计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(10):1−7 [doi: 10.15888/j.cnki.csa.007093]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于深层声学特征的端到端语音分离①

李娟娟[1], 王　丹[2], 李子晋[3]

[1](复旦大学 计算机科学技术学院, 上海 201203)
[2](盲信号处理国家级重点实验室, 上海 200434)
[3](中国音乐学院 音乐科技系, 北京 100101)
通讯作者: 李娟娟, E-mail: liergou0828@163.com

摘　要: 提出基于深层声学特征的端到端单声道语音分离算法, 传统声学特征提取方法需要经过傅里叶变换、离散余弦变换等操作, 会造成语音能量损失以及长时间延迟. 为了改善这些问题, 提出了以语音信号的原始波形作为深度神经网络的输入, 通过网络模型来学习语音信号的更深层次的声学特征, 实现端到端的语音分离. 客观评价实验说明, 本文提出的分离算法不仅有效地提升了语音分离的性能, 也减少了语音分离算法的时间延迟.

关键词: 语音分离; 声学特征; 深度神经网络; 语音原始波形; 端到端模型

引用格式: 李娟娟,王丹,李子晋.基于深层声学特征的端到端语音分离.计算机系统应用,2019,28(10):1–7. http://www.c-s-a.org.cn/1003-3254/7093.html

End-to-End Speech Separation Based on Deep Acoustic Feature

LI Juan-Juan[1], WANG Dan[2], LI Zi-Jin[3]

[1](School of Computer Science, FudanUniversity, Shanghai 201203, China)
[2](National Key Laboratory of Blind Signal Processing, Shanghai 200434, China)
[3](Department of Music Technology, China Conservatory of Music, Beijing 100101, China)

Abstract: An end-to-end single channel speech separation algorithm based on deep acoustic feature is proposed. The traditional acoustic feature extraction methods require the Fourier transform, discrete cosine transform and other operations. This will cause speech energy loss and long latency. In order to improve these problems, the original waveform of the speech signal is used as an input to a deep neural network, deeper acoustic features of the speech signal are learned through a network model. Objective evaluation shows that the proposed algorithm not only improves the performance of speech separation effectively, but also reduces the time delay of speech separation algorithm.

Key words: speech separation; acoustic feature; deep neural network; speech original waveform; end-to-end model

语音作为一项最为便捷的交流工具, 实现了人类社会高效快速的信息交换, 成为人类文明的一个重要助力. 然而在现实环境中, 感兴趣的语音信号通常会被其他声源干扰, 严重损害了语音的可懂度, 降低了语音交互的性能. 为了解决以上问题, 语音分离是最为关键的技术之一.

语音分离是指从多个说话人的混合语音中分离得到想要的语音数据, 源于著名的“鸡尾酒会问题”[1], 主要是研究如何能够从混合的语音信号中同时得到目标和干扰语音信号, 它在语音识别、残疾人助听领域具有广泛的应用. 本文主要探究两个说话人混合的情况. 图 1 是语音分离技术的示意图, 图中左边的两张语谱图分别是两个说话人的语音的语谱图, 经过混合后得到中间的混合语音的语谱图, 而经过语音分离以后得

① 基金项目: 国家自然科学基金 (61671156); 北京市社会科学基金 (17YTC028)
Foundation item: National Natural Science Foundation of China (61671156); Social Science Foundation of Beijing Municipality (17YTC028)
收稿时间: 2019-03-12; 修改时间: 2019-04-04; 采用时间: 2019-04-16; csa 在线出版时间: 2019-10-15

到的是右边分离出的语音的语谱图. 从图 1 可以看出, 由于不同的说话人的语音的发音特性有差异和说话内容、语速等不同, 以及语音信号这种时变信号本身具有一定的短时平稳特性, 从而使得语音分离具有可行性[2].

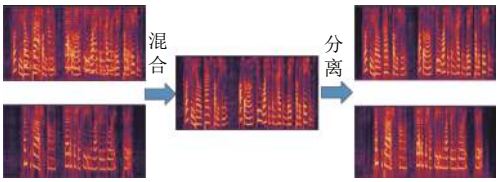

图 1 语音分离技术示意图

语音分离作为一个重要的研究领域, 几十年来, 受到国内外研究者的广泛关注和重视. 近年来, 监督性语音分离技术取得了重要的研究进展, 特别是深度学习的应用, 极大地促进了语音分离的发展. 在基于深度神经网络的语音分离算法中, 特征提取是至关重要的步骤. 傅里叶变换域特征是最常用的语音分离特征, Xu[3]、Huang[4]、Weninger[5]等使用傅里叶幅度谱或者傅里叶对数幅度谱作为语音分离的输入特征. Wang 等在文献[6]中总结了 Gammatone 滤波变换域特征, 并且利用 Group Lasso 的特征选择方法得到 AMS+RASTA-PLP+MFCC 的特征组合. Chen 等在文献[7]中提取的多分辨率特征 MRCG 具有明显的优势, 逐渐取代了组合特征成为语音分离中最常用的特征之一. 然而以上这些传统声学特征的提取需要经一系列复杂的操作, 会造成语音能量损失以及长时间延迟.

近年来, 端到端的方法已经用于语音识别、语音合成和语音增强等语音任务中, 并在这些任务中取得了较优的效果. Luo 等人在文献[8]中首次提出了基于非负矩阵分解思想的端到端语音分离, 并取得了较优的效果. 为了进一步说明端到端的方法在语音分离这一方向的可行性, 本文提出以语音信号的原始波形作为深度神经网络的输入, 通过网络模型来学习语音信号的更深层次的深层声学特征, 实现端到端的语音分离.

1 基于传统声学特征的语音分离

语音分离旨在分离混合语音信号中的信号, 这个过程能够很自然地表达成一个监督性学习问题[3–5]. 一个典型的监督性语音分离系统通常通过监督性学习算法, 例如深度神经网络, 学习一个从混合语音的传统声学特征到分离目标的映射函数[9]. 算法 1 为基于传统声学特征的语音分离算法.

算法 1. 基于传统声学特征的语音分离算法

1) 时频分解, 通过信号处理的方法将输入的时域信号分解成二维的时频信号表示;
2) 特征提取, 提取帧级别或者时频单元级别的听觉特征 (短时傅里叶变换谱或者短时傅里叶功率谱等);
3) 模型训练, 利用大量的输入输出训练对通过机器学习算法学习一个从混合语音特征到分离目标 (理想二值掩蔽或者理想比例掩蔽等) 的映射函数;
4) 波形合成, 利用估计的分离目标以及混合信号, 通过逆变换 (逆傅里叶变换或者逆听觉滤波) 获得目标语音的波形信号.

在提取传统声学特征时, 先要进行时频分解, 一般都是将时域信号通过短时离散傅里叶变换 (Short-time Fourier Transform, STFT)、离散余弦变换 (Discrete Cosine Transform, DCT) 或者通过一些听觉滤波器组 (如 Gammatone 滤波器组) 得到二维的时频域表示. 在这个过程中产生了两个问题. 一是忽略了在提取特征的过程中造成语音的高频部分以及相位信息的损失, 以及在变换过程中可能会引入虚假的信息, 从而对语音分离的性能造成影响. 二是由于变换域中的有效语音分离对高频分辨率的需求, 导致相对较大的时间窗口长度, 对于语音通常超过 32 毫秒[3, 10–12], 音乐分离超过 90 毫秒[13]. 因为系统的最小延迟受 STFT 时间窗的长度限制, 所以当需要非常短的延迟时, 这限制了此类系统的使用, 例如电信系统或可听设备这类实时性系统. 克服这些问题的一种自然方法是直接建模时域中的信号. 有研究结果表明, 语音原始波形相比基于傅里叶变换的梅尔倒谱系数等特征, 在某些研究领域具有更好的语音性能[14]. 所以本文选择以语音信号的原始波形作为深度神经网络的输入, 通过网络模型来学习语音信号深层次的深层声学特征 (Deep Acoustic Feature, *DAF*), 实现端到端的语音分离.

2 基于深层声学特征的端到端语音分离

图 2 是基于深层声学特征的端到端语音分离算法的整体流程, 主要分为 4 个部分: (1) 信号预处理, 对混合信号的原始波形进行分段及规整. (2) 深层声学特征提取, 提取时域信号的 DAF 作为分离模型的输入. (3) 分离模型, 训练分离模型得到各个信号的特征掩蔽值. (4) 信号重建, 利用得到的信号的特征掩蔽值及混合信号的 DAF, 通过信号重建得到各个分离信号的时域波形.

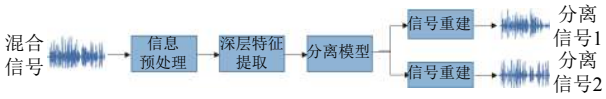

图 2 算法整体流程

2.1 信号预处理

数据预处理在许多机器学习算法中起着很重要的作用, 如果输入的特征向量在整个训练集上均值接近零, 那么模型的收敛速度会很快. 语音信号的预处理模块包括两部分, 分段、规整.

首先将混合信号分成 K 段, 每段长度为 L, 再对每段使用单元 L^2 规整, X_k 是分段后的信号, 规整方式如下:

$$X_{k-norm}=\frac{X_k}{X_k}=\frac{X_k}{\sqrt{X_k^{\mathrm{T}}X_k}}=\frac{X_k}{\sqrt{\sum_{i=1}^{L}X_{ki}^2}} \tag{1}$$

单元 L^2 规整即可以削弱时不变信道的影响, 还能减少加性噪声的影响, 同时时域信号被缩放到相似的动态范围内, 使得后续模型的学习过程也能取得较好的效果.

2.2 深层特征提取

在基于深度神经网络的语音分离算法中, 语音分离任务能够被表达成一个学习问题, 对于深度学习问题, 特征提取是至关重要的步骤. 提取好的特征能够极大地提高语音分离的性能[15].

针对传统声学特征提取方法需要经过傅里叶变换、离散余弦变换等操作, 提取复杂特征作为输入, 会造成能量损失的问题, 本文选择以语音信号的原始波形作为深度神经网络的输入, 通过网络模型来学习语音信号深层次的声学特征, *DAF* 提取过程如图 3 所示.

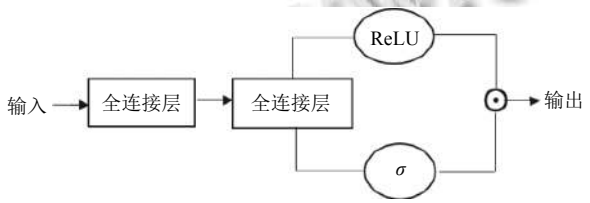

图 3 *DAF* 提取过程

在 *DAF* 的提取过程中, 参考语言建模[16]中的门限卷积方法, 在第二层全连接层后引入门限机制如下:

$$h=\mathrm{ReLU}(w_1X+b_1)\odot\sigma(w_2X+b_2) \tag{2}$$

其中, ReLU 为线性整流函数, σ 为 Sigmoid 激活函数, $\odot$表示逐元素乘积操作. 引入门限机制可以控制模型中的信息流动, 帮助模型的神经元之间有更加复杂的联系. 相比于语音建模中的门限卷积, 本文中使用全连接代替卷积操作, 虽然使用卷积操作能减少训练参数从而缩短训练时间, 但是使用全连接操作能减少语音损失的能量, 提取的特征也能更多地挖掘深层次的声学特征, 提升语音分离的性能.

2.3 分离模型

双向长短时记忆网络 (Bi-derectional Long Short Term Memory, BiLSTM) 结构能够有效抓住音频数据中的长时依赖, 对语音建模非常有效[17,18]. 本文中, 分离网络由 4 层深度 BiLSTM 后面接着一个全连接层构成, 在第二层隐藏层的输出与第四层隐藏层的输出之间增加了跳跃连接[19], 改善了多层网络反向传播的梯度消散问题, 提升网络性能.

网络的输入是混合信号的 *DAF*, 网络的输出是各个信号的掩蔽值. 已有研究证明在语音分离任务中把掩蔽值 (mask) 作为分离目标能显著地提高语音分离的可懂度和感知质量. 其中, 最常使用的分离目标之一为理想比例掩蔽 (Ideal Ratio Mask, IRM)[20]. 基于 IRM 的定义, 本文中使用的信号的掩蔽值, 特征比例掩蔽 (Feature Ratio Mask, *FRM*) 的定义如下:

$$FRM=\frac{DAF_i}{DAF_{\mathrm{mix}}},i=1,2 \tag{3}$$

使用掩蔽值作为分离模型的输出比使用特征 *DAF* 的效果更好. 全连接层的激活函数为 Softmax 函数. 为了加速训练进程及维持训练过程中的稳定性, 对分离网络的输入即混合信号的 *DAF* 要进行层级归一化.

2.4 信号重建

将混合信号的 *DAF* 逐元素乘以各个信号的 *FRM*, 经过一层全连接层后. 得到规整的目标信号的时域波形, 最后通过逆规整和整合, 重建各个信号的时域信号.

2.5 损失函数

网络模型的最终输出是估计的干净信号的时域波形, 由于模型效果的重要评价指标之一是尺度不变信噪比 (Scale-invariant Source-to-noise Ratio, *SI-SNR*)[8], 所以在这里不使用估计语音的时域波形和干净的时域波形的均方误差, 而是基于 SI-SNR 来设计损失函数. *SI-SNR* 的定义如下:

$$S_{\mathrm{target}}=\frac{\langle\hat{S},S\rangle S}{\|S\|^2} \tag{4}$$

$$e_{\text{noise}} = \hat{S} - S_{\text{target}} \tag{5}$$

$$SI-SNR = 10\log_{10}\frac{\|S_{\text{target}}\|^2}{\|e_{\text{noise}}\|^2} \tag{6}$$

其中, $\hat{S}$是估计的目标信号, S 是干净的目标信号, 注意 $\hat{S}$和 S 都要做均值为 0 的归一化来确保尺度不变性. 由于 *SI-SNR* 的值越大, 语音质量越好, 而在训练的过程中使用梯度下降来训练模型, 所以实际的损失函数 *loss* 的定义取 *SI-SNR* 的倒数.

3 实验结果和分析

3.1 实验配置

华尔街日报语料库 (Wall Street Journal, WSJ0) 是语音分离任务常用的数据集[11–13], 每条语音大约在 5 s 左右. 混合语音由随机选取 WSJ0 训练集 si_tr_s 中的任意两个说话人, 以随机选取的 0–5 dB 信噪比混合而成, 最终形成 30 个小时的训练集和 10 小时的验证集. 测试集使用 WSJ0 的 si_dt_05 和 si_et_05 的未知说话人以相同的混合方式产生, 最终生成 5 小时的测试集.

实验中所使用的语音波形文件具有 8 kHz 的采样频率. 分段时的长度 L=40 (5 ms), 每段之间有 50% 的重叠, 提取的 *DAF* 长度为 500. 深度 BiLSTM 采用 4 层隐藏层, 每层隐藏层的结点是 500, 在第二层隐藏层的输出与第四层隐藏层的输出之间有跳跃连接, 最后一层全连接层的结点数为 1000, 使用 Softmax 激活函数. 在训练过程中, 使用随机初始化的网络, 采用的最小批训练方法中每个最小批的训练集包含 128 个样本. 初始的学习率设置为 $1e^{-3}$, 当验证集上的损失在连续 3 个迭代次数 (epoch) 没有降低时, 就将学习率设置为当前学习率的一半. 当验证集上的损失在连续 10 个 epoch 都没有降低时停止训练. 选用 Adam 优化函数, Adam 优化器的超参数具有很好的解释性, 通常无需调整或仅需很少的微调, 适用于大规模数据及参数的场景.

3.2 评价指标

本实验中采用的评价指标为 BSS-EVAL 指标. BSS-EVAL 工具箱通常用来评估模型的分离性能, 它是由 Vincent 等人在 2006 年提出的语音分离指标[21], 并开源的语音分离评估工具箱, 广泛被研究者用于语音分离评价中. 根据 BSS-EVAL 指标, 语音分离评估使用 3 个定量值分别是, 信噪干扰比 (Source to Interference Ratio, SIR), 信噪伪影比 (Source to Artifact Ratio, SAR) 和信噪失真比 (Source to Distortion Ratio, SDR). 3 个值均是越高越好. 其中, SDR 计算分离声音中存在多少总失真, SDR 值越高表示语音分离系统整体上的失真越小, 语音分离系统性能越好. SIR 直接比较非目标声源噪音与目标声音的分离程度. SAR 是指在语音分离过程中引入的人工误差, SAR 值越高, 表明引入误差对语音分离系统影响越小.

3.3 实验结果和分析

(1) 基于 *DAF* 的语音分离算法的效果

表 1 为所提的基于 *DAF* 的语音分离算法在测试集 (3000 条语音) 上的分离语音的平均 SDR、SIR 及 SAR 值, 分别为 11.60、22.58 和 12.38. 从客观评价指标来看, 本文所提出的语音分离算法在测试集上的有效性.

表 1 测试集平均 SDR、SIR、SAR 值

参数	SDR	SIR	SAR
数值	11.60	22.58	12.38

图 4 是本文所提语音分离算法在测试集 (3000 条语音) 上的 SDR 值 (每条混合语音分离出来的两条语音的 SDR 值取平均) 的分布. 其中分离后语音的 SDR 值大于 10 的有 75%, 分离效果很好, 语音质量清晰可懂. SDR 值在 5 到 10 范围内的有 8%, 分离效果较好, 语音不够清晰, 但是可懂. SDR 值在 0 到 5 范围内的有 10%, 分离效果一般, 不明显. SDR 值<0 的有 7%, 分离效果差, 分离前与分离后没有差别. 经观察分析, 这 7% 的混合语音, 混合的两个不同的说话人基本是同性别并且发音特性较为相似, 导致分离算法在这部分数据上处理效果不好.

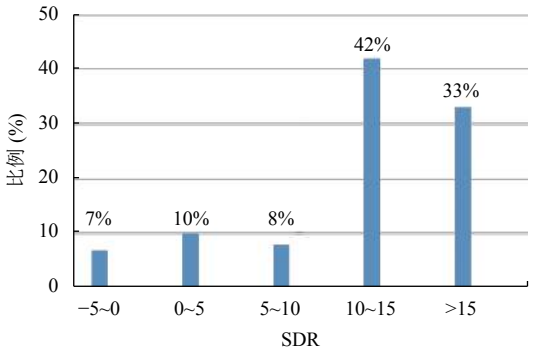

图 4 测试集上 SDR 值分布

图 5 分别是混合语音、分离语音 1 和 2 的 *DAF* 的可视图. 从图中可以看出, *DAF* 中有一条条的类似于

频谱图中的“声纹”的东西, 并且不同的说话人对应的“声纹”的位置不同, 说明深度网络确实可以从语音的时域信号中学习到不同说话人的声音特性并且能做出相应的区分.

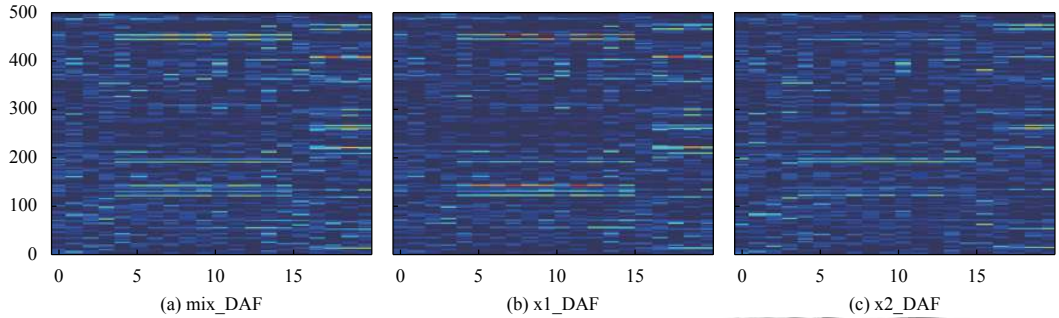

图 5 语音 *DAF* 的可视图

图 6 是所提语音分离算法的一个效果示例, 每张小图的上方是语音信号的原始波形, 下方是其对应的语谱图. 图中左边的两张小图分别是测试集中的两个说话人的语音, 以 0.27 dB 的信噪比经过混合后得到中间的混合语音, 右边是分离出的两个说话人语音, 分离后的 SDR 值分别为 14.20 和 12.39. 无论是从客观评价指标 SDR, 还是从主观地比较分离前后的语音原始波形和语谱图, 均能看出所提出语音分离算法的有效性.

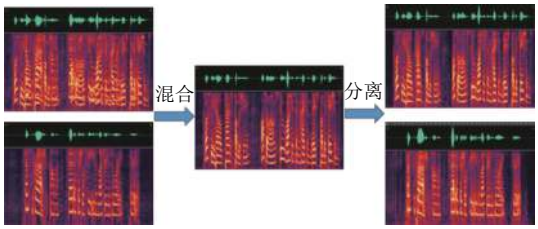

图 6 一个语音分离效果示例

(2) 不同声学特征的效果对比

在这一部分实验中, 为了探究本文使用的深层声学特征的有效性, 与语音分离任务中最常用的传统声学特征, 经过 STFT 变换的 257 维对数功率谱特征 (Log Power Spectrum, LPS) 做对比. 同时为了验证 *DAF* 中使用的门限机制的有效性, 与单独使用 ReLU、Sigmoid 激活函数做对比, 其他实验配置与 3.1 小节的配置相同.

表 2 为使用不同声学特征的测试集上的平均 SDR、SIR 和 SAR 值. 使用门限机制 DAF、单独使用 ReLU 评价指标比使用 LPS 特征高, 说明使用网络去学习语音信号深层特征比使用传统基于 STFT 的特征有效. 而单独使用 Sigmoid 的深层特征比使用 LPS 评价指标低, 说明了提取深层特征中选取恰当激活函数的重要性, 选取不当会导致没有传统特征效果好. 另外, 使用 DAF 特征比使用单独 ReLU 和单独使用 Sigmoid 的评价指标高, 说明本文所提出的深度声学特征中使用门限机制的有效性.

表 2 不同声学特征的测试集平均 SDR、SIR、SAR 值

声学特征	SDR	SIR	SAR
DAF	11.60	22.58	12.38
ReLU	11.29	21.07	12.15
Sigmoid	5.08	11.47	9.40
LPS	7.05	16.20	10.74

(3) 不同分离网络的效果对比

在这一部分实验中, 为了验证分离网络中使用的 BiLSTM 的双向的有效性, 使用普通 LSTM(非双向) 与之做对比. 深度 LSTM 网络有 4 层隐藏层, 每层隐藏层的结点为 1000. 其他实验配置与 3.1 小节的配置相同.

表 3 为使用不同分离网络 (BiLSTM vs 普通 LSTM) 在测试集上的平均 SDR、SIR 和 SAR 值. 使用 BiLSTM 比使用普通 LSTM 的分离网络的 SDR 值高了 5 左右. 因为普通 LSTM 在时序上处理序列没有考虑未来的上下文信息, 忽略了未来时刻的影响. 而使用 BiLSTM 看到未来信息对当前时刻的影响, 更适用于本算法中的分离网络.

表 3 不同分离网络的测试集平均 SDR、SIR、SAR 值

分离网络	SDR	SIR	SAR
BiLSTM	11.60	22.58	12.38
普通 LSTM	6.55	16.63	7.58

(4) 不同损失函数的效果对比

在本实验中采用了 1/SI-SNR 的损失函数, 其他最常用的损失函数是直接基于时域信号的最小均方差

(Minimum Mean Squared Error, MMSE) 损失函数, 直接优化估计语音与干净语音的时域信号差. 该损失函数定义如下:

$$E=\frac{1}{N}\sum_{n=1}^{N}\left(\hat{S}_n(x_n,w,b)-S_n\right)^2+\lambda\|w\|^2 \tag{7}$$

其中, $\hat{S}$是估计的目标信号, S 是干净的目标信号, λ是 L_2 范数正则化系数. 在这一部分实验中, 为了验证本文中使用基于 *SI-SNR* 损失函数的的有效性, 使用 MMSE 损失函数与之做对比. 其他实验配置与 3.1 小节的配置相同.

表 4 为使用不同损失函数 (1/*SI-SNR* vs MMSE) 在测试集上的平均 SDR、SIR 和 SAR 值. 使用基于 *SI-SNR* 的损失函数比使用 MMSE 的 SDR 值高了 4 左右. 因为 *SI-SNR* 本身就是评价语音分离效果的重要指标, *SI-SNR* 越高则语音质量越高, 相对于直接优化语音原始波形的损失, 使用基于 *SI-SNR* 的损失函数更适用于本算法的模型优化.

表 4 不同损失函数的测试集平均 SDR、SIR、SAR 值

损失函数	SDR	SIR	SAR
1/*SI-SNR*	11.60	22.58	12.38
MMSE	7.44	17.75	8.40

(5) 不同语音分离算法的效果对比

在这一部分实验中, 为了探究所提算法在语音分离任务上的性能优劣, 使用目前四种具有代表性的语音分离算法与之做对比, 分别为深度聚类 (Deep Clustering, DC) 语音分离算法[10], 置换不变性 (Permutation Invariant Training, PIT) 语音分离算法[11]、时域语音分离算法 Tasnet[8]和在音乐分离任务上表现很好的多任务 Chimera 模型[13]. 这四种方法中有基于时域的方法, 也有基于频域的方法. 在测试集上的测试结果如表 5 所示. 这可以发现, 本文所提出的算法的在语音分离任务上的有效性.

表 5 不同语音分离算法的测试集平均 SDR 值

参数	本文算法	DC	Chimera	PIT	Tasnet
SDR	11.6	10.6	10.8	10.0	11.1

(6) 时间延迟

在这部分实验中, 为了探究基于传统声学特征的分离算法和本文所提算法的时间延迟, 选用最常用的 STFT 特征与之做对比, 实验结果如表 6 所示. 算法延迟 T 等于建模所需的时域波形时间 T_1、特征提取所需的时间 T_2、分离网络的时间 T_3 和波形重建的时间 T_4 的和. 实验中保证分离网络的结构相同, 即 T_3 相同, T_4 与 T_2 成正比. 所以实际的时间延迟由 T_1 和 T_2 决定. 实验所使用的 GPU 为 GTX1070. 在 8 kHz 的采样率下, 提取 STFT 特征时, 每帧的采样点数最少为 256, 对应时域波形为 32 ms. 本文对 5 ms 的时域波形进行建模, 通过模型对 5 ms 提取 DAF 特征的时间为 0.002 ms. 5.002 ms 远小于 32 ms, 本文所提算法能极大地降低时间延迟.

表 6 时间延迟实验 (单位: ms)

特征	T_1	T_2
DAF	5	0.002
STFT	32	-

4 总结与展望

本文提出了基于深层声学特征的语音分离算法, 该算法通过网络模型来学习语音信号的更深层次的深层声学特征, 实现端到端的语音分离. 在实验部分, 选取了 SDR、SIR 和 SAR 作为客观评价指标在 WSJ0 数据集上进行了一系列对比实验. 结果表明, 本文提出的深层声学特征在语音分离任务中的有效性, 提出的算法提升了语音分离的性能. 并且本文对 5 ms 的时域波形进行建模, 极大地降低了时间延迟. 但是测试集中仍然有 7% 的数据分离效果不好, 对于这部分发音特性较为相似的语音分离任务, 是今后的研究重点.

参考文献

1 Cherry EC. Some experiments on the recognition of speech, with one and with two ears. The Journal of the Acoustical Society of America, 1953, 25(5): 975–979. [doi: 10.1121/1.1907229]

2 王燕南. 基于深度学习的说话人无关单通道语音分离[博士学位论文]. 合肥: 中国科学技术大学, 2017.

3 Xu Y, Du J, Dai LR, *et al*. An experimental study on speech enhancement based on deep neural networks. IEEE Signal Processing Letters, 2014, 21(1): 65–68. [doi: 10.1109/LSP.2013.2291240]

4 Huang PS, Kim M, Hasegawa-Johnson M, *et al*. Deep learning for monaural speech separation. Proceedings of 2014 IEEE International Conference on Acoustics, Speech and Signal Processing. Florence, Italy. 2014. 1562–1566.

5 Weninger F, Hershey JR, Le Roux J, *et al*. Discriminatively trained recurrent neural networks for single-channel speech

separation. Proceedings of 2014 IEEE Global Conference on Signal and Information Processing. Atlanta, GA, USA. 2014. 577–581.

6 Wang YX, Han K, Wang DL. Exploring monaural features for classification-based speech segregation. IEEE Transactions on Audio, Speech, and Language Processing, 2013, 21(2): 270–279. [doi: 10.1109/TASL.2012.2221459]

7 Chen JT, Wang YX, Wang DL. A feature study for classification-based speech separation at low signal-to-noise ratios. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014, 22(12): 1993–2002. [doi: 10.1109/TASLP.2014.2359159]

8 Luo Y, Mesgarani N. TaSNet: Time-domain audio separation network for real-time, single-channel speech separation. Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing. Calgary, AB, Canada. 2018. 696–700.

9 Wang YX, Narayanan A, Wang DL. On training targets for supervised speech separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014, 22(12): 1849–1858. [doi: 10.1109/TASLP.2014.2352935]

10 Isik Y, Le Roux J, Chen Z, *et al*. Single-channel multi-speaker separation using deep clustering. Proceedings of the Annual Conference of International Speech Communication Association. San Francisco, CA, USA. 2016. 545–549.

11 Kolbæk M, Yu D, Tan ZH, *et al*. Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2017, 25(10): 1901–1913. [doi: 10.1109/TASLP.2017.2726762]

12 Chen Z, Luo Y, Mesgarani N. Deep attractor network for single-microphone speaker separation. Proceedings of 2017 IEEE International Conference on Acoustics, Speech and Signal Processing. New Orleans, LA, USA. 2017. 246–250.

13 Luo Y, Chen Z, Hershey JR, *et al*. Deep clustering and conventional networks for music separation: Stronger together. Proceedings of 2017 IEEE International Conference on Acoustics, Speech and Signal Processing. New Orleans, LA, USA. 2017. 61–65.

14 Sainath TN, Weiss RJ, Senior A, *et al*. Learning the speech front-end with raw waveform CLDNNs. Proceedings of the 16th Annual Conference of the International Speech Communication Association. Dresden, Germany. 2015. 1–5.

15 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. 自动化学报, 2016, 42(6): 819–833.

16 Dauphin YN, Fan A, Auli M, *et al*. Language modeling with gated convolutional networks. Proceedings of the International Conference on Machine Learning. Sydney, Australia. 2017. 933–941.

17 Sak H, Senior A, Beaufays F. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. Proceedings of the 15th Annual Conference of the International Speech Communication Association. Singapore. 2014. 338–342.

18 Chen JT, Wang DL. Long short-term memory for speaker generalization in supervised speech separation. Proceedings of 17th Annual Conference on International Speech Communication Association. San Francisco, CA, USA. 2016. 3314–3318.

19 He KM, Zhang XY, Ren SQ, *et al*. Identity mappings in deep residual networks. Proceedings of the 14th European Conference on Computer Vision. Amsterdam, The Netherlands. 2016. 630–645.

20 Narayanan A, Wang DL. Ideal ratio mask estimation using deep neural networks for robust speech recognition. Proceedings of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver, BC, Canada. 2013. 7092–7096.

21 Vincent E, Gribonval R, Fevotte C. Performance measurement in blind audio source separation. IEEE Transactions on Audio, Speech, and Language Processing, 2006, 14(4): 1462–1469. [doi: 10.1109/TSA.2005.858005]