想二值掩码音源分离[23]

Fig.1　Ideal binary mask-based sound source separation[23]

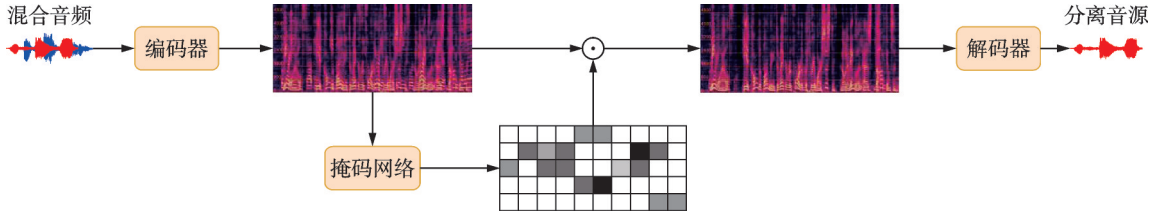


图2　基于时频掩码的音源分离系统框图[23]

Fig.2　Block diagram of a time-frequency mask-based sound source separation system[23]

首先,编码器负责将输入的时域混合音频信号$\boldsymbol{x}$转换至一个更适合分析的变换域或特征空间,得到其特征表示为

$$\boldsymbol{X}=\mathrm{Encoder}(\boldsymbol{x}) \tag{3}$$

最常用的编码器之一是短时傅里叶变换,此时$\boldsymbol{X}$为复数形式的时频谱图,包含幅度谱$|\boldsymbol{X}|$与相位谱$\phi_X$。另一种趋势是采用参数可学习的一维卷积神经网络直接从原始波形提取特征,形成多通道的二维特征图,这种方法能自适应地从数据中学习更具判别性的表示[24]。

随后,掩码估计网络以前一步得到的特征$\boldsymbol{X}$作为输入,其功能是计算并输出对应于目标声源的时频掩码

$$\boldsymbol{M}=\mathrm{MaskNet}(\boldsymbol{X}) \tag{4}$$

掩码估计网络通常是一个可学习的深度神经网络,早期工作常使用多层感知机或卷积神经网络,而近年来研究者多使用Transformer等结构来更好地建模音频信号的长距离依赖关系,以期提升掩码估计的精度[25]。

最后,解码器执行信号重建。将估计的掩码$\boldsymbol{M}$与混合特征$\boldsymbol{X}$逐点相乘,得到目标声源的估计表示$\hat{\boldsymbol{X}}$,再经解码器恢复为时域波形$\hat{\boldsymbol{x}}$

$$\hat{\boldsymbol{x}}=\mathrm{Decoder}(\boldsymbol{M}\odot\boldsymbol{X}) \tag{5}$$

解码器与编码器需配对设计:若编码器为短时傅里叶变换,则解码器为其逆变换,重建时通常沿用混合信号的相位谱;若编码器为可学习的卷积模块,则解码器通常对应为转置卷积模块。

为训练上述分离模型,需要定义合适的损失函数以衡量估计信号与真实目标信号之间的差距。常见的目标函数包括在时域直接计算的误差,如平均绝对误差(Mean absolute error, MAE)

$$\mathcal{L}_{\mathrm{MAE}}=\frac{1}{N}\|\boldsymbol{x}-\hat{\boldsymbol{x}}\|_1 \tag{6}$$

以及均方误差(Mean square error, MSE)

$$\mathcal{L}_{\mathrm{MSE}}=\frac{1}{N}\|\boldsymbol{x}-\hat{\boldsymbol{x}}\|_2^2 \tag{7}$$

式中：$\|\cdot\|_1$代表$L_1$范数，$\|\cdot\|_2$代表$L_2$范数，$\boldsymbol{x}$、$\hat{\boldsymbol{x}}$分别表示目标音频与估计音频。

这些损失函数计算简便，但可能无法完全对应听觉感知质量。信号失真比（Signal-to-distortion ratio, SDR）损失可以更好地反映目标音频与估计音频间的能量关系，更接近于客观语音质量指标

$$\mathcal{L}_{\mathrm{SDR}}=-\mathrm{SDR}(\hat{\boldsymbol{x}},\boldsymbol{x})=-10\times\lg\frac{\|\boldsymbol{x}\|_2^2}{\|\boldsymbol{x}-\hat{\boldsymbol{x}}\|_2^2} \tag{8}$$

考虑到听觉感知具有尺度不变性，尺度不变的信噪比（Scale invariant signal-to-distortion ration, SI-SDR）损失被提出并证明更为有效[26]。它通过将估计信号投影到目标信号方向上来消除幅度影响。

$$\mathcal{L}_{\mathrm{SI\_SDR}}=-\mathrm{SI\_SDR}(\hat{\boldsymbol{x}},\boldsymbol{x})=-10\times\lg\frac{\left\|\frac{\hat{\boldsymbol{x}}^{\mathrm{T}}\boldsymbol{x}}{\|\boldsymbol{x}\|_2^2}\boldsymbol{x}\right\|_2^2}{\left\|\frac{\hat{\boldsymbol{x}}^{\mathrm{T}}\boldsymbol{x}}{\|\boldsymbol{x}\|_2^2}\boldsymbol{x}-\hat{\boldsymbol{x}}\right\|_2^2} \tag{9}$$

此外，为了进一步提升分离结果的听觉自然度与细节保真度，研究者还提出了结合多分辨率时频分析损失等更为复杂的训练目标[27]。

对于语音分离或音乐源分离等任务，由于混合声源具有同质性，监督学习面临标签置换问题的挑战。深度聚类[28]与置换不变训练[3]是解决该问题的两大关键技术，它们共同构成了现代深度学习音源分离的基石。近年来，音源分离研究持续演进。Luo等[24]提出的时域语音分离网络直接以波形为输入，提高了模型设计的灵活性。随着Transformer架构的兴起，Subakan等[25]将其应用于音频分离，提出了SepFormer模型，利用自注意力机制学习长距离依赖以提升性能。在音乐源分离方面，Défossez等[29]提出的Demucs模型实现了全带宽立体声音乐的源分离，而Luo等[30]则通过频分循环神经网络（Band split recurrent neural network, BSRNN）来平衡高采样率音乐分离时的性能与计算效率。

### 2.2 目标音频处理方法概述

目标音频处理技术旨在从包含多个声源的混合信号中根据用户指令提取或处理特定目标声音，其发展建立在音源分离技术的基础之上，并针对语音、音乐及广义声音事件等不同领域演化出多个专门的研究方向。接下来回顾作者团队主要关注的目标音频提取、目标说话人语音提取和识别领域的传统研究方法与发展现状。

（1）目标音频提取

目标音频提取旨在依据用户指令从混合音频中提取指定类型的声音，与旨在分离所有声源的无条件普适音频分离模型[31-32]不同，该技术通过引入用户查询来明确目标，简化任务。典型系统由查询网络和分离网络构成，支持标签[33-35]、音频[36-38]、视觉[39-40]及自然语言[41-42]等多模态查询。然而，现有方法存在诸多局限。基于预定义标签的方法泛化性弱[34]，基于自然语言查询的方法虽更灵活，但面临联合训练困难与泛化能力不足的挑战[41]。一种改进方案是采用冻结的预训练跨模态模型（如CLIP[39]或CLAP[16-17]）作为查询网络以利用其语义先验知识，但分离网络仍需从头训练，未能充分继承知识[43]。此外，现有系统大多仅支持正向“提取”指令，对包含“去除”等复杂查询的支持不足[44]。最后，现有方法严重依赖海量音频-文本配对数据[39, 41-42]。探索利用对比学习预训练模型的模态对齐能力以减少对配对数据的依赖[45-47]，并设计能紧密耦合预训练知识、支持灵活多价态查询的架构，是提升系统泛化性与实用性的关键。

(2)目标说话人语音提取

目标说话人语音提取旨在从混合语音中直接重建目标说话人的纯净音频。现有方法可分为判别式与生成式两类。判别式方法通常通过预测时频掩码并对混合频谱进行滤波来实现,其核心思想可追溯至理想二值掩码[22]。Žmolíková 等[48]和 Delcroix 等[49]最早探索利用卷积神经网络(Convolutional neural network, CNN)或长短时记忆网络(Long short-term memory, LSTM)进行掩码建模。近年来,Transformer[50]等先进架构及频谱分割[51]等新技术的引入进一步提升了掩码估计精度。这类方法通常利用从注册语音中提取的全局声纹嵌入来引导分离,但可能忽略局部信息,因此近期研究开始探索引入更细粒度的目标说话人提示[52-53]。

由于掩码预测不完美可能导致语音过抑制或欠抑制,生成式方法逐渐受到关注。此类方法直接建模目标语音的条件分布,通过采样生成而非滤波来重建语音,从而普遍提升了感知质量[54]。Kamo 等[55]提出了基于条件扩散模型的生成式语音提取方法。Yu 等[54]和 Tang 等[56]则关注通过语言建模方法生成高音质的语音标记。然而,当前的生成式方法在追求高音质的同时,有时可能忽视重生成语音的可懂度,导致语义信息损失。

(3)目标说话人语音识别

目标说话人语音识别是一个融合了音源分离与自动语音识别(Automatic speech recognition, ASR)的交叉领域。最初的方案多为级联系统,即先分离语音再进行语音识别。尽管级联系统易于构建,但存在误差累积问题。因此,能够进行端到端联合优化的方法更受青睐。近年来,一些研究者结合置换不变训练[3, 57-58]与序列化输出训练[59-60]的技术,使得模型能够直接转写混合语音中的所有说话人内容。

进一步地,为专注于特定目标说话人,Kanda 等[61]提出了基于说话人属性的自动语音识别(Speaker attributed-ASR, SA-ASR),而 TS-ASR[5-8]作为其子方向,仅需目标说话人的少量注册信息,任务定义更为聚焦。然而,现有方法大多需要从头训练专用模型[5-6]或对预训练模型进行全参数微调[7-8],计算与存储成本高昂。随着参数高效微调(Parameter-efficient fine-tuning, PEFT)技术[19, 62-64]的发展,特别是提示微调(Prompt tuning, PT)[62, 64]在自然语言处理领域[65]的成功,如何将其有效迁移至音频领域,以实现对大规模预训练 ASR 模型的高效适配,成为一个具有潜力的研究方向。

### 2.3　预训练模型与参数高效微调方法概述

在目标音频处理领域,面对复杂多变的真实声学场景与有限的任务特定标注数据,传统的专用模型常面临泛化能力不足与训练成本高昂的挑战。预训练模型技术的出现为应对这些挑战提供了新思路。通过在超大规模、多样化的数据集上进行前置学习,模型能够捕获通用且鲁棒的多模态特征表示。模型丰富的先验知识极大地增强了对下游任务的适应能力与性能,特别是在数据稀缺的情况下。

所谓的“预训练”环节是指在模型正式适配特定目标任务之前,先行在海量的通用数据集上进行大规模特征学习的过程。这一阶段的核心目标是使模型能够捕获到具有普遍意义的数据表征,从而在后续的特定任务迁移中表现出更强的泛化潜力和稳定性。特别是在下游任务数据体量稀缺的情况下,通过预训练模型进行的迁移学习能够显著弥补样本不足导致的性能瓶颈。对于目标音频处理任务来说,迁移预训练音频模型中包含的音频先验知识对提升方法性能和泛化能力具有重要意义。

通常,不同的预训练方案主要在以下两个维度呈现差异:(1)模型架构设计,即模型内部可训练参数的具体拓扑与组织形态;(2)代理任务设置,这决定了模型参数的优化路径,涵盖了具体的训练目标函数以及优化器的配置细节。在音频领域,典型的预训练模型包括 wav2vec[66]、HuBERT[67]、Wav-LM[68]等,这些模型基于 Transformer 结构通过掩码令牌预测(Masked token prediction, MTP)代理任务学习通用语音表征,在 ASR[69-71]、说话人验证(Speaker verification, SV)[72]、语音情感识别(Speech emotion recognition, SER)[73-74]等下游任务中表现出了优秀的迁移能力。

本节重点介绍在目标音频处理领域具有广泛应用场景的预训练音频模型,即具备强鲁棒性的大规

模弱监督语音识别模型Whisper[10]以及基于对比学习机制的语言-音频预训练模型CLAP[16-17]。这些模型是作者团队基于预训练模型的目标音频处理相关研究的基础。此外，本节简要介绍作者团队所使用的提示微调方法。

(1)大规模弱监督鲁棒语音识别模型

由OpenAI推出的Whisper模型是目前弱监督语音学习领域的杰出代表。该模型采用了经典的Transformer编码器-解码器对称结构，并将语音转写及其派生子任务定义为预训练代理任务。其整体架构由音频编码器(Audio encoder, AE)与文本解码器(Text decoder, TD)协同构成，如图3所示[23]。图中Prev、SOT、EN、Transcribe、No timestamps均为特殊token，用以指示模型所执行的具体任务。具体地，SOT为开始转录(Start of transcript)；第4个token的位置(即EN的位置)表示语言标签(Language tag)，EN为English的语言代码，表示该文本输入为英文。

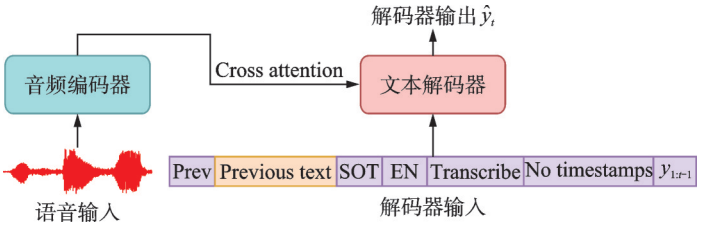


图3 Whisper模型示意图[23]

Fig.3 Whisper model schematic diagram[23]

在处理流程中，Whisper首先接收一个对数梅尔频谱图作为原始音频输入。音频编码器通过两层一维卷积神经网络实现维度的升维以及时域上的降采样，随后利用堆叠的Transformer编码器块将语音特征转化为深层隐藏向量。文本解码器则负责在隐藏向量的基础上执行词元的逐个生成过程。在预测第$t$个词元时，解码器不仅参考了已生成的历史序列，还受到如语种识别、时间戳预测标记等特定任务控制指令的约束，从而综合这些信息来预测当前时刻的输出词元。

通过在多达$68\times10^4$ h的多语言语音数据上进行训练，Whisper家族构建了从Tiny(约40M参数)到Large(约1.5B参数)等多个尺度的模型矩阵，展现了极强的跨语言识别、翻译及长文本对齐能力。

(2)对比学习语言-音频预训练模型

CLAP模型代表了音频建模的另一条重要技术路线——跨模态语义对齐。该模型采用了一种“双塔”结构，分别利用两个独立的Transformer编码器来处理音频信号与自然语言文本，如图4所示[23]。

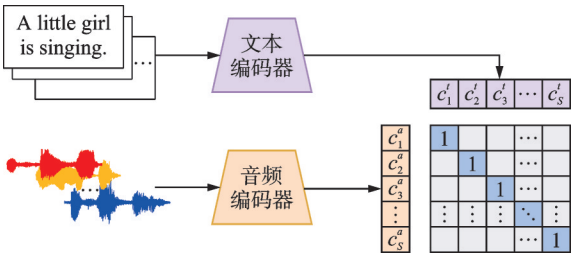


图4 对比学习语言音频预训练模型示意图[23]

Fig.4 Schematic diagram of contrastive language audio pre-training model[23]

CLAP模型的核心代理任务是对比学习，其本质是通过最大化匹配对(正样本)的余弦相似度，并同步最小化不匹配对(负样本)的相似度，从而在共享的嵌入空间中实现音频语义与文本描述的深度对齐。具体而言，音频特征序列与文本词元序列分别经由各自的编码器及多层感知机(Multi-layer perceptron, MLP)映射为音频和文本向量，其损失函数拉近匹配的音频及其描述文本特征向量，推远不匹配的音频及其描述文本特征向量，确保了模型能够学习到跨模态的通用语义表征。这种对齐后的嵌入空间为后续的音频检索、自动标注以及多模态生成任务提供了坚实的语义基础。

(3)提示微调

高效地将预训练模型迁移到目标音频处理任务中是基于预训练模型的目标音频处理方向另一个核心研究点。常用的参数轻量化微调方法包括适配器[63]、低秩适配[19]、提示微调[62]等，其中提示微调将“提示词工程”的思想从手工设计层面上升到了参数优化层面。该技术的核心在于通过在预训练模型的输入端附加特定的“提示信息”，以此引导模型激活与特定任务相关的潜在能力。

按照提示信息的性质,该技术可细分为硬提示与软提示两种。硬提示又称离散提示,指由人工设计的自然语言指令,其本质是固定的词元嵌入;软提示是将一组可学习的连续向量,直接级联在输入序列的前端,通过反向传播算法针对下游任务进行端到端的优化。具体来说,若原始输入为$\boldsymbol{X}$,则提示微调后的输入序列表示为$\boldsymbol{X}'=[\boldsymbol{P},\boldsymbol{X}]$,其中$\boldsymbol{P}$在软提示方案中代表可训练的参数向量。通过优化$\boldsymbol{P}$并冻结预训练模型的原始参数可以使模型在仅训练极少数参数的情况下高效适配下游任务。

提示微调不仅在参数规模上表现出极致的高效性,还具备卓越的推理优势。由于它不改变模型内部结构,仅在输入侧进行调整,因此能够支持多任务并行推理:即在一个批次中,将属于不同任务的多个提示向量分别拼接至对应的输入前,通过单次前向传播同时完成多个任务的处理,极大地提升了模型在实际应用场景中的吞吐量。

## 3　基于预训练模型的目标音频处理方法

本节系统阐述了作者团队在基于预训练模型的目标音频处理领域所取得的核心研究成果。针对现有技术在跨模态泛化能力弱、大规模精标数据依赖度高以及生成音频语义易丢失等瓶颈问题,作者团队从语义保持、参数效率及开放域适配等多个维度出发,提出了一系列创新性的解决方案。

具体来说,本节将重点介绍以下4个方向的突破:首先是利用对比语言-音频预训练模型构建的CLAPSep框架,实现了支持文本与音频双模态灵活查询的通用音频提取。其次是针对互联网级无标签数据,提出了无需配对数据的检索增强型训练范式,显著降低了模型对人工标注的依赖。在语音提取方面,通过多任务学习将Whisper的语义先验引入生成式架构,在提升听感质量的同时解决了语义失真问题。最后,针对目标说话人识别任务,探索了基于提示微调的轻量化适配方案,实现了在极低参数更新量下的高效精准转写。这些研究共同推动了目标音频处理技术向更具泛化性与实用性的方向演进。

### 3.1　基于对比学习的多模态查询目标音频提取方法

在开放声学场景中,由于潜在音源类别繁多,对所有音源进行逐一分离极为困难。为应对这一挑战,查询驱动的目标音频提取范式应运而生,其核心思路是根据用户提供的参考信息(即查询),从混合信号中仅分离出所需的目标声音。该类系统通常由查询网络与分离网络两部分构成:前者负责将用户查询转化为可供分离网络理解的条件特征,后者据此执行目标音源的提取。

早期该类方法如文献[41]采用BERT[75]等预训练语言模型作为查询编码器,但此类模型缺乏对文本-音频跨模态关联的建模能力,导致系统需同时学习模态对齐与音源分离,训练收敛困难且易过拟合。对比语言-音频预训练模型CLAP的出现有效解决了这一问题。CLAP通过音频编码器与文本编码器将两种模态映射至共享嵌入空间,天然具备跨模态对齐能力,可直接作为查询网络使用而无需微调,并支持文本与音频查询的灵活切换。AudioSep[42]利用了这一策略,取得了相较于LASS[41]的显著提升,但其分离网络仍从随机初始化开始训练,缺乏与查询网络之间的语义先验关联。

本节重点综述作者团队在该方向上的最新进展CLAPSep[18]。如图5所示[18,23]该模型由多模态查询网络、层次化音频编码器和分离解码器3部分构成。查询网络直接复用CLAP预训练的文本编码器与音频编码器,将文本和音频查询映射至共享的多模态对齐嵌入空间,并通过随机线性插值策略[76]缓解模态差异问题[77],使模型在推理时能够灵活响应文本、音频或二者的组合查询。此外,该方法同时支持正向查询与负向查询,正向查询用于描述需要提取的目标音源,负向查询则用于指示需要抑制的干扰成分,两类查询嵌入拼接后构成条件嵌入。层次化音频编码器则复用CLAP中基于HTS-AT[78]的预训练音频编码器,从混合音频的梅尔频谱中逐层提取多尺度声学特征,并通过LoRA[19]进行轻量级参数更新以适配下游任务。分离解码器采用U-Net[79]结构,利用FiLM[80]机制将条件嵌入注入各层级音频特征,经跳跃连接聚合后估计时频掩码,最终结合混合音频的相位信息通过逆短时傅里叶变换重建目

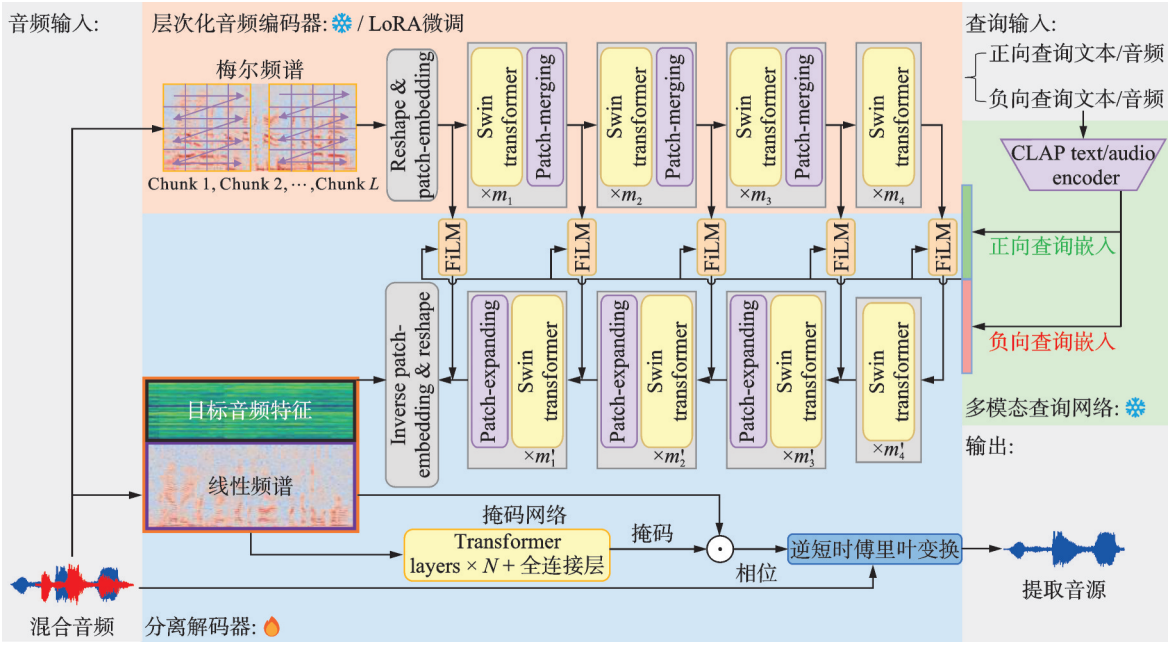


图5 CLAPSep的整体架构[18, 23]

Fig.5 Overall architecture of CLAPSep[18, 23]

标音源波形。得益于在查询网络与分离网络中同时引入预训练CLAP编码器所形成的紧密耦合架构，CLAPSep仅需约$5\times10^4$段音频进行训练，这一数据规模远少于AudioSep[42]等未利用预训练模型的方法所使用的两百余万段大规模标注数据，却在AudioCaps、ESC-50和FSDKaggle2018等多个基准数据集上取得了更优的分离性能。

为进一步验证模型在开放场景下的主观听感与分离鲁棒性，作者团队在线演示页面(https://aisa-ka0v0.github.io/CLAPSep_demo/)公开了丰富的目标音频提取示例。该平台不仅提供了大量基于真实复杂声学环境的目标音频提取音频样例，还针对前文所述的文本、音频及二者组合的灵活查询方式，以及正负向查询机制在抑制干扰与锁定目标时的协同作用进行了详细的听觉效果对比。

### 3.2 无需配对数据的语言查询目标音频提取方法

尽管预训练模型和大规模配对数据有助于增强目标音频提取系统的泛化性能，互联网上大量音频数据缺少文本标注这一现实仍然制约着模型的进一步扩展。类似问题已在文本生成图像[45]、文本生成音频[46]等跨模态任务中得到关注，其共同策略是借助对比学习预训练模型的模态对齐空间来降低对标注数据的需求。

在全监督场景下，训练数据以音频-文本配对形式存在，即$\boldsymbol{d}=\{\boldsymbol{x},\boldsymbol{y}\}$，其中$\boldsymbol{x}$为音频信号，$\boldsymbol{y}$为对应的自然语言描述。训练时，从数据集中随机采样若干音频片段，选取其中之一作为目标音源$\boldsymbol{x}$，将其余片段叠加为干扰信号$\boldsymbol{v}$，由此合成混合信号$\tilde{\boldsymbol{x}}=\boldsymbol{x}+\boldsymbol{v}$。提取模型$\mathcal{F}(\cdot)$以混合信号与条件嵌入$\boldsymbol{c}$为输入，学习恢复目标音源的映射关系：$\hat{\boldsymbol{x}}=\mathcal{F}(\tilde{\boldsymbol{x}},\boldsymbol{c};\boldsymbol{\theta})$，其中条件嵌入由预训练CLAP文本编码器对查询文本编码获得，即$\boldsymbol{c}=\mathrm{CLAP}_{\mathrm{text}}(\boldsymbol{y})$。该框架虽然直观有效，但其可行性完全依赖于大规模音频-文本配对标注数据的可获得性，而现实中此类数据的采集成本高昂且规模有限。

当缺少音频-文本配对标注时，核心挑战在于如何在没有文本描述的前提下构造有效的条件嵌入。一种直观方案是利用预训练CLAP音频编码器直接对目标音频编码以获取条件嵌入，即$\boldsymbol{c}=$

$\mathrm{CLAP}_{\mathrm{audio}}(\boldsymbol{x})$,并在推理阶段切换为文本编码器处理用户查询。该方案依赖CLAP预训练所建立的跨模态对齐嵌入空间,但由于音频与文本嵌入之间的对齐并不完美,两种模态各自保留了特有信息,导致训练与推理阶段的模态不一致[77, 81],进而影响分离性能。

为缓解上述模态差距,已有多项工作[45, 47, 82]采用高斯噪声注入方法,将音频嵌入与文本嵌入之间的偏差建模为零均值、方差为$\varepsilon$的高斯分布,并在训练阶段对音频条件嵌入添加随机扰动:$\boldsymbol{c}=\mathrm{CLAP}_{\mathrm{audio}}(\boldsymbol{x})+\boldsymbol{n}$,其中$\boldsymbol{n}\sim N(0,\varepsilon\boldsymbol{I})$。该策略已在多种跨模态任务中展现出有效性,但与基于人工标注配对数据的全监督训练相比仍存在一定差距[47]。其根本原因在于高斯噪声仅能粗略模拟模态间的统计偏移,无法精确刻画音频嵌入向文本嵌入的语义映射关系。

为解决这一问题,作者团队的最新研究进一步提出了基于检索的无配对数据训练方案,其整体流程如图6所示[20,23]。在训练开始之前,首先利用大语言模型生成大量多样化的音频描述文本(无需真实音频参与),经CLAP文本编码器编码后存入嵌入缓存。训练阶段,CLAP音频编码器提取目标音频嵌入,并通过检索策略从缓存中匹配语义最近的文本嵌入,将其附加高斯噪声注入后作为分离网络的条件输入;推理阶段则直接由CLAP文本编码器对用户查询进行编码。该检索机制使训练与推理时的条件嵌入均来自文本模态,从根本上消除了模态不一致问题,同时规避了音频嵌入中细粒度声学信息的泄露。实验表明,该方案通过大规模无标注的数据集增强训练,在多个基准评测上实现了1~2 dB的性能提升,优于传统的全监督配对数据训练方案。

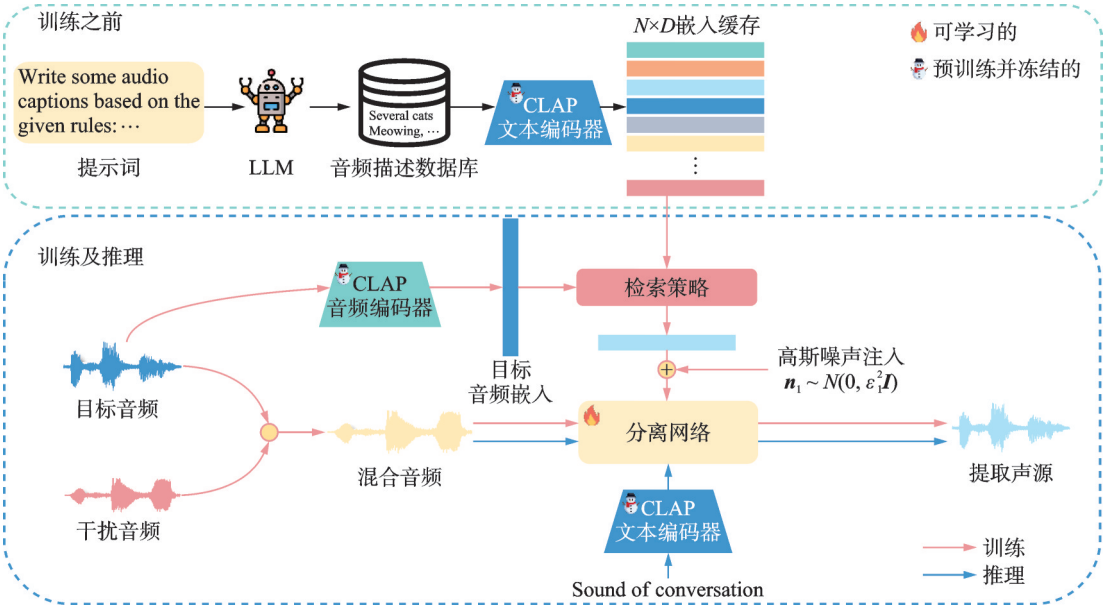


图6　基于检索的无配对数据训练方案流程[20, 23]

Fig.6　Search-based unpaired data training scheme workflow[20, 23]

### 3.3　基于多任务学习的目标说话人语音提取方法

在多说话人场景中,除了获取转写文本之外,直接从混合信号中还原目标说话人的原始语音同样具有重要的应用价值。目标说话人语音提取旨在借助目标说话人的辅助参考信息,将其语音从混合信号中分离出来。

目前,TSE方法大致分为两大类:判别式方法与生成式方法。判别式方法[48-49, 83]的核心思路是在时频域中估计目标语音的掩码,并将其施加于混合信号的频谱表示上,从而直接缩小模型输出与目标

语音之间的差距。然而,受限于掩码估计的精度,此类方法难以避免对目标信号的过度压制或对干扰信号的残留泄漏,进而影响重建语音的感知质量。生成式方法[54-56]则通过对干净目标语音的条件概率分布进行直接建模来生成语音,通常能获得更优的听感。然而,已有的生成式方案较少关注重建语音的可懂度,存在语义内容缺失或失真的风险。

针对上述问题,作者团队提出了一种多任务联合训练范式,在生成式TSE框架下兼顾高感知质量与高可懂度。该方法的核心灵感来源于Whisper模型的多说话人语义聚焦能力。尽管Whisper仅在单说话人数据上训练,但已有研究[84-85]表明,通过注入目标说话人参考信息,其编码器能够在重叠语音中有效捕捉目标说话人的语义表征。基于这一发现,该方法以预训练Whisper为骨干网络,将其语义建模优势引入生成式语音提取流程。

如图7所示[15,23],多任务学习的目标说话人语音提取方法以预训练Whisper音频编码器为基础构建共享的目标语音编码器,通过说话人嵌入和原始注册语音引导编码器聚焦目标信号。编码器输出的目标语音标记被送入两个并行分支:(1)基于OT-CFM[12-14]的语音合成模块,用于生成高质量梅尔谱;(2)预训练文本解码器,用于转写目标语音文本,为语音可懂度提供额外的训练监督。最终由HiFi-GAN[86]声码器将梅尔谱还原为波形。

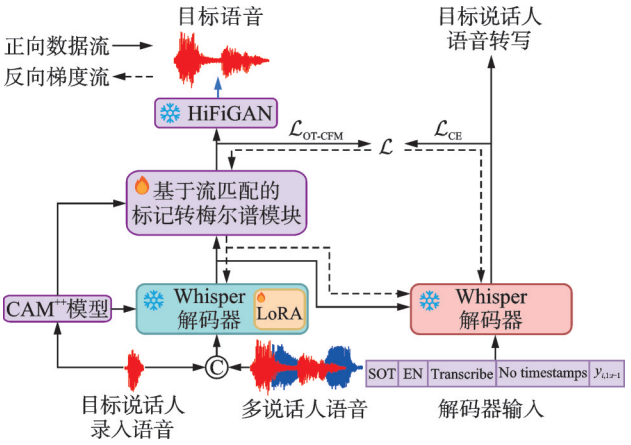


图7 多任务学习的目标说话人语音提取方法[15,23]

Fig.7 Multi-task learning approach for target speaker speech extraction[15,23]

具体而言,该模型的目标语音编码器以Whisper预训练音频编码器为基础构建。为使其适配目标语音提取任务,该方法在编码器输入端采用联合提示方案:将目标说话人嵌入经线性映射后与注册语音的梅尔频谱拼接在混合语音特征之前,共同作为编码器输入,其中注册语音采用额外的可学习位置编码以区别于混合语音。编码器的注意力模块通过低秩适配技术进行高效微调,在大幅减少可训练参数量的同时保留了预训练模型的原有知识。

编码器提取的目标语音标记随后被送入两个并行分支。第1个分支为基于最优传输条件流匹配[12]的梅尔谱合成器,该技术已在语音合成[13-14]、说唱生成[87]及声音转换[88]等多种语音生成任务中展现出出色的合成品质。其核心思想是构建从先验高斯分布到目标语音梅尔谱分布的概率密度演化路径,以神经网络学习驱动该演化过程的向量场,从而生成高质量的目标语音梅尔谱,最终经HiFiGAN[86]声码器转换为时域波形。第2个分支为Whisper预训练文本解码器,以自回归方式预测目标语音的文本转写,作为辅助任务迫使共享编码器保留充分的语义信息,从而缓解生成式方法中常见的语义丢失问题。训练时,模型联合最小化流匹配损失与交叉熵损失,仅更新流匹配合成器、LoRA增量参数及注册语音位置编码,其余预训练参数保持冻结。

在Libri2Mix和WSJ0-2mix两个标准基准上的实验表明,该方法相较于判别式基线在深度噪声抑制平均意见分(Deep noise suppression mean opinion score,DNSMOS)感知质量指标上取得一致提升,相较于离散域生成式基线TSELM[56]在感知质量、可懂度和音色一致性方面均表现出全面优势,同时也优于先将混合语音转写为文本再经语音合成后端CosyVoice[14]重建的级联方案。消融实验显示,移除文本解码分支后词错误率恶化近10%,证实了语义监督的关键作用;移除注册语音后词错误率和音色一致性均显著下降,凸显了局部声学信息对目标说话人识别的重要性。不过,该方法在余弦相似度上略逊

于判别式方法,反映出生成式模型在非语言细节保持方面仍有提升空间。

鉴于单纯的客观量化指标难以完全反映生成式模型在听觉维度的全貌,读者可通过访问演示平台(https://aisaka0v0.github.io/GenerativeTSE_demo/)获取直观的目标说话人语音提取结果。该页面系统性地展示了本方法相较于其他方法在音色还原和语义清晰度上的显著主观优势,从而为本节的实验结论提供了全面、客观的主观听感参照。

### 3.4　基于提示微调的目标说话人语音识别方法

在多说话人声学场景中,语音重叠现象广泛存在且说话人特征分布复杂,对混合语音中特定目标说话人的语音进行精准识别极为困难。为应对这一挑战,目标说话人语音识别范式应运而生,其核心思路是根据用户提供的目标说话人声纹参考信息,从混合语音信号中仅转写所需的目标说话人语音内容。该类系统通常由说话人信息编码模块与语音识别模块两部分构成:前者负责将目标说话人声纹嵌入转化为可供识别模块理解的条件特征,后者据此执行目标说话人语音的精准转写。

早期工作[5-6]采用从头训练的方式构建目标说话人语音识别模型,但此类方法缺乏对大规模语音先验知识的利用,导致模型训练数据需求大、收敛困难且泛化能力有限。大规模预训练语音识别模型Whisper的出现有效解决了这一问题。Whisper通过音频编码器与文本编码器将语音与文本模态映射至共享语义空间,天然具备强大的语音识别与跨模态建模能力,可直接作为基础识别模型使用。部分研究[7-8]利用该预训练模型进行全量微调以适配目标说话人识别任务,取得了相较于从头训练方法的显著提升,但其全量微调的方式会大幅增加计算成本,且易造成预训练先验知识的灾难性遗忘。

本节重点介绍作者团队在该领域的最新进展,将提示微调技术与预训练Whisper模型结合,构建了说话人信息编码与语音识别模块在语义层面紧密耦合的架构。该方法复用Whisper预训练编码器与解码器作为核心识别网络,将目标说话人声纹嵌入与可学习软提示结合作为条件信息注入模型,并引入深度提示[89]与重参数化技术[65]提升模型性能与训练稳定性。借助提示微调这一参数高效适配策略[62, 64],模型仅需优化少量任务相关参数即可适配目标说话人识别任务,显著降低了数据与计算需求,同时有效保留了预训练模型中关于语音识别的先验知识。这一设计借鉴了自然语言处理中复用预训练大模型并通过提示微调适配下游任务的成功经验。

所提基于提示微调的目标说话人语音识别方法的整体架构如图8所示[11,23],该模型在用户给定的目标说话人声纹提示条件下,从多说话人混合语音信号中转写目标说话人语音。系统支持声纹嵌入作

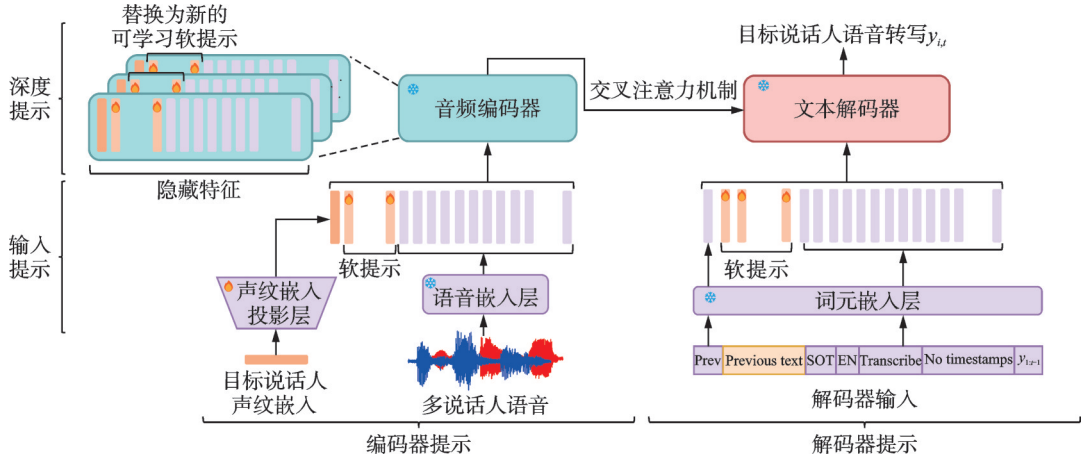


图8　基于提示微调的目标说话人语音识别方法框架[11, 23]

Fig.8　Framework of prompt-based fine-tuning of target speaker speech recognition methods[11, 23]

为核心提示信息,并允许在编码器与解码器双端注入可学习软提示,同时引入深度提示与重参数化机制优化模型性能。整个模型由说话人提示编码网络、预训练语音识别骨干网络以及提示注入与优化模块3个子模块构成:提示编码网络负责将目标说话人声纹嵌入转化为与模型维度对齐的条件特征并生成可学习软提示,预训练骨干网络复用Whisper的音频编码器与文本解码器完成语音特征提取与文本序列生成,提示注入与优化模块则通过双端提示注入、深度提示替换与重参数化操作,实现提示信息与识别网络的深度融合,最终输出目标说话人的语音转写文本。

该方法的核心思路源自提示微调技术[62, 64],即在冻结预训练模型全部权重的前提下,仅通过在模型输入端添加少量可学习的连续向量(软提示)来适配下游任务。在编码器端,首先利用可训练的线性投影层将预训练声纹识别模型提取的目标说话人嵌入向量映射至与Whisper一致的特征维度,再将其与一组编码器软提示拼接于语音特征之前送入音频编码器;在解码器端,另一组解码器软提示被插入Whisper原始任务控制词元序列的特定位置,引导文本解码器经交叉注意力机制聚焦于与目标说话人相关的编码特征,完成目标语音的转写。

然而,仅在输入层注入软提示面临两方面局限:Transformer的二次计算复杂度限制了软提示序列长度,可优化参数量因此受限;且输入层提示经多层传播后信号易衰减,难以直接影响模型输出。为此,该方法进一步引入深度提示机制[89],在每个Transformer中间层以新的独立可学习向量替换上一层对软提示位置的前向传播结果,使提示信息在网络各层持续发挥引导作用。同时,为缓解直接优化软提示带来的训练不稳定问题,采用残差重参数化策略[65]:通过前馈网络对原始软提示进行非线性变换后与其残差相加,训练完成后仅保存优化后的提示向量即可丢弃该网络,不增加推理开销。训练阶段,以最小化目标说话人转写文本的负对数似然为优化目标,仅优化软提示向量,说话人嵌入投影矩阵和重参数化网络参数,而模型其他部分的预训练权重保持不变。在Libri2Mix[90]数据集上的实验表明,该方法仅微调约1%的任务相关参数即可达到与全量微调方法[7-8]相当的识别性能,同时保留了Whisper原有的逆文本规范化与时间戳预测能力。

## 4 研究展望

本文围绕基于预训练模型的目标音频处理方法进行了系统性概述,涵盖了预训练模型的高效适配与知识迁移、生成式语音提取中的语义保持与感知质量提升、多模态查询驱动的开放场景音频分离,以及摆脱配对标注数据依赖的无监督训练范式等关键研究方向。现有工作已表明,预训练模型的引入显著提升了目标音频处理系统在开放声学环境下的泛化能力与实用性,多模态对齐嵌入空间的构建为灵活的用户交互提供了坚实基础,而基于检索的无配对训练方案则为大规模无标注数据的有效利用开辟了新路径。

尽管上述研究已取得显著进展,该领域仍面临诸多开放性挑战,以下方向值得未来研究重点关注:

(1)推理效率与轻量化部署。当前基于大规模预训练模型的方法普遍存在推理时延高、计算资源需求大的问题,难以直接应用于会议转录、智能耳机、嵌入式终端等实时性要求较高的场景。未来可借助知识蒸馏、结构化剪枝和动态推理等模型压缩技术,在尽量保持性能的前提下大幅降低模型的计算与存储开销,推动目标音频处理技术从实验室走向实际产品部署。

(2)更深层次的多模态融合。现有多模态查询方法主要围绕文本与音频两种模态展开,对视觉(如说话人唇动、场景图像)、脑电信号(如听觉注意力解码)等模态信息的利用仍然有限。未来可结合跨模态预训练模型,探索构建统一的音频-文本-视觉-脑电联合嵌入空间,实现更自然、更鲁棒的人机交互方式,并提升系统在复杂多源场景下的目标音频处理能力。

(3)开放域与长尾场景的泛化能力。尽管基于预训练模型的方法在常见声学事件上已展现出较强

的零样本泛化能力,但对于罕见声学事件(如突发性异常声、低资源语种语音)的处理仍存在明显不足。增量学习、元学习以及基于检索增强的持续学习策略有望使模型在部署后持续适应新场景和新类别,实现动态知识更新与性能提升。

(4)迈向通用目标音频处理大模型。当前研究针对不同任务(如语音识别、语音提取和声音事件分离)通常设计独立的模型与训练流程。随着大规模多任务预训练范式的成熟,未来有望构建统一的目标音频处理大模型,在单一框架下同时支持目标说话人语音识别、语音增强与提取和通用音频分离等多种任务,通过自然语言或多模态指令实现灵活的任务切换,从而显著降低系统复杂度并提升用户体验。

(5)数据驱动范式的进一步突破。无配对数据训练方案已初步证明了大规模无标注数据的利用价值,但如何更高效地挖掘互联网海量音频资源中的隐含语义信息、如何结合自监督与弱监督信号构建更优质的训练数据,仍是值得深入探索的方向。此外,合成数据与真实数据的协同利用、数据质量自动筛选机制等问题也亟待研究。

综上,基于预训练模型的目标音频处理技术正处于快速发展阶段,未来研究应在提升效率、深化多模态融合、增强泛化能力和构建统一框架等方面持续突破,推动其在人机交互、无障碍通信、多媒体内容创作及司法取证等实际应用领域的广泛落地。

## 5　结束语

近年预训练模型技术快速发展为语音与自然音频处理任务提供了新的视点与机遇,也为音频处理领域的各种下游任务提供了强大的基础能力,逐渐受到了越来越多研究者的关注。本文针对这一机遇,结合作者团队的研究,概述了基于预训练模型的目标音频处理方法。首先介绍了目标音频处理的任务定义与问题建模,随后从音源分离、目标音频提取、目标说话人语音提取和识别等方面梳理了传统方法的研究进展,并对Whisper、CLAP等预训练模型及提示微调技术进行了系统介绍。在此基础上,重点概述了作者团队在基于预训练模型的目标音频处理方法研究方面取得的进展,分别从多模态查询与泛化、无配对数据训练、生成式语音提取中的语义保持以及参数高效的目标说话人识别等角度展开论述,并对各方法的实验结果和性能进行了简要分析。最后,对该领域面临的关键挑战与未来发展方向进行了讨论。预期通过轻量化的模块设计及参数高效微调方案,基于预训练模型可以构建出具备高性能与可靠泛化能力的目标音频处理系统。提升音频处理技术的实用性与普适性,改善人机智能系统的交互体验。

**参考文献:**

[1] BRONKHORST A W. The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions[J]. Acta Acustica United with Acustica, 2000, 86(1): 117-128.

[2] BELL A J, SEJNOWSKI T J. An information-maximization approach to blind separation and blind deconvolution[J]. Neural Computation, 1995, 7(6): 1129-1159.

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

[4] ARAKI S, ITO N, HAEB-UMBACH R, et al. $30^{+}$ years of source separation research: Achievements and future challenges [C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2025: 1-5.

[5] KANDA N, HORIGUCHI S, TAKASHIMA R, et al. Auxiliary interference speaker loss for target-speaker speech recognition

[C]//Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH). Graz, Austria: ISCA, 2019: 236-240.

[6] ZHANG Y, PUVVADA K C, LAVRUKHIN V, et al. Conformer-based target-speaker automatic speech recognition for single-channel audio[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2023: 1-5.

[7] HUANG Z, RAJ D, GARCÍA P, et al. Adapting self-supervised models to multi-talker speech recognition using speaker embeddings[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2023: 1-5.

[8] ZHANG W, QIAN Y. Weakly-supervised speech pre-training: A case study on target speech recognition[C]//Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH). Dublin, Ireland: ISCA, 2023: 3517-3521.

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

[10] RADFORD A, KIM J W, XU T, et al. Robust speech recognition via large-scale weak supervision[C]//Proceedings of International Conference on Machine Learning. Hawaii, USA: [s.n.], 2023: 28492-28518.

[11] MA H, PENG Z, SHAO M, et al. Extending whisper with prompt tuning to target-speaker ASR[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2024: 12516-12520.

[12] LIPMAN Y, CHEN R T, BEN-HAMU H, et al. Flow matching for generative modeling[C]//Proceedings of 11th International Conference on Learning Representations. Kigali, Rwanda: [s.n.], 2023.

[13] MEHTA S, TU R, BESKOW J, et al. Matcha-TTS: A fast TTS architecture with conditional flow matching[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2024: 11341-11345.

[14] DU Z, CHEN Q, ZHANG S, et al. CosyVoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens[J]. arXiv preprint arXiv: 2407.05407, 2024.

[15] MA H, CHEN R, ZHANG X L, et al. Enhancing intelligibility for generative target speech extraction via joint optimization with target speaker ASR[J]. IEEE Signal Processing Letters, 2025, 32: 2309-2313.

[16] WU Y, CHEN K, ZHANG T, et al. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2023: 1-5.

[17] KIM J, JUNG J, LEE J, et al. EnCLAP: Combining neural audio codec and audio-text joint embedding for automated audio captioning[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2024: 6735-6739.

[18] MA H, PENG Z, LI X, et al. CLAPSep: Leveraging contrastive pre-trained model for multi-modal query-conditioned target sound extraction[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2024, 32: 4945-4960.

[19] HU E J, SHEN Y, WALLIS P, et al. LoRA: Low-rank adaptation of large language models[C]//Proceedings of International Conference on Learning Representations. [S.l.]: ICLR, 2022.

[20] MA H, PENG Z, LI X, et al. Language-queried target sound extraction without parallel training data[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2025: 1-5.

[21] CICHOCKI A, ZDUNEK R, AMARI S. New algorithms for non-negative matrix factorization in applications to blind source separation[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2006.

[22] WANG Y, WANG D. Towards scaling up classification-based speech separation[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2013, 21(7): 1381-1390.

[23] 马豪.基于预训练模型的目标音频处理方法研究[D].青岛: 山东大学, 2025.

MA Hao. Research on target sound processing with pre-trained models [D]. Qingdao: Shandong University, 2025.

[24] 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.

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

[26] ROUX J L, WISDOM S, ERDOGAN H, et al. SDR-half-baked or well done? [C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2019: 626-630.

[27] YU J, CHEN H, LUO Y, et al. High fidelity speech enhancement with band-split RNN[C]//Proceedings of Interspeech 2023. Dublin, Ireland: [s.n.], 2023: 2483-2487.

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

[29] DÉFOSSEZ A, USUNIER N, BOTTOU L, et al. Music source separation in the waveform domain[J]. arXiv preprint arXiv: 1911.13254, 2019.

[30] LUO Y, YU J. Music source separation with band-split RNN[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023, 31: 1893-1901.

[31] KAVALEROV I, WISDOM S, ERDOGAN H, et al. Universal sound separation[C]//Proceedings of 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). [S.l.]: IEEE, 2019: 175-179.

[32] WISDOM S, TZINIS E, ERDOGAN H, et al. Unsupervised sound separation using mixture invariant training[J]. Advances in Neural Information Processing Systems, 2020, 33: 3846-3857.

[33] OCHIAI T, DELCROIX M, KOIZUMI Y, et al. Listen to what you want: Neural network-based universal sound selector [C]//Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH). Shanghai, China: [s.n.], 2020: 1441-1445.

[34] VELURI B, CHAN J, ITANI M, et al. Real-time target sound extraction[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2023: 1-5.

[35] DELCROIX M, VÁZQUEZ J B, OCHIAI T, et al. SoundBeam: Target sound extraction conditioned on sound-class labels and enrollment clues for increased performance and continuous learning[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022, 31: 121-136.

[36] KONG Q, CHEN K, LIU H, et al. Universal source separation with weakly labelled data[J]. arXiv preprint arXiv: 2305.07447, 2023.

[37] KILGOUR K, GFELLER B, HUANG Q, et al. Text-driven separation of arbitrary sounds[C]//Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH). Incheon, Korea:[s.n.], 2022: 5403-5407.

[38] CHEN K, DU X, ZHU B, et al. Zero-shot audio source separation via query-based learning from weakly-labeled data[C]//Proceedings of the AAAI Conference on Artificial Intelligence. [S.l.]: AAAI, 2022: 4441-4449.

[39] DONG H W, TAKAHASHI N, MITSUFUJI Y, et al. CLIPSep: Learning text-queried sound separation with noisy unlabeled videos[C]// Proceedings of 11th International Conference on Learning Representations. Kigali, Rwanda:[s.n.], 2023.

[40] ZHAO H, GAN C, ROUDITCHENKO A, et al. The sound of pixels[C]//Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany: Springer, 2018: 570-586.

[41] LIU X, LIU H, KONG Q, et al. Separate what you describe: Language-queried audio source separation[C]//Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH). Incheon, Korea: [s.n.], 2022: 1801-1805.

[42] LIU X, KONG Q, ZHAO Y, et al. Separate anything you describe[J]. IEEE Transactions on Audio, Speech and Language Processing, 2024, 33: 458-471.

[43] LI C, QIAN Y, CHEN Z, et al. Target sound extraction with variable cross-modality clues[C]//Proceedings of IEEE

International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2023: 1-5.

[44] JIANG X, HAN C, LI Y A, et al. Listen, chat, and remix: Text-guided soundscape remixing for enhanced auditory experience [J]. IEEE Journal of Selected Topics in Signal Processing, 2025, 19(4): 635-645.

[45] ZHOU Y, ZHANG R, CHEN C, et al. Towards language-free training for text-to-image generation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2022: 17886-17896.

[46] LIU H, CHEN Z, YUAN Y, et al. AudioLDM: Text-to-audio generation with latent diffusion models[C]//Proceedings of International Conference on Machine Learning. Honolulu, Hawaii, USA: [s.n.], 2023: 21450-21474.

[47] DESHMUKH S, ELIZALDE B, EMMANOUILIDOU D, et al. Training audio captioning models without audio[C]// Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2024: 371-375.

[48] ŽMOLÍKOVÁ 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.

[49] DELCROIX M, OCHIAI T, ZMOLIKOVA K, et al. Improving speaker discrimination of target speech extraction with time-domain speakerbeam[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2020: 691-695.

[50] 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 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2023:1-5.

[51] LE X, CHEN L, HE C, et al. Personalized speech enhancement combining band-split RNN and speaker attentive module[C]// Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2023:1-2.

[52] ZHANG K, LI J, WANG S, et al. Multi-level speaker representation for target speaker extraction[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2025: 1-5.

[53] HE S, ZHANG H, RAO W, et al. Hierarchical speaker representation for target speaker extraction[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2024: 10361-10365.

[54] YU L, ZHANG W, DU C, et al. Generation-based target speech extraction with speech discretization and vocoder[C]// Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2024: 12612-12616.

[55] KAMO N, DELCROIX M, NAKATANI T. Target speech extraction with conditional diffusion model[C]//Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH). Dublin, Ireland: [s.n.], 2023: 176-180.

[56] TANG B, ZENG B, LI M. TSELM: Target speaker extraction using discrete tokens and language models[C]//Proceedings of National Conference on Man-Machine Speech Communication. Zhenjiang, China: Springer, 2025: 459-469.

[57] QIAN Y, CHANG X, YU D. Single-channel multi-talker speech recognition with per-mutation invariant training[J]. Speech Communication, 2018, 104: 1-11.

[58] MENG L, KANG J, CUI M, et al. A sidecar separator can convert a single-talker speech recognition system to a multi-talker one[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2023: 1-5.

[59] KANDA N, GAUR Y, WANG X, et al. Serialized output training for end-to-end over-lapped speech recognition[C]// Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH). Shanghai, China: [s.n.], 2020: 2797-2801.

[60] LI C, QIAN Y, CHEN Z, et al. Adapting multi-lingual ASR models for handling multiple talkers[C]//Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH). Dublin, Ireland: [s.n.], 2023: 1314-1318.

[61] KANDA N, YE G, GAUR Y, et al. End-to-end speaker-attributed asr with transformer[C]//Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH). Brno, Czechia: [s.n.], 2021: 4413-4417.

[62] LESTER B, AL-RFOU R, CONSTANT N. The power of scale for parameter-efficient prompt tuning[C]//Proceedings of

the 2021 Conference on Empirical Methods in Natural Language Processing. Punta Cana, Dominican Republic: [s.n.], 2021: 3045-3059.

[63] HOULSBY N, GIURGIU A, JASTRZEBSKI S, et al. Parameter-efficient transfer learning for NLP[C]//Proceedings of International Conference on Machine Learning. Long Beach, CA, USA: [s.n.], 2019: 2790-2799.

[64] LI X L, LIANG P. Prefix-tuning: Optimizing continuous prompts for generation[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. [S.l.]: [s.n.], 2021: 4582-4597.

[65] RAZDAIBIEDINA A, MAO Y, KHABSA M, et al. Residual prompt tuning: Improving prompt tuning with residual reparameterization[C]//Proceedings of Findings of the Association for Computational Linguistics. Toronto, Canada: [s.n.], 2023: 6740-6757.

[66] BAEVSKI A, ZHOU Y, MOHAMED A, et al. wav2vec 2.0: A framework for self-supervised learning of speech representations[J]. Advances in Neural Information Processing Systems, 2020, 33: 12449-12460.

[67] HSU W N, BOLTE B, TSAI Y H H, et al. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3451-3460.

[68] CHEN S, WANG C, CHEN Z, et al. WavLM: Large-scale self-supervised pre-training for full stack speech processing[J]. IEEE Journal of Selected Topics in Signal Processing, 2022, 16(6): 1505-1518.

[69] AO J, WANG R, ZHOU L, et al. SpeechT5: Unified-modal encoder-decoder pre-training for spoken language processing[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Dublin, Ireland: [s.n.], 2022: 5723-5738.

[70] RUBENSTEIN P K, ASAWAROENGCHAI C, NGUYEN D D, et al. AudioPaLM: A large language model that can speak and listen[J]. arXiv preprint arXiv: 2306.12925, 2023.

[71] SONG Z, ZHUO J, YANG Y, et al. LoRA-Whisper: Parameter-efficient and extensible multilingual ASR[C]//Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH). Kos, Greece: [s.n.], 2024: 3934-3938.

[72] CHEN S, WU Y, WANG C, et al. UniSpeech-SAT: Universal speech representation learning with speaker aware pre-training [C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2022: 6152-6156.

[73] PEPINO L, RIERA P, FERRER L. Emotion recognition from speech using wav2vec 2.0 embeddings[C]//Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH). Brno, Czechia: [s.n.], 2021: 3400-3404.

[74] SIRIWATDHANA S, REIS A, WEERASEKERA R, et al. Jointly fine-tuning "BERT-like" self supervised models to improve multimodal speech emotion recognition[C]//Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH). Shanghai, China: [s.n.], 2020: 3755-3759.

[75] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding [C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, Minnesota: [s.n.], 2019: 4171-4186.

[76] LÜDDECKE T, ECKER A. Image segmentation using text and image prompts[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2022: 7086-7096.

[77] ZHANG Y, SUI E, YEUNG-LEVY S. Connect, collapse, corrupt: Learning cross-modal tasks with uni-modal data[C]//Proceedings of The Twelfth International Conference on Learning Representations. Vienna, Austria: [s.n.], 2024.

[78] CHEN K, DU X, ZHU B, et al. HTS-AT: A hierarchical token-semantic audio transformer for sound classification and detection[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2022: 646-650.

[79] RONNEBERGER O, FISCHER P, BROX T. U-Net: Convolutional networks for biomedical image segmentation[C]//Proceedings of International Conference on Medical Image Computing and Computer-assisted Intervention. Munich, Germany: Springer International Publishing, 2015: 234-241.

[80] PEREZ E, STRUB F, DE VRIES H, et al. FiLM: Visual reasoning with a general conditioning layer[C]//Proceedings of the AAAI Conference on Artificial Intelligence. [S.l.]: AAAI, 2018: 3942-3951.

[81] LIANG V W, ZHANG Y, KWON Y, et al. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning[C]//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, Louisiana, USA: MIT Press, 2022: 17612-17625.

[82] NUKRAI D, MOKADY R, GLOBERSON A. Text-only training for image captioning using noise-injected CLIP[C]//Proceedings of Findings of the Association for Computational Linguistics. Abu Dhabi, United Arab Emirates: [s.n.], 2022: 4055-4063.

[83] 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.

[84] MENG L, KANG J, WANG Y, et al. Empowering whisper as a joint multi-talker and target-talker speech recognition system [C]//Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH). Kos, Greece: [s.n.], 2024: 4653-4657.

[85] GUO P, CHANG X, LV H, et al. SQ-Whisper: Speaker-querying based Whisper model for target-speaker ASR[J]. IEEE Transactions on Audio, Speech and Language Processing, 2024, 33: 175-185.

[86] KONG J, KIM J, BAE J. HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis[C]//Proceedings of the 34th Annual Conference on Neural Information Processing Systems (NeurIPS). [S.l.]: MIT Press, 2020: 17022-17033.

[87] NING Z, WANG S, JIANG Y, et al. Drop the beat! freestyler for accompaniment conditioned rapping voice generation[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Philadelphia, Pennsylvania: AAAI, 2025: 24966-24974.

[88] YAO J, YAN Y, PAN Y, et al. StableVC: Style controllable zero-shot voice conversion with conditional flow matching[C]//Proceedings of the AAAI Conference on Artificial Intelligence. [S.l.]: AAAI, 2025, 39(24): 25669-25677.

[89] LIU X, JI K, FU Y, et al. P-Tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Dublin, Ireland: [s.n.], 2022: 61-68.

[90] COSENTINO J, PARIENTE M, CORNELL S, et al. LibriMix: An open-source dataset for generalizable speech separation [J]. arXiv preprint arXiv: 2005.11262, 2020.

**作者简介:**

**刘琚**(1965-),**通信作者**,男,教授,博士生导师,研究方向:盲信号处理、通信信号处理、数字图像处理,E-mail: juliu@sdu.edu.cn。

**马豪**(2000-),男,博士研究生,研究方向:音频信号处理。

**李晓航**(1999-),男,博士研究生,研究方向:深度学习、光计算。

**李玉楷**(2000-),男,硕士研究生,研究方向:预训练大模型、多模态数据对齐。

**司媛**(2002-),女,硕士研究生,研究方向:医学影像处理。

**邢志坤**(2003-),男,硕士研究生,研究方向:医学图像处理。

**王芷涵**(2003-),女,硕士研究生,研究方向:语音识别。

**邵明杰**(1992-),男,副研究员,博士生导师,研究方向:最优化方法、信号处理与机器学习。

(编辑:陈珺)

# Research Progress in Target Audio Processing Methods Based on Pre-trained Models

LIU Ju[1*], MA Hao[1], LI Xiaohang[1], LI Yukai[1], SI Yuan[1], XING Zhikun[1], WANG Zhihan[1], SHAO Mingjie[1,2]

(1. School of Information Science and Engineering, Shandong University, Qingdao 266237, China; 2. Academy of Mathematics and Systems Science of the Chinese Academy of Sciences, Beijing 100190, China)

**Abstract**: Target audio processing aims to recover or identify a specific target sound source from mixed audio signals based on user-provided cues. As an important branch of audio signal processing and machine listening, it plays a vital role in a wide range of applications, including human-computer interaction, smart office environments, assistive technologies, and multimedia forensics. In recent years, the emergence of large-scale pre-trained models has opened up new possibilities for target audio processing by significantly improving representation learning, cross-modal understanding, and adaptation to low-resource conditions. This paper presents an overview of the recent research progress made by our team in this area, with particular emphasis on the integration of pre-trained models into target audio processing frameworks. First, we review the research status of several related tasks, including target speaker automatic speech recognition, speech extraction, target audio extraction, and sound source separation, and introduce representative pre-trained models such as Whisper and contrastive language-audio pretraining (CLAP) together with parameter-efficient fine-tuning strategies. Focusing on the tasks of target audio extraction and target speaker recognition, we then summarize our recent studies, including a contrastive-learning-based multimodal query method for target audio extraction, a language-queried target audio extraction method that removes the reliance on paired training data, a multitask-learning-based method for target speaker speech extraction, and a prompt-tuning-based method for target speaker automatic speech recognition. These studies have achieved substantial advances in multimodal generalization, reduction of labeled-data dependence, preservation of target semantic information, and parameter-efficient model adaptation. We further show that the combination of pre-trained models and task-oriented fine-tuning provides an effective pathway toward more robust and flexible target audio processing systems. Finally, we discuss several future research directions, including improving inference efficiency, promoting deeper multimodal fusion, enhancing open-domain generalization, and developing universal foundation models for target audio processing.

**Highlights**:

1. This work systematically elaborates on how semantic priors from models like Whisper and CLAP bridge the generalization gap in open-world acoustic environments.
2. It delves into the root challenges of deploying large-scale models, introducing prompt tuning and retrieval-augmented unpaired learning to enable high-efficiency domain transfer with minimal computational overhead.
3. Targeting the core gap of semantic distortion, this paper proposes a multi-task learning framework that integrates linguistic priors into flow-matching architectures, ensuring high-fidelity audio extraction without compromising intelligibility.

**Key words**: target audio processing; pre-trained model; parameter-efficient fine-tuning; target sound extraction; target speaker speech recognition; contrastive learning

**Received**: 2026-01-19; **Revised**: 2026-03-09

***Corresponding author, E-mail**: juliu@sdu.edu.cn.