基于时域的基频感知语音分离方法*

王 凯，李鸣鹤，黄志华，黄 浩†
(新疆大学 信息科学与工程学院，新疆 乌鲁木齐 830017)

摘 要： 传统的单通道语音分离方法主要采用混音作为输入，对其进行分离得到目标说话人的语音. 最近的研究表明，将预估计的基频信息注入到原始混音信号中能够提高分离效果，但这种方法最初应用于时频域. 近年来，基于时域的语音分离方法已经被验证优于早期的时频域分离方法. 基于上述出发点，本文提出基于辅助基频的时域语音分离方法. 该方法首先将时域信号输入预分离模块生成预分离语音，并从预分离语音中提取基频；然后将提取的基频与原始混音拼接，作为后分离模块的输入进行第二次分离. 本文评估了不同的基频提取方法和训练策略. 语音分离实验结果表明：在训练后分离模块时，先使用理想基频与混音融合训练一个理想分离网络，然后用RAPT方法对预分离源提取估计基频注入混音，再进行理想分离网络的微调，能够获得最佳的语音分离性能，比Conv-TasNet基线方法提高了0.5 dB. 这说明显式地注入辅助基频信息不仅在时频域语音分离中表现出了有效性，同时也适用于时域语音分离.

关键词： 语音分离；单通道；基频；时域
DOI： 10.13568/j.cnki.651094.651316. 2021.01.07.0002
中图分类号： TN912.3 **文献标识码：** A **文章编号：** 2096-7675(2022)02-0182-07
引文格式： 王凯, 李鸣鹤, 黄志华, 黄浩. 基于时域的基频感知语音分离方法[J]. 新疆大学学报(自然科学版)(中英文), 2022, 39(2): 182-188.
英文引文格式： WANG Kai, LI Minghe, HUANG Zhihua, HUANG Hao. Time domain speech separation using auxiliary pitch information[J]. Journal of Xinjiang University(Natural Science Edition in Chinese and English), 2022, 39(2): 182-188.

Time Domain Speech Separation Using Auxiliary Pitch Information

WANG Kai, LI Minghe, HUANG Zhihua, HUANG Hao

(*School of Information Science and Engineering, Xinjiang University, Urumqi Xinjiang 830017, China*)

Abstract： In most speech separation methods, only the mixture is used as the input. Pitch-aware architecture injects pitch information into the original mixture to improve the separation result, which was originally applied in time-frequency(T-F) domain. Based on the fact that speech separation in time domain has achieved much better performance than that in T-F domain, we investigate into the effectiveness on the utilization of auxiliary pitch information in time domain speech separation. Firstly, a pre-separation module is trained to generate pre-separated sources, from which pitches are extracted. The extracted pitches are then spliced with the original mixture as the input to a post-separation module. We evaluate different pitch trackers and training strategies. It is shown that, for training the post-separation module, the combination of pre-training on ideal pitches and then fine-tuning on estimated pitches extracted from pre-separated sources using RAPT gives the best result, achieving 0.5 dB improvement over the Conv-TasNet baseline. This indicates that the auxiliary pitch information which has shown effectiveness in T-F domain speech separation is also applicable to time domain speech separation.

Key words： speech separation; single-channel; pitch tracking; time domain

0 引 言

语音分离是语音识别、说话人分类、说话人识别等语音处理任务中必不可少的前端组成部分. 与多通道语音分离任务相比，单通道语音分离由于其所提供的信息有限而更具挑战性，也是当前语音处理研究的热点. 近年来，语音分离的研究主要集中在基于深度学习的方法上，并取得了突破性的进展. 根据对输入混音的处理方式，可将语音分离技术分为两类：基于时频域的分离方法和基于时域的分离方法. 前者的提出时间较早[1−6]，其

* **收稿日期：** 2021-01-07
基金项目： 新疆多语种信息技术重点实验室开放课题（2020D04047）；国家重点研发项目（2020AAA0107902）；国家自然科学基金项目（61663044；61761041）.
作者简介： 王凯（1982-），男，博士生，从事语音分离、深度学习的研究，E-mail: terry_wang@stu.xju.edu.cn.
† **通讯作者：** 黄浩（1976-），男，博士，教授，主要从事语音识别、多媒体人机交互技术的研究，E-mail: huanghao@xju.edu.cn.

中深度聚类[1−2]和PIT[5−6]分别解决了语音分离中的排列问题[1]，为后续的研究工作打下基础.

时频域语音分离方法中通常首先用短时傅立叶变换（STFT）得到混音的幅度谱，然后将混音以幅度谱或对数幅度谱的形式作为输入，用神经网络估计分离后的各说话人语音，在重构估计语音的时域信号时采用混音相位. 由于混音相位对于各个语音相位来说加入了干扰，会导致次优的分离结果. 为此，有研究提出了相位重建算法[7]和复数域的STFT表示[8]，但这仍需要额外的模型或处理步骤. 为了避免相位的问题，近期有文献研究使用时域表示的语音分离方法[9−12]. 这种方法将时域中的混合波形转换为非负的特定空间，在该空间中进行分离处理以计算估计的各语音. 由于相位信息隐含在原始波形中，时域方法避免了相位重建的困难，其结果通常优于基于时频域的方法.

除了将混音作为输入之外，一些研究试图添加其他先验信息来帮助改善分离结果[13−18]. 有些先验知识与原始混音正交，即它们的信息之间互相独立，例如预先收集说话人的音频、视频、图片等信息（称为预注册信息）来帮助语音分离[13]，以及记录说话人基于位置的方位角特征，以备后续分离步骤[14]. 而另一些先验信息是隐含在原始混音信号中的，例如时域混音信号的时频域表示[16]，以及各目标语音的基频信息[17]. 显式地添加隐藏在原始信号中的信息可以提高性能，究其原因是由于训练数据集的有限性和神经网络能力的局限性，使得输出结果无法完全与理想情况相同，而先验知识可以帮助神经网络进一步提升语音分离性能.

基频感知语音分离[17]是一种在时频域利用基频频率辅助信息的语音分离方法，分为预分离和后分离两个阶段. 预分离阶段训练一个基于深度聚类（DC）的模型来预分离估计语音，过程中用前馈网络代替DC中的K-means聚类过程来提升聚类效果. 然后利用一个基频提取神经网络，从预先分离的语音中提取基频. 后分离阶段将估计的基频与原始混音进行拼接作为新的输入，送入基于uPIT[6]的分离网络来计算最终的目标语音.

上述方法的不足在于DC和uPIT都是较早期的时频域分离方法，因此模型不可避免地存在相位重构问题，从而限制了其性能. 基于这一原因，我们设计了一种新的时域语音分离框架，继承了其附加基频信息的特性. 具体而言，我们采用的模型也分为预分离阶段和后分离阶段. 预分离阶段采用Conv-TasNet[10]进行分离，然后对分离后的语音分别用传统算法RAPT[19]和基于深度学习的方法估计基频. 我们发现对于预分离后的语音，RAPT比基于深度学习的方法效果更好. 后分离阶段采用Conv-TasNet的变体训练一个新的分离模型，其输入采用原始混音和相应基频的组合. 我们首先将理想的基频信息注入原始混音中，以确定基频信息是否也有助于时域语音分离. 结果表明，本文的研究方向是正确的：如果能够提取到理想的基频信息，时域语音分离的性能也可以得到改善. 然后，我们研究了不同基频跟踪方法与后分离模块的组合，以及联合训练、预训练加微调[20]等不同训练方法的效果. 实验结果表明，如果后分离模块先使用理想基频输入进行预训练，再使用RAPT算法得到的预分离语音基频进行微调，将取得最好的效果，比仅基于Conv-TasNet而没有基频信息的分离模型提高了0.5 dB.

1　基于时域的基频感知语音分离

1.1　整体架构

所提出方法的框架如图1所示，由三个模块组成：预分离模块、基频跟踪模块和后分离模块. 预分离模块用于将混合波形分离成预先分离的信号源，基频跟踪模块从中提取估计的基频. 提取的基频和原始混音进行拼接，然后输入到后分离模块进行最终分离. 由于基频频率和混音采样值之间的数量级差异较大，在拼接之前需要进行数量级的重新缩放操作. 预分离模块采用Conv-TasNet网络结构，网络配置与文献[10]中描述的相同. 对于基频跟踪模块，参考文献[17]，我们提出新的基于深度学习的基频跟踪模型，另外也验证了传统的基于RAPT的基频跟踪方法. 后续的实验部分将对它们的结果进行比较. 最后参考Conv-TasNet设计了后分离模块.

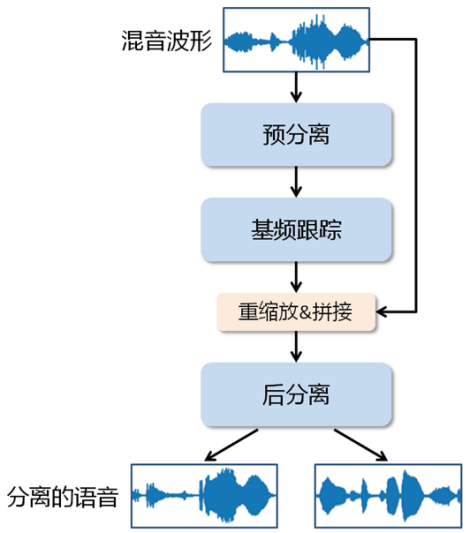

图 1　本文所提出方法的架构

1.2 Conv-TasNet简介

本文在预分离和后分离阶段采用Conv-TasNet[10]作为分离模块，该网络是一个基于时间卷积网络TCN[21]的端到端训练的全卷积网络，其输入是由时域采样得到的混音波形信号，输出为分离之后的各语音波形信号. 其结构包括三个部分：编码器、分离模块和解码器. 编码器采用一维卷积层，将混音分段并转换为非负的潜在空间的表示. 分离模块采用多个一维卷积块堆叠的结构，为混音中每个语音的每个时间步估计掩蔽，然后将各掩蔽乘以混音得到潜在空间的估计语音. 分离模块中使用多层空洞卷积[21−22]和逐层增加的膨胀因子，可以得到较大的感受野，从而对长序列建模. 解码器采用转置卷积操作，将潜在空间的语音特征转换到时域，重建估计的语音信号. Conv-TasNet的目标函数直接采用尺度不变信噪比$SI-SNR$，定义如下：

$$s_{\text{target}} := \frac{\langle \hat{s}, s\rangle s}{\|s\|^2} \tag{1}$$

$$e_{\text{noise}} := \hat{s} - s_{\text{target}} \tag{2}$$

$$SI-SNR = 10\log_{10}\frac{\|s_{\text{target}}\|^2}{\|e_{\text{noise}}\|^2} \tag{3}$$

其中：$\hat{s}$和s分别是估计语音和源干净语音，$\|s\|^2 = \langle s, s\rangle$指信号能量. 更多关于Conv-TasNet的细节可参阅文献[10].

1.3 基频跟踪模块

基频是周期信号的固有特性，基频跟踪（又称基频估计）是估计基频轮廓的任务[23]. 经典的基频估计方法包括RAPT[19]、PRAAT[24]、YIN[25]等，通过计算语音或音乐的局部极大值或极小值来估计基频. 其中RAPT方法针对语音的基频跟踪任务设计，对语音有特别的适用性. 近年来，数据驱动的基于深度神经网络的基频提取模型被提出，并显示出良好的性能[26−29]，受到了更多的关注，但会遇到缺乏基频标签的问题.

对于预分离的语音， 我们分别采用传统方法RAPT和基于深度神经网络的方法提取基频. 对于后者，我们也分别尝试了基于分类和基于回归这两种方法，如图2所示. 基于分类的模型有400个输出，对应的基频频率从1到400，基频估计是一个分类任务. 基于回归的模型的输出是一维的，每一时间帧的基频频率通过回归任务计算得到. 本文的回归模型可以看作文献[17]的增强版本. 所提出的两个基频跟踪模型都采用前馈神经网络FNN和长短期记忆神经网络LSTM，然后分类模型后面是log-softmax和argmax操作，而回归模型后面是一个ReLU激活函数层. 对于浊音（V）和轻音（UV）标志预测，考虑到VUV标志在语音分离过程中没有显式的应用，我们采用一个简单的设定阈值方法，即：小于阈值的基频为轻音（UV），大于阈值的基频为浊音（V）.

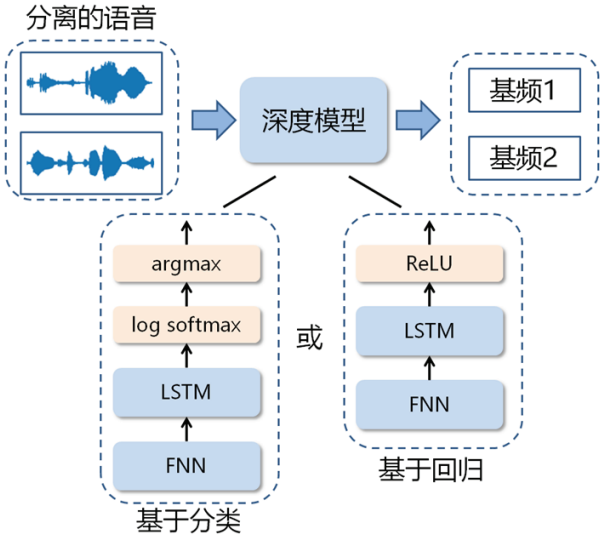

图 2 基于深度学习的基频跟踪模型的结构

另外，对基频跟踪模型的输入进行简单的预处理. 具体来说，在计算某时间帧的基频时，相邻时间帧可能含有有益的信息，我们将输入帧向前后对称地扩充，从而生成一个较长的输入时间帧.

1.4 拼接时的基频排列

在后分离阶段，我们将估计的基频和原始混音进行拼接操作，以生成后分离模块新的输入. 图3展示了三种不同的拼接方法[17]：

（1）Oracle：干净语音的信息是已知的，具有最大SNR的干净语音对应的估计基频首先与混音拼接.

（2）任意顺序：估计基频按任意顺序与混音拼接.

（3）能量：具有最大能量的预分离语音对应的估计基频首先与混音拼接.

因为估计的语音和干净语音较接近，“Oracle”和“能量”方法可以近似看作相同. 事实上，文献[17]中的结果显示，“能量”方法稍微优于“Oracle”方法，且两者皆比“任意”方法好得多. 因此，我们采用“能量”方法作为本文的拼接方法. 又因估计基频的长度与语音波形长度不匹配，需要对基频序列长度进行成比例的缩放，使其与语音长度相同，便于拼接操作.

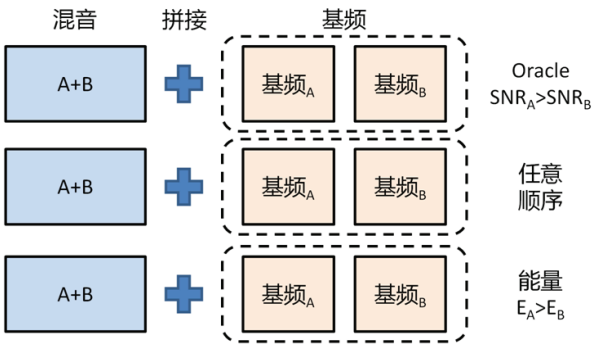

图 3　估计基频与原始混音的拼接方法

2　实验结果

2.1　实验设置

本文采用WSJ0-2mix数据集[1]验证所提出方法的有效性. WSJ0-2mix数据集广泛应用于单通道语音分离模型的验证. 本文采用8 kHz采样的版本，具体做法是在Wall Street Journal （WSJ0）数据集中的训练集si_tr_s中任选2个说话人的语音，按-5～5 dB中的任意SNR混合，以生成30小时的训练集混音和10小时的验证集混音. 从WSJ0开发集si_dt_05和验证集si_et_05的16个说话人中，任选2个说话人的语音，同样按上述方式生成5小时的验证集. 这样测试集的说话人与训练集和验证集不同，即是一个开放数据集.

如前所述，Conv-TasNet在预分离模块和后分离模块中都有应用，而且后者的输入维度是基频与混音拼接的新维度. 在Conv-TasNet中，编码器的滤波器长度设为16. 文献[10]和文献[30]表明，编码器滤波器的长度越短，语音分离的结果越好. 因本文的目的是验证基频信息对时域语音分离系统的提高作用，而不是追求更高的分离结果，因此我们固定此滤波器长度为16.

实验采用平均尺度不变信噪比提升（$SI-SNRi$）[10,31]作为语音分离的评价指标，使用Adam算法作为模型的优化器. 传统训练时的学习率设置为1×10^{-3}，调优时训练的学习率设置为1×10^{-4}.

2.2　基频跟踪结果

本文使用传统方法RAPT和基于深度神经网络的方法进行基频估计任务. 其中后者包括2种方法，即分类方法和回归方法，基频跟踪模型的输入是预分离阶段得到的估计语音. 我们采用4个隐层的FNN，每层512个单元，后面连接一个双向LSTM层，512个隐层单元. 基于分类模型的训练目标为传统交叉熵，基于回归模型的训练目标为MSE. 深度模型的输入维度为3 600，即每一帧的窗长为400个采样点，向两边各扩展4个帧. 相邻帧的移动距离为64个采样点，即8 ms. 另外，对于基于深度神经网络模型还有一个问题需要考虑，即监督训练时的参考标签. 因为对于WSJ0数据集来说没有基频的参考标签，参照文献[17]的方法，我们采用RAPT结果作为参考标签. 对于RAPT方法，输入音频为8 kHz采样，每一帧的移动距离为8 ms；其最小和最大基频频率分别设置为30 Hz和400 Hz. 对于预分离得到的估计语音，相比原干净语音是有失真的，以上设置可以比较传统RAPT和深度学习方法哪一个抗失真的能力更优.

结果如表1所示，其中CLS表示基于分类的基频跟踪模型，REG指基于回归的基频跟踪模型. 我们也展示了基频感知系统[17]中的基频跟踪结果，标记为PA. $VUVError$表示浊音和清音标志预测的错误率. MAE_{global}和$RMSE_{\text{global}}$分别指预分离语音基频与干净语音基频的平均绝对误差和均方根误差，MAE_{F0}和$RMSE_{\text{F0}}$指对于VUV标志预测正确的帧，预分离语音基频与干净语音基频的平均绝对误差和均方根误差.

表 1　不同基频跟踪方法的结果比较（对预分离语音）

指标	RAPT	CLS	REG	PA[17]
$VUVError$/%	**3.3**	4.7	4.2	5.4
MAE_{global}/Hz	**5.9**	8.5	8.8	9.9
$RMSE_{\text{global}}$/Hz	**25.8**	29.9	27.1	27.3
MAE_{F0}/Hz	**1.2**	1.8	4.0	4.8
$RMSE_{\text{F0}}$/Hz	**6.4**	8.8	12.9	13.7

由表1可知，RAPT的基频预测结果普遍优于基于深度学习的方法，也优于文献[17]的方法. 原因在于，基于时域的预分离模块可以达到足够的分离精度，使通过RAPT方法估计的基频相对失真误差较少. 还可以发现本文采用的三种方法，RAPT、CLS和REG都优于PA[17]，而且CLS和REG在不同的指标上各有优势. 在后续实验中，我们将只使用RAPT、CLS和REG的基频估计结果用于后分离模型.

2.3 语音分离结果

2.3.1 使用理想基频的语音分离

首先通过假设目标语音已知，来评估添加理想基频的结果. 用RAPT方法计算出理想的基频，并与原始混音拼接成后分离模块的输入. 我们复现Conv-TasNet并将其作为基线，其分离结果为15.2 dB，比文献[10]低0.1 dB. 原因可能是参数初始化和生成的数据集的随机性.

表 2 使用理想基频的语音分离结果

方法	输入	模型	指标	dB
时频域[17]	混音	DC[1]	$SDRi$	10.4
	混音	uPIT[6]	$SDRi$	9.5
	+理想基频	DC	$SDRi$	12.2
	+理想基频	uPIT	$SDRi$	13.4
时域	混音	Conv-TasNet[10]	$SI-SNRi$	15.2
	+理想基频	Conv-TasNet	$SI-SNRi$	**17.6**

表2首先给出了利用理想基频辅助信息进行时频域分离的结果[17]，说明了基频感知方法在时频域语音分离中的有效性. 然后给出了本文的时域结果，结果表明，利用理想的基频信息，可以获得显著的改善（2.4 dB）. 2.4 dB的改进可以看作理想情况下的值，或者说是要追求的上限. 需要说明的是，文献[17]中使用平均$SDRi$作为指标，而本文使用平均$SI-SNRi$. 平均$SDRi$和平均$SI-SNRi$通常只相差一个常量[4,7,9,10,30]，因此考查性能提升程度时，是可以互相比较的.

我们注意到时频域的基频感知系统[17]采用DC或uPIT作为分离模块. 由表2可知，加入基频后，基于uPIT的分离模型可以得到更大的性能提升，即3.9 dB. 这与表2中时域的分离方法相比(2.4 dB)，可知时频域的提升程度更大. 原因可能是时频域方法缺乏相位信息，加入基频可带来更多的信息补偿；而时域的输入已经隐式地包含基频信息，因此显式加入基频会带来相对有限的增强.

2.3.2 使用非理想基频的语音分离

与训练阶段不同的是，在推理阶段，我们没有理想基频提取的干净语音，我们所掌握的只是预分离的估计语音. 良好的预分离结果会得到精确的基频跟踪结果，反过来足够接近理想情况的基频跟踪结果也会使后分离的语音质量得到提升. 这就导致了“鸡和蛋”的问题. 我们使用从预分离语音中提取的基频来近似理想基频，表示为非理想基频. 将非理想基频与原始混合语音相结合作为后分离网络的输入. 为了解决“鸡和蛋”的问题，我们利用理想基频的训练成果辅助非理想基频网络的训练. 具体来说，对于后分离网络，我们不进行随机的参数初始化，而是使用理想基频训练的网络参数，在此基础上，使用非理想基频进行调优.

表3显示了使用非理想基频的分离结果. 我们将文献[17]中时频域的基频感知分离结果作为对比. 其采用基于DC的模型作为预分离模块，将聚类过程从原来的K-means替换为前馈网络. 对于后分离，分别验证了基于DC和uPIT的模型. 可以发现预分离和后分离均采用DC方法，结果与只使用DC进行预分离结果相同；而将后分离改为uPIT方法后，结果有较大改善.

对于时域的语音分离，我们复现Conv-TasNet作为预分离网络，其分离结果为15.2 dB. 对于后分离，我们评估了多种基频提取方法（RAPT，分类CLS和回归REG），对应的结果分别为15.3 dB、15.2 dB和15.2 dB，说明相比无基频辅助信息的Conv-TasNet没有明显提升. 我们也将CLS和REG的基频提取模型与后分离网络进行联合训练，表示为CLS+POST+Joint和REG+POST+Joint，对应的结果分别为14.8 dB和15.3 dB. 由此可知联合训练中基于分类的基频提取方法失败了. 最后我们展示了使用理想基频预训练的后分离网络进行调优的结果，表示为RAPT+POST+Tune、CLS+POST+Tune和REG+POST+Tune，对应结果分别为15.7 dB、15.6 dB和15.6 dB，均优于基线Conv-TasNet. 而且，使用RAPT的调优方法比其他两个深度模型高0.1 dB，这与基频跟踪结果中，RAPT优于其他两个方法的情况相对应.

表 3　使用非理想基频的语音分离结果

方法	预分离	后分离	指标	dB
时频域[17]	DC	None	$SDRi$	10.8
		DC	$SDRi$	10.8
		uPIT	$SDRi$	12
时域	Conv-TasNet[10]	None	$SI-SNRi$	15.2
		RAPT+POST	$SI-SNRi$	15.3
		CLS+POST	$SI-SNRi$	15.2
		REG+POST	$SI-SNRi$	15.2
		CLS+POST+Joint	$SI-SNRi$	14.8
		REG+POST+Joint	$SI-SNRi$	15.3
		RAPT+POST+Tune	$SI-SNRi$	**15.7**
		CLS+POST+Tune	$SI-SNRi$	15.6
		REG+POST+Tune	$SI-SNRi$	15.6

通过对时频域和时域分离结果的比较，我们发现时域方法相对基线的提升小于时频域的提升（0.5 dB对1.2 dB）. 原因在于：首先，基频在时域上提供的信息增量有限；其次，基频感知系统[17]使用基于DC的模型作为预分离模块，uPIT作为后分离模块（即DC+uPIT），这利用了它们之间的互补性. 然而，本文中预分离模块和后分离模块本质上是相同的，均为Conv-TasNet.

3　结论和展望

本文研究了在时域语音分离方法中，利用辅助基频信息来改善语音分离的性能. 实验结果表明，辅助基频信息对时域语音分离也有一定的帮助. 在训练过程中，我们发现传统的训练方法并不能明显提高训练效果，而使用理想基频进行预训练，再对预训练的模型进行微调，可以获得最佳性能. 未来的工作包括扩展到其他先验信息，如说话人表示等. 我们还发现，深度模型由于缺乏参考基频而落后于传统的RAPT算法，因此可以采用一种更精确的基频跟踪算法，或者创建带有参考基频的数据集，来进一步提高所提出方法的效果.

参考文献：

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

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

[3] CHEN Z, LUO Y, MESGARANI N. Deep attractor network for single-microphone speaker separation[C]. New Orleans: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). IEEE, 2017: 246-250.

[4] LUO Y, CHEN Z, MESGARANI N. Speaker-independent speech separation with deep attractor network[J]. IEEE/ACM Transactions on Audio Speech & Language Processing, 2018, 26(4): 787-796.

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

[6] KOLBK 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 & Language Processing, 2017, 25(10): 1901-1913.

[7] WANG Z Q, TAN K, WANG D L. Deep learning based phase reconstruction for speaker separation: a trigonometric perspective[C]. Brighton: 2019 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). IEEE, 2019: 71-75.

[8] LE ROUX J, WICHERN G, WATANABE S, et al. Phasebook and friends: leveraging discrete representations for source separation[J]. IEEE Journal of Selected Topics in Signal Processing, 2019, 13(2): 370-382.

[9] LUO Y, MESGARANI N. Tasnet: time-domain audio separation network for real-time, single-channel speech separation[C]. Calgary: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). IEEE, 2018: 696-700.

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

[11] SHI Z, LIN H, LIU L, et al. End-to-end monaural speech separation with multi-scale dynamic weighted gated dilated convolutional pyramid network[C]. Graz: Interspeech 2019, 2019: 4614-4618.

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

[13] YOSHIOKA T, ABRAMOVSKI I, AKSOYLAR C, et al. Advances in online audio-visual meeting transcription[C]. Singapore: 2019 IEEE Automatic Speech Recognition and Understanding Workshop(ASRU). IEEE, 2019: 276-283.

[14] CHEN Z, XIAO X, YOSHIOKA T, et al. Multi-channel overlapped speech recognition with location guided speech extraction network[C]. Athens: 2018 IEEE Spoken Language Technology Workshop(SLT). IEEE, 2018: 558-565.

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

[16] YANG G P, TUAN C I, LEE H Y, et al. Improved speech separation with time-and-frequency cross-domain joint embedding and clustering[C]. Graz: Interspeech 2019, 2019: 1363-1367.

[17] WANG K, SOONG F, XIE L. A pitch-aware approach to single-channel speech separation[C]. Brighton: 2019 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). IEEE, 2019: 296-300.

[18] CHEN Z, YOSHIOKA T, LU L, et al. Continuous speech separation: dataset and analysis[C]. Online: 2020 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). IEEE, 2020: 7284-7288.

[19] TALKIN D, KLEIJN W B. A robust algorithm for pitch tracking(RAPT)[J]. Speech Coding and Synthesis, 1995, 495: 518.

[20] YOSINSKI J, CLUNE J, BENGIO Y, et al. How transferable are features in deep neural networks?[C]. Montreal: Advances in Neural Information Processing Systems, 2014: 3320-3328.

[21] LEA C, VIDAL R, REITER A, et al. Temporal convolutional networks: a unified approach to action segmentation[C]. Amsterdam: European Conference on Computer Vision, 2016: 47-54.

[22] LI L, LIN G L, MA S B. Research of single image super-resolution reconstruction with sawtooth dilated residual convolution[J]. Journal of Xinjiang University(Natural Science Edition in Chinese and English), 2021, 38(2): 174-190.

[23] GERHARD D. Pitch extraction and fundamental frequency: history and current techniques[M]. Regina: Department of Computer Science, University of Regina, 2003.

[24] BOERSMA P. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound[J]. Proceedings of the Institute of Phonetic Sciences, 1993, 17(1193): 97-110.

[25] DE CHEVEIGNÉ A, KAWAHARA H. YIN, a fundamental frequency estimator for speech and music[J]. The Journal of the Acoustical Society of America, 2002, 111(4): 1917-1930.

[26] HAN K, WANG D L. Neural network based pitch tracking in very noisy speech[J]. IEEE/ACM Transactions on Audio Speech & Language Processing, 2014, 22(12): 2158-2168.

[27] KIM J W, SALAMON J, LI P, et al. Crepe: a convolutional representation for pitch estimation[C]. Calgary: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). IEEE, 2018: 161-165.

[28] XU S, SHIMODAIRA H. Direct f0 estimation with neural-network-based regression[C]. Graz: Interspeech 2019, 2019: 1995-1999.

[29] GFELLER B, FRANK C, ROBLEK D, et al. SPICE: self-supervised pitch estimation[J]. IEEE/ACM Transactions on Audio Speech & Language Processing, 2020, 28: 1118-1128.

[30] HEITKAEMPER J, JAKOBEIT D, BOEDDEKER C, et al. Demystifying TasNet: a dissecting approach[C]. Online: 2020 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). IEEE, 2020: 6359-6363.

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

责任编辑：张自强