基于深度学习语音分离技术的研究现状与进展

刘文举[1] 聂 帅[1] 梁 山[1] 张学良[2]

摘 要 现阶段, 语音交互技术日益在现实生活中得到广泛的应用, 然而, 由于干扰的存在, 现实环境中的语音交互技术远没有达到令人满意的程度. 针对加性噪音的语音分离技术是提高语音交互性能的有效途径, 几十年来, 全世界范围内的许多研究者为此投入了巨大的努力, 提出了很多实用的方法. 特别是近年来, 由于深度学习研究的兴起, 基于深度学习的语音分离技术日益得到了广泛关注和重视, 显露出了相当光明的应用前景, 逐渐成为语音分离中一个新的研究趋势. 目前已有很多基于深度学习的语音分离方法被提出, 但是, 对于深度学习语音分离技术一直以来都缺乏一个系统的分析和总结, 不同方法之间的联系和区分也很少被研究. 针对这个问题, 本文试图对语音分离的主要流程和整体框架进行细致的分析和总结, 从特征、模型以及目标三个方面对现有的前沿研究进展进行全面而深入的综述, 最后对语音分离技术进行展望.

关键词 神经网络, 语音分离, 计算听觉场景分析, 机器学习

引用格式 刘文举, 聂帅, 梁山, 张学良. 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

DOI 10.16383/j.aas.2016.c150734

Deep Learning Based Speech Separation Technology and Its Developments

LIU Wen-Ju[1] NIE Shuai[1] LIANG Shan[1] ZHANG Xue-Liang[2]

Abstract Nowadays, speech interaction technology has been widely used in our daily life. However, due to the interferences, the performances of speech interaction systems in real-world environments are far from being satisfactory. Speech separation technology has been proven to be an effective way to improve the performance of speech interaction in noisy environments. To this end, decades of efforts have been devoted to speech separation. There have been many methods proposed and a lot of success achieved. Especially with the rise of deep learning, deep learning-based speech separation has been proposed and extensively studied, which has been shown considerable promise and become a main research line. So far, there have been many deep learning-based speech separation methods proposed. However, there is little systematic analysis and summary on the deep learning-based speech separation technology. We try to give a detail analysis and summary on the general procedures and components of speech separation in this regard. Moreover, we survey a wide range of supervised speech separation techniques from three aspects: 1) features, 2) targets, 3) models. And finally we give some views on its developments.

Key words Neural network, speech separation, computational auditory scene analysis, machine learning

Citation Liu Wen-Ju, Nie Shuai, Liang Shan, Zhang Xue-Liang. Deep learning based speech separation technology and its developments. *Acta Automatica Sinica*, 2016, **42**(6): 819−833

现实环境中, 感兴趣的语音信号通常会被噪音干扰, 严重损害了语音的可懂度, 降低了语音识别的性能. 针对噪音, 前端语音分离技术是最常用的方法之一. 一个好的前端语音分离模块能够极大地提高语音的可懂度和自动语音识别系统的识别性能[1−6]. 然而, 在真实环境中, 语音分离技术的性能远没有达到令人满意的程度, 特别是在非平稳噪音和单声道的情况下, 语音分离依然面临着巨大的挑战. 本文重点探讨单声道条件下语音分离技术.

几十年来, 单声道条件下的语音分离问题被广泛地研究. 从信号处理的角度来看, 许多方法提出估计噪音的功率谱或者理想维纳滤波器, 比如谱减法[7] 和维纳滤波法[8−9]. 其中维纳滤波是最小均方误差意义下分离纯净语音的最优滤波器[9]. 在假定语音和噪音的先验分布的条件下, 给定带噪语音, 它能推断出语音的谱系数. 基于信号处理的方法通常假设噪音是平稳的或者是慢变的[10]. 在满足假设条件的情况下, 这些方法能够取得比较好的分离性能. 然而, 在现实情况下, 这些假设条件通常很难满足, 其分离性能会严重地下降, 特别在低信噪比条件下, 这些方法通常会失效[9]. 相比于信号处理的方法, 基于模型的方法利用混合前的纯净信号分别构建语音

收稿日期 2015-11-04 录用日期 2016-04-01
Manuscript received November 4, 2015; accepted April 1, 2016
国家自然科学基金 (61573357, 61503382, 61403370, 61273267, 91120303, 61365006) 资助
Supported by National Natural Science Foundation of China (61573357, 61503382, 61403370, 61273267, 91120303, 61365006)
本文责任编委 柯登峰
Recommended by Associate Editor KE Deng-Feng
1. 中国科学院自动化研究所模式识别国家重点实验室 北京 100190 2. 内蒙古大学计算机系 呼和浩特 010021
1. National Laboratory of Patten Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190 2. College of Computer Science, Inner Mongolia University, Huhhot 010021

和噪音的模型, 例如文献 [11−13], 在低信噪比的情况下取得了重要的性能提升. 但是基于模型的方法严重依赖于事先训练的语音和噪音模型, 对于不匹配的语音或者噪音, 其性能通常会严重下降. 在基于模型的语音分离方法中, 非负矩阵分解是常用的建模方法, 它能挖掘非负数据中的局部基表示, 目前已被广泛地应用到语音分离中[14−15]. 然而非负矩阵分解是一个浅层的线性模型, 很难挖掘语音数据中复杂的非线性结构. 另外, 非负矩阵分解的推断过程非常费时, 很难应用到实际应用中. 计算听觉场景分析是另一个重要的语音分离技术, 它试图模拟人耳对声音的处理过程来解决语音分离问题[16]. 计算听觉场景分析的基本计算目标是估计一个理想二值掩蔽, 根据人耳的听觉掩蔽来实现语音的分离. 相对于其他语音分离的方法, 计算听觉场景分析对噪音没有任何假设, 具有更好的泛化性能. 然而, 计算听觉场景分析严重依赖于语音的基音检测, 在噪音的干扰下, 语音的基音检测是非常困难的. 另外, 由于缺乏谐波结构, 计算听觉场景分析很难处理语音中的清音成分.

语音分离旨在从被干扰的语音信号中分离出有用的信号, 这个过程能够很自然地表达成一个监督性学习问题[17−20]. 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, 学习一个从带噪特征到分离目标 (例如理想掩蔽或者感兴趣语音的幅度谱) 的映射函数[17]. 最近, 监督性语音分离得到了研究者的广泛关注, 取得了巨大的成功. 作为一个新的研究趋势, 相对于传统的语音增强技术[9], 监督性语音分离不需要声源的空间方位信息, 且对噪音的统计特性没有任何限制, 在单声道, 非平稳噪声和低信噪比的条件下显示出了明显的优势和相当光明的研究前景[21−23].

从监督性学习的角度来看, 监督性语音分离主要涉及特征、模型和目标三个方面的内容. 语音分离系统通常利用时频分解技术从带噪语音中提取时频域特征, 常用的时频分解技术有短时傅里叶变换 (Short-time Fourier transform, STFT)[24] 和 Gammatone 听觉滤波模型[25]. 相应地, 语音分离特征可以分为傅里叶变换域特征和 Gammatone 滤波变换域特征. Wang 和 Chen 等在文献 [26−27] 中系统地总结和分析了 Gammatone 滤波变换域特征, 提出了一些列组合特征和多分辨率特征. 而 Mohammadiha、Xu、Weninger、Le Roux、Huang 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[14, 18, 20, 23, 28−29]. 从建模单元来区分, 语音分离的特征又可分为时频单元级别的特征和帧级别的特征. 时频单元级别的特征从一个时频单元的信号中提取, 帧级别的特征从一帧信号中提取, 早期, 由于模型学习能力的限制, 监督性语音分离方法通常对时频单元进行建模, 因此使用时频单元级别的特征, 例如文献 [1] 和文献 [30−34]. 现阶段, 监督性语音分离主要使用帧级别的特征[17−21, 23, 35−36]. 监督性语音分离系统的学习模型主要分为浅层模型和深层模型. 早期的监督性语音分离系统主要使用浅层模型, 比如高斯混合模型 (Gaussian mixture model, GMM)[1]、支持向量机 (Support vector machine, SVM)[26, 30, 32] 和非负矩阵分解 (Nonnegative matrix factorization, NMF)[14]. 然而, 语音信号具有明显的时空结构和非线性关系, 浅层结构在挖掘这些非线性结构信息的能力上非常有限. 而深层模型由于其多层次的非线性处理结构, 非常擅长于挖掘数据中的结构信息, 能够自动提取抽象化的特征表示, 因此, 近年来, 深层模型被广泛地应用到语音和图像处理中, 并取得了巨大的成功[37]. 以深度神经网络 (Deep neural network, DNN) 为代表的深度学习[37] 是深层模型的典型代表, 目前已被广泛应用到语音分离中[5, 18, 20, 22, 29, 38−39]. 最近, Le Roux、Hershey 和 Hsu 等将 NMF 扩展成深层结构并应用到语音分离中, 取得了巨大的性能提升[23, 40−41], 在语音分离中显示了巨大的研究前景, 日益得到人们的重视. 理想时频掩蔽和目标语音的幅度谱是监督性语音分离的常用目标, 如果不考虑相位的影响, 利用估计的掩蔽或者幅度谱能够合成目标语音波形, 实验证明利用这种方法分离的语音能够显著地抑制噪音[42−43], 提高语音的可懂度和语音识别系统的性能[38, 44−49]. 但是, 最近的一些研究显示, 相位信息对于语音的感知质量是重要的[50]. 为此, 一些语音分离方法开始关注相位的估计, 并取得了分离性能的提升[51−52]. 为了将语音的相位信息考虑到语音分离中, Williamson 等将浮值掩蔽扩展到复数域, 提出复数域的掩蔽目标, 该目标在基于深度神经网络的语音分离系统中显著地提高了分离语音的感知质量[53].

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. 然而, 对监督性语音分离方法一直以来缺乏一个系统的分析和总结, 尽管有一些综述性的工作被提出, 但是它们往往局限于其中的一个方面, 例如, Wang 等在文献 [17] 中侧重于监督性语音分离的目标分析, 而在文献 [26] 中主要比较了监督性语音分离的特征, 并没有一个整体的总结和分析, 同时也没有对这些工作的相互联系以及区别进行研究. 本文从监督性语音分离涉及到的特征、模型和目标三个主要方面

对语音分离的一般流程和整体框架进行了详细的介绍、归纳和总结. 以此希望为该领域的研究及应用提供一个参考.

本文的组织结构如下: 第 1 节概述了语音分离的主要流程和整体框架; 第 2～5 节分别介绍了语音分离中的时频分解、特征、目标、模型等关键模块; 最后, 对全文进行了总结和展望, 并从建模单元、目标和训练模型三个方面对监督性语音分离方法进行了比较和分析.

1　系统结构

图 1 给出了语音分离的一般性结构框图, 主要分为 5 个模块: 1) 时频分解, 通过信号处理的方法 (听觉滤波器组或者短时傅里叶变换) 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, 比如, 短时傅里叶变换谱 (FFT-magnitude)、短时傅里叶变换对数谱 (FFT-log)、Amplitude modulation spectrogram (AMS)、Relative spectral transform and perceptual linear prediction (RASTA-PLP)、Mel-frequency cepstral coefficients (MFCC)、Pitch-based features 以及 Multi-resolution cochleagram (MRCG). 3) 分离目标, 利用估计的分离目标以及混合信号合成目标语音的波形信号. 针对语音分离的不同应用特点, 例如针对语音识别, 语音分离在分离语音的过程中侧重减少语音畸变和尽可能地保留语音成分. 针对语音通讯, 语音分离侧重于提高分离语音的可懂度和感知质量. 常用的语音分离目标主要分为时频掩蔽的目标、目标语音幅度谱估计的目标和隐式时频掩蔽目标, 时频掩蔽目标训练一个模型来估计一个理想时频掩蔽, 使得估计的掩蔽和理想掩蔽尽可能相似; 目标语音幅度谱估计的方法训练一个模型来估计目标语音的幅度谱, 使得估计的幅度谱与目标语音的幅度谱尽可能相似; 隐式时频掩蔽目标将时频掩蔽技术融合到实际应用的模型中, 用来增强语音特征或估计目标语音, 隐式掩蔽并不直接估计理想掩蔽, 而是作为中间的一个计算过程来得到最终学习的目标, 隐式掩蔽作为一个确定性的计算过程, 并没有参数需要学习, 最终的目标误差通过隐式掩蔽的传导来更新模型参数. 4) 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带噪特征到分离目标的映射函数, 应用于语音分离的学习模型大致可分为浅层模型 (GMM, SVM, NMF) 和深层模型 (DNN, DSN, CNN, RNN, LSTM, Deep NMF). 5) 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 Gammatone 滤波) 获得目标语音的波形信号.

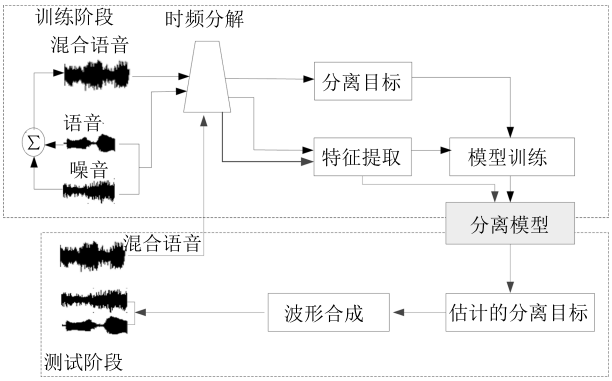

图 1　监督性语音分离系统的结构框图

Fig. 1　A block diagram of the supervised speech separation system

2　时频分解

时频分解是整个系统的前端处理模块, 通过时频分解, 输入的一维时域信号能够被分解成二维的时频信号. 常用的时频分解方法包括短时傅里叶变换[24] 和 Gammatone 听觉滤波模型[25].

假设 $w(t) = w(-t)$ 是一个实对称窗函数, $X(t,f)$ 是一维时域信号 $x(k)$ 在第 t 时间帧、第 f 个频段的短时傅里叶变换系数, 则

$$X(t,f) = \int_{-\infty}^{+\infty} x(k)w(k-t)\exp(-\mathrm{j}2\pi fk)\mathrm{d}k \tag{1}$$

对应的傅里叶能量幅度谱 $p_x(t,f)$ 为

$$p_x(t,f) = |X(t,f)| \tag{2}$$

其中, $|\cdot|$ 表示复数域的取模操作. 为了简化符号表示, 用向量 $\boldsymbol{p} \in \mathbf{R}_+^{F\times 1}$ 表示时间帧为 t 的幅度谱, 这里 F 是傅里叶变换的频带数. 短时傅里叶变换是完备而稳定的[54], 可以通过短时傅里叶逆变换 (Inverse short-time Fourier transform, ISTFT) 从 $X(t,f)$ 精确重构 $x(k)$. 也就是说, 可以通过估计目标语音的短时傅里叶变换系数来实现语音的分离或者增强, 用 $\hat{Y}_s(t,f)$ 来表示估计的目标语音的短时傅里叶变换系数, 那么目标语音的波形 $\hat{s}(k)$ 可以通过 ISTFT 计算

$$\hat{s}(k) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \hat{Y}_s(k)w(k-t)\exp(\mathrm{j}2\pi fk)\mathrm{d}f\mathrm{d}k \tag{3}$$

如果不考虑相位的影响, 语音分离过程可以转换为目标语音幅度谱的估计问题, 一旦估计出目标

语音的幅度谱 $\hat{\boldsymbol{y}}_s \in \mathbf{R}_+^{F\times 1}$, 联合混合语音的相位, 通过 ISTFT, 能得到目标语音的估计波形 $\hat{s}(k)$[17].

Gammatone 听觉滤波使用一组听觉滤波器 $g(t)$ 对输入信号进行滤波, 得到一组滤波输出 $G(k,f)$. 滤波器组的冲击响应为

$$g(t)=\begin{cases}t^{l-1}\exp(-2\pi bt)\cos(2\pi ft), & t\geq 0\\ 0, & \text{其他}\end{cases} \tag{4}$$

其中, 滤波器阶数 $l=4$, b 为等效矩形带宽 (Equivalent rectangle bandwidth, ERB), f 为滤波器的中心频率, Gammatone 滤波器组的中心频率沿对数频率轴等间隔地分布在 [80 Hz, 5 kHz]. 等效矩形带宽与中心频率一般满足式 (5) , 可以看出随着中心频率的增加, 滤波器带宽加宽.

$$ERB(f)=24.7(0.0043f+1.0) \tag{5}$$

对于 4 阶的 Gammatone 滤波器, Patterson 等[25] 给出了带宽的计算公式

$$b=1.093ERB(f) \tag{6}$$

然后, 采用交叠分段的方法, 以 20 ms 为帧长, 10 ms 为偏移量对每一个频率通道的滤波响应做分帧加窗处理. 得到输入信号的时频域表示, 即时频单元. 在计算听觉场景分析系统中, 时频单元被认为是处理的最小单位, 用 T-F 表示. 通过计算时频单元内的内毛细胞输出 (或者听觉滤波器输出) 能量, 就得到了听觉谱 (Cochleagram) , 本文用 $GF(t,f)$ 表示时间帧 t 频率为 f 的时频单元 T-F 的听觉能量.

3 特征

语音分离能够被表达成一个学习问题, 对于机器学习问题, 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. 从特征提取的基本单位来看, 主要分为时频单元级别的特征和帧级别的特征. 时频单元级别的特征是从一个时频单元的信号中提取特征, 这种级别的特征粒度更细, 能够关注到更加微小的细节, 但是缺乏对语音的全局性和整体性的描述, 无法获取到语音的时空结构和时序相关性, 另外, 一个时频单元的信号, 很难表征可感知的语音特性 (例如, 音素). 时频单元级别的特征主要用于早期以时频单元为建模单元的语音分离系统中, 例如, 文献 [1, 26, 30−32], 这些系统孤立地看待每个时频单元, 在每一个频带上训练二值分类器, 判断每一个频带上的时频单元被语音主导还是被噪音主导; 帧级别的特征是从一帧信号中提取的, 这种级别的特征粒度更大, 能够抓住语音的时空结构, 特别是语音的频带相关性, 具有更好的全局性和整体性, 具有明显的语音感知特性. 帧级别的特征主要用于以帧为建模单元的语音分离系统中, 这些系统一般输入几帧上下文帧级别特征, 直接预测整帧的分离目标, 例如, 文献 [17−20, 27, 35]. 近年来, 随着语音分离研究的深入, 已有许多听觉特征被提出并应用到语音分离中, 取得了很好的分离性能. 下面, 我们简要地总结几种常用的听觉特征.

1) 梅尔倒谱系数 (Mel-frequency cepstral coefficient, MFCC). 为了计算 MFCC, 输入信号进行 20 ms 帧长和 10 ms 帧移的分帧操作, 然后使用一个汉明窗进行加窗处理, 利用 STFT 计算能量谱, 再将能量谱转化到梅尔域, 最后, 经过对数操作和离散余弦变换 (Discrete cosine transform, DCT) 并联合一阶和二阶差分特征得到 39 维的 MFCC.

2) PLP (Perceptual linear prediction). PLP 能够尽可能消除说话人的差异而保留重要的共振峰结构, 一般认为是与语音内容相关的特征, 被广泛应用到语音识别中. 和语音识别一样, 我们使用 12 阶的线性预测模型, 得到 13 维的 PLP 特征.

3) RASTA-PLP (Relative spectral transform PLP). RASTA-PLP 引入了 RASTA 滤波到 PLP[55], 相对于 PLP 特征, RASTA-PLP 对噪音更加鲁棒, 常用于鲁棒性语音识别. 和 PLP 一样, 我们计算 13 维的 RASTA-PLP 特征.

4) GFCC (Gammatone frequency cepstral coefficient). GFCC 特征是通过 Gammatone 听觉滤波得到的. 我们对每一个 Gammatone 滤波输出按照 100 Hz 的采样频率进行采样. 得到的采样通过立方根操作进行幅度压制. 最后, 通过 DCT 得到 GFCC. 根据文献 [56] 的建议, 一般提取 31 维的 GFCC 特征.

5) GF (Gammatone feature). GF 特征的提取方法和 GFCC 类似, 只是不需要 DCT 步骤. 一般提取 64 维的 GF 特征.

6) AMS (Amplitude modulation spectrogram). 为了计算 AMS 特征, 输入信号进行半波整流, 然后进行四分之一抽样, 抽样后的信号按照 32 ms 帧长和 10 ms 帧移进行分帧, 通过汉明窗加窗处理, 利用 STFT 得到信号的二维表示, 并计算 STFT 幅度谱, 最后利用 15 个中心频率均匀分布在 15.6 ∼ 400 Hz 的三角窗, 得到 15 维的 AMS 特征.

7) 基于基音的特征 (Pitch-based feature). 基于基音的特征是时频单元级别的特征, 需要对每一个时频单元计算基音特征. 这些特征包含时频单元被目标语音主导的可能性. 我们计算输入信号的 Cochleagram, 然后对每一个时频单元计算 6 维的基音特征, 详细的计算方法可以参考文献 [26, 57].

8) MRCG (Multi-resolution cochleagram).

MRCG 的提取是基于语音信号的 Cochleagram 表示的. 通过 Gammatone 滤波和加窗分帧处理, 我们能得到语音信号的 Cochleagram 表示, 然后通过以下步骤可以计算 MRCG.

步骤 1. 给定输入信号, 计算 64 通道的 Cochleagram, CG1, 对每一个时频单元取对数操作.

步骤 2. 同样地, 用 200 ms 的帧长和 10 ms 的帧移计算 CG2.

步骤 3. 使用一个长为 11 时间帧和宽为 11 频带的方形窗对 CG1 进行平滑, 得到 CG3.

步骤 4. 和 CG3 的计算类似, 使用 23×23 的方形窗对 CG1 进行平滑, 得到 CG4.

步骤 5. 串联 CG1, CG2, CG3 和 CG4 得到一个 64×4 的向量, 即为 MRCG.

MRCG 是一种多分辨率的特征, 既有关注细节的高分辨率特征, 又有把握全局性的低分辨率特征.

9) 傅里叶幅度谱 (FFT-magnitude). 输入的时域信号进行分帧处理, 然后对每帧信号进行 STFT, 得到 STFT 系数, 然后对 STFT 进行取模操作即得到 STFT 幅度谱.

10) 傅里叶对数幅度谱 (FFT-log-magnitude). STFT 对数幅度谱是在 STFT 幅度谱的基础上取对数操作得到的, 主要目的是凸显信号中的高频成分.

以上介绍的听觉特征是语音分离的主要特征, 这些特征之间既存在互补性又存在冗余性. 研究显示, 对于单个独立特征, GFCC 和 RASTA-PLP 分别是噪音匹配条件和噪音不匹配条件下的最好特征[26]. 基音反映了语音的固有属性, 基于基音的特征对语音分离具有重要作用, 很多研究显示基于基音的特征和其他特征进行组合都会显著地提高语音分离的性能, 而且基于基音的特征非常鲁棒, 对于不匹配的听觉条件具有很好的泛化性能. 然而, 在噪音条件下, 准确地估计语音的基音是非常困难, 又因为缺乏谐波结构, 基于基音的特征仅能用于浊音段的语音分离, 而无法处理清音段, 因此在实际应用中, 基于基音的特征很少应用到语音分离中[26], 实际上, 语音分离和基音提取是一个 "鸡生蛋, 蛋生鸡" 的问题, 它们之间相互促进而又相互依赖. 针对这一问题, Zhang 等巧妙地将基音提取和语音分离融合到深度堆叠网络 (Deep stacking network, DSN) 中, 同时提高了语音分离和基音提取的性能[34]. 相对于基音特征, AMS 同时具有清音和浊音的特性, 能够同时用于浊音段和清音段的语音分离, 然而, AMS 的泛化性能较差[58]. 针对各个特征之间的不同特性, Wang 等利用 Group Lasso 的特征选择方法得到 AMS + RASTA − PLP + MFCC 的最优特征组合[26], 这个组合特征在各种测试条件下取得了稳定的语音分离性能而且显著地优于单独的特征, 成为早期语音分离系统最常用的特征. 在低信噪比条件下, 特征提取对于语音分离至关重要, 相对于其他特征或者组合特征, Chen 等提取的多分辨率特征 MRCG 表现了明显的优势[27], 逐渐取代 AMS + RASTA − PLP + MFCC 的组合特征成为语音分离最常用的特征之一. 在傅里叶变换域条件下, FFT-magnitude 或 FFT-log-magnitude 是最常用的语音分离特征, 由于高频能量较小, 相对于 FFT-magnitude, FFT-log-magnitude 能够凸显高频成分, 但是, 一些研究表明, 在语音分离中, FFT-magnitude 要略好于 FFT-log-magnitude[28].

语音分离发展到现阶段, MRCG 和 FFT-magnitude 分别成为 Gammtone 域和傅里叶变换域下最主流的语音分离特征. 此外, 为了抓住信号中的短时变化特征, 一般还会计算特征的一阶差分和二阶差分, 同时, 为了抓住更多的信息, 通常输入特征会扩展上下文帧. Chen 等还提出使用 ARMA 模型 (Auto-regressive and moving average model) 对特征进行平滑处理, 来进一步提高语音分离性能[27].

4 目标

语音分离有许多重要的应用, 总结起来主要有两个方面: 1) 以人耳作为目标受体, 提高人耳对带噪语音的可懂度和感知质量, 比如应用于语音通讯; 2) 以机器作为目标受体, 提高机器对带噪语音的识别准确率, 例如应用于语音识别. 对于这两个主要的语音分离目标, 它们存在许多密切的联系, 例如, 以提高带噪语音的可懂度和感知质量为目标的语音分离系统通常可以作为语音识别的前端处理模块, 能够显著地提高语音识别的性能[59], Weninger 等指出语音分离系统的信号失真比 (Signal-to-distortion ratio, SDR) 和语音识别的字错误率 (Word error rate, WER) 有明显的相关性[5], Weng 等将多说话人分离应用于语音识别中也显著地提高了识别性能[6]. 尽管如此, 它们之间仍然存在许多差别, 以提高语音的可懂度和感知质量为目标的语音分离系统侧重于去除混合语音中的噪音成分, 往往会导致比较严重的语音畸变, 而以提高语音识别准确率为目标的语音分离系统更多地关注语音成分, 在语音分离过程中尽可能保留语音成分, 避免语音畸变. 针对语音分离两个主要目标, 许多具体的学习目标被提出, 常用的分离目标大致可以分为三类: 时频掩蔽、语音幅度谱估计和隐式时频掩蔽. 其中时频掩蔽和语音幅度谱估计的目标被证明能显著地抑制噪音, 提高语音的可懂度和感知质量[17−18]. 而隐式时频掩蔽通常将掩蔽技术融入到实际应用模型中, 时频掩蔽作为中间处理过程来提高其他目标的性能, 例如语音识别[5, 60]、目标语音波形的估计[21].

4.1 时频掩蔽

时频掩蔽是语音分离的常用目标, 常见的时频掩蔽有理想二值掩蔽和理想浮值掩蔽, 它们能显著地提高分离语音的可懂度和感知质量. 一旦估计出了时频掩蔽目标, 如果不考虑相位信息, 通过逆变换技术即可合成目标语音的时域波形. 但是, 最近的一些研究显示, 相位信息对于提高语音的感知质量具有重要的作用[50]. 为此, 一些考虑相位信息的时频掩蔽目标被相继提出, 例如复数域的浮值掩蔽 (Complex ideal ratio mask, CIRM)[53].

1) 理想二值掩蔽 (Ideal binary mask, IBM). 理想二值掩蔽 (IBM) 是计算听觉场景分析的主要计算目标[61], 已经被证明能够极大地提高分离语音的可懂度[44−47, 62]. IBM 是一个二值的时频掩蔽矩阵, 通过纯净的语音和噪音计算得到. 对于每一个时频单元, 如果局部信噪比 $SNR(t,f)$ 大于某一局部阈值 (Local criterion, LC), 掩蔽矩阵中对应的元素标记为 1, 否则标记为 0. 具体来讲, IBM 的定义如下:

$$IBM(t,f)=\begin{cases}1, & 若\ SNR(t,f)>LC\\ 0, & 其他\end{cases} \tag{7}$$

其中, $SNR(t,f)$ 定义了时间帧为 t 和频率为 f 的时频单元的局部信噪比. LC 的选择对语音的可懂度具有重大的影响[63], 一般设置 LC 小于混合语音信噪比 5 dB, 这样做的目的是为了保留足够多的语音信息. 例如, 混合语音信噪比是 −5 dB, 则对应的 LC 设置为 −10 dB.

2) 目标二值掩蔽 (Target binary mask, TBM). 类似于 IBM, 目标二值掩蔽 (TBM) 也是一个二值的时频掩蔽矩阵. 不同的是, TBM 是通过纯净语音的能量和一个固定的参照噪音 (Speech-shaped noise, SSN) 的能量计算得到的. 也就是说, 在式 (7) 中的 $SNR(t,f)$ 项是用参考的 SSN 而不是实际的噪音计算的. 尽管 TBM 的计算独立于噪音, 但是实验测试显示 TBM 取得了和 IBM 相似的可懂度提高[63]. TBM 能够提高语音的可懂度的原因是它保留了与语音感知密切相关的时空结构模式, 即语音能量在时频域上的分布. 相对于 IBM, TBM 可能更加容易学习.

3) Gammtone 域的理想浮值掩蔽 (Gammtone ideal ratio mask, IRM_Gamm)

理想浮值掩蔽定义如下:

$$IRM_{gamm}(t,f)=\left(\frac{S^2(t,f)}{S^2(t,f)+N^2(t,f)}\right)^{\beta}=\left(\frac{SNR(t,f)}{SNR(t,f)+1}\right)^{\beta} \tag{8}$$

其中, $S^2(t,f)$ 和 $N^2(t,f)$ 分别定义了混合语音中时间帧为 t 和频率为 f 的时频单元的语音和噪音的能量. β 是一个可调节的尺度因子. 如果假定语音和噪音是不相关的, 那么 IRM 在形式上和维纳滤波密切相关[9, 64]. 大量的实验表明 $\beta=0.5$ 是最好的选择, 此时, 式 (8) 和均方维纳滤波器非常相似, 而维纳滤波是最优的能量谱评估[9].

4) 傅里叶变换域的理想浮值掩蔽 (FFT ideal ratio mask, IRM_FFT). 类似于 Gammtone 域的理想浮值掩蔽, 傅里叶域的理想浮值掩蔽 IRM_FFT 的定义如下:

$$IRM_{FFT}(t,f)=\frac{|Y_s(t,f)|^2}{|Y_s(t,f)|^2+|Y_n(t,f)|^2}=\frac{P_s(t,f)}{P_s(t,f)+P_n(t,f)} \tag{9}$$

其中, $Y_s(t,f)$ 和 $Y_n(t,f)$ 是混合语音中纯净的语音和噪音的短时傅里叶变换系数. $P_s(t,f)$ 和 $P_n(t,f)$ 分别是它们对应的能量密度.

5) 短时傅里叶变换掩蔽 (Short-time Fourier transform mask, FFT-Mask). 不同于 IRM_FFT, FFT-Mask 的定义如下:

$$Mask_{FFT}(t,f)=\frac{|Y_s(t,f)|}{|X(t,f)|} \tag{10}$$

其中, $Y_s(t,f)$ 和 $X(t,f)$ 是纯净的语音和混合语音的短时傅里叶变换系数. IRM_FFT 的取值范围在 $[0,1]$, 显然 FFT-Mask 的取值范围可以超过 1.

6) 最优浮值掩蔽 (Optimal ratio time-frequency mask, ORM). 理想浮值掩蔽 (IRM) 是假定在语音和噪音不相关的条件下, 能够取得最小均方误差意义下最大信噪比增益[42−43]. 然而在真实环境中, 语音和噪音通常存在一定的相关性, 针对这个问题, Liang 等[42−43] 推导出一般意义下的最小均方误差的最优浮值掩蔽, 定义如下:

$$ORM(t,f)=\frac{P_s(t,f)+\Re(Y_s(t,f)Y_n^*(t,f))}{P_s(t,f)+P_n(t,f)+2\Re(Y_s(t,f)Y_n^*(t,f))} \tag{11}$$

其中, $\Re(\cdot)$ 表示取复数的实部, * 表示共轭操作. 相对于 IRM, ORM 考虑了语音和噪音的相关性, 其变化范围更大, 估计难度也更大.

7) 复数域的理想浮值掩蔽 (Complex ideal ratio mask, CIRM). 传统的 IRM 定义在幅度域, 而 CIRM 定义在复数域. 其目标是通过将 CIRM 作用到带噪语音的 STFT 系数得到目标语音的 STFT

系数. 具体地, 给定带噪语音在时间帧为 t 和频率为 f 的时频单元的 STFT 系数 X, 那么目标语音在对应时频单元的 STFT 系数 Y 可以通过下式计算得到:

$$Y = M \times X \tag{12}$$

其中, $\times$ 定义复数乘法操作, M 定义时间帧为 t 和频率为 f 的时频单元的 CIRM. 通过数学推导我们能计算得到:

$$M = \frac{X_r Y_r + X_i Y_i}{X_r^2 + X_i^2} + \mathrm{j}\frac{X_r Y_i - X_i Y_r}{X_r^2 + X_i^2} \tag{13}$$

其中, X_r 和 Y_r 分别是 X 和 Y 的实部, X_i 和 Y_i 分别是 X 和 Y 的虚部, j 是虚数单位. 定义 M_r 和 M_i 分别是 M 的实部和虚部, 那么,

$$M_r = \frac{X_r Y_r + X_i Y_i}{X_r^2 + X_i^2} \tag{14}$$

$$M_i = \frac{X_r Y_i - X_i Y_r}{X_r^2 + X_i^2} \tag{15}$$

在实际语音分离中通常不会直接估计复数域的 M, 而是通过估计其实部 M_r 和虚部 M_i. 但 M_r 和 M_i 的取值可能会超过 $[0, 1]$, 这往往会增大 M_r 和 M_i 估计难度. 因此, 在实际应用中通常会利用双曲正切函数对 M_r 和 M_i 进行幅度压制.

4.2　语音幅度谱估计

如果不考虑相位的影响, 语音分离问题可以转化为目标语音幅度谱的评估, 一旦从混合语音中评估出了目标语音的幅度谱, 利用混合语音的相位信息, 通过逆变换即可得到目标语音的波形. 常见的幅度谱包括 Gammtone 域幅度谱和 STFT 幅度谱.

1) Gammtone 域幅度谱 (Gammatone frequency power spectrum, GF-POW). 时域信号经过 Gammtone 滤波器组滤波和分帧加窗处理, 可以得到二维的时频表示 Cochleagram. 直接估计目标语音的 Gammtone 域幅度谱 (GF-POW) 能够实现语音的分离. 对于 Gammtone 滤波, 由于没有直接的逆变换方法, 我们可以通过估计的 GF-POW 和混合语音的 GF-POW 构造一个时频掩蔽 $\sqrt{S_{GF}^2(t,f)/X_{GF}^2(t,f)}$ 来合成目标语音的波形, 其中, $S_{GF}^2(t,f)$ 和 $X_{GF}^2(t,f)$ 分别是纯净语音和混合语音在 Gammtone 域下时间帧为 t 和频带为 f 的时频单元的能量.

2) 短时傅里叶变换幅度谱 (Short-time Fourier transform spectral magnitude, FFT-magnitude). 时域信号经过分帧加窗处理, 然后通过 STFT, 可以得到二维的时频表示, 如果不考虑相位的影响, 我们可以直接估计目标语音的 STFT 幅度谱, 利用原始混合语音的相位, 通过 IFTST 可以估计得到目标语音的时域波形.

4.3　隐式时频掩蔽

语音分离旨在从混合语音中分离出语音成分, 尽管可以通过估计理想时频掩蔽来分离目标语音, 但理想时频掩蔽是一个中间目标, 并没有针对实际的语音分离应用直接优化最终的实际目标. 针对这些问题, 隐式时频掩蔽被提取, 在这些方法中, 时频掩蔽作为一个确定性的计算过程被融入到具体应用模型中, 例如识别模型或者分离模型, 它们并没有估计理想时频掩蔽, 其最终的目标是估计目标语音的幅度谱甚至是波形, 或者提高语音识别的准确率.

Huang 等提出将掩蔽融合到目标语音的幅度谱估计中[19, 28]. 在文献 [29] 中, 深度神经网络作为语音分离模型, 时频掩蔽函数作为额外的处理层加入到网络的原始输出层, 如图 2 所示, 通过时频掩蔽, 目标语音的幅度谱从混合语音的幅度谱中估计出来. 其中时频掩蔽函数 M_s 和 M_n 通过神经网络原始输出 (语音和噪音幅度谱的初步估计) 计算得到的, 如下:

$$M_s = \frac{|\hat{\boldsymbol{y}}_s|}{|\hat{\boldsymbol{y}}_s| + |\hat{\boldsymbol{y}}_n|}, \quad M_n = \frac{|\hat{\boldsymbol{y}}_n|}{|\hat{\boldsymbol{y}}_s| + |\hat{\boldsymbol{y}}_n|} \tag{16}$$

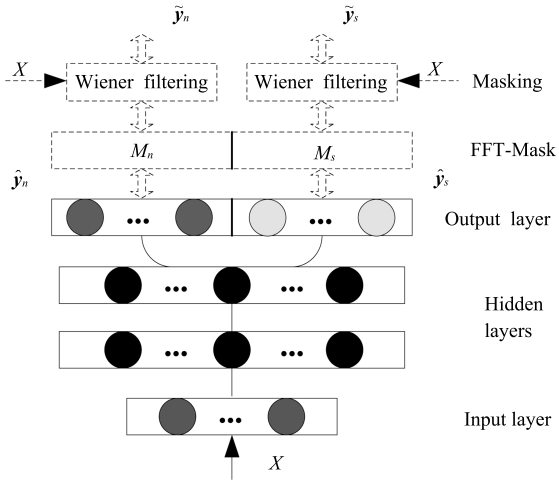

图 2　Huang 等提出的声源分离系统的网络结构[28]

Fig. 2　The network structure of the proposed source separation system by Huang et al.[28]

一旦时频掩蔽被计算出来, 就可以通过掩蔽技术从混合语音的幅度谱 X 估计出语音和噪音的幅度谱 $\tilde{\boldsymbol{y}}_s$ 和 $\tilde{\boldsymbol{y}}_n$:

$$\tilde{\boldsymbol{y}}_s = M_s \otimes X, \quad \tilde{\boldsymbol{y}}_n = M_n \otimes X \tag{17}$$

需要注意的是原始的网络输出并不用来计算误差函数, 仅仅用来计算时频掩蔽函数, 时频掩蔽函数是确定性的, 并没有连接权重, 掩蔽输出用来计算误差并以此来更新模型参数.

Wang 等提出将时频掩蔽融合到目标语音波形估计中[21], 在文献 [21] 中, 时频掩蔽作为神经网络的一部分, 掩蔽函数从混合语音的 STFT 幅度谱估计目标语音的 STFT 幅度谱, 然后通过 ISTFT, 利用混合语音的相位信息和估计的 STFT 幅度谱合成目标语音的时域波形, 如图 3 所示, 估计的时域波形与目标波形计算误差, 最后通过反向传播更新网络权重. 假设 $\boldsymbol{m}$ 是最后一个隐层的输出, 它可以看成是估计的掩蔽, 用来从混合语音的 STFT 幅度谱 $\boldsymbol{x}$ 中估计目标语音的 STFT 幅度谱 $\tilde{\boldsymbol{y}}_s$.

$$\tilde{\boldsymbol{y}}_s = \boldsymbol{m} \otimes \boldsymbol{x} \tag{18}$$

评估的目标语音幅度谱加上对应的混合语音的相位 $\boldsymbol{p}$ 输入到逆傅里叶变换层能够得到目标语音的时域波形信号.

$$\hat{\boldsymbol{s}} = \mathrm{IFFT}([\boldsymbol{c}, \mathrm{flipud}((\boldsymbol{c}_{2:d})^*)]^{\mathrm{T}}) \tag{19}$$

其中, $\boldsymbol{c}$ 是目标语音的傅里叶变换系数, 它能够通过下面的公式得到:

$$\boldsymbol{c} = \tilde{\boldsymbol{y}}_s \otimes \mathrm{e}^{\mathrm{j}\boldsymbol{p}} \tag{20}$$

$d = F/2$, F 是傅里叶变换的分析窗长. flipud 定义了向量的上下翻转操作, * 是复数的共轭操作. 下标 $m:n$ 取向量从 m 到 n 的元素的操作.

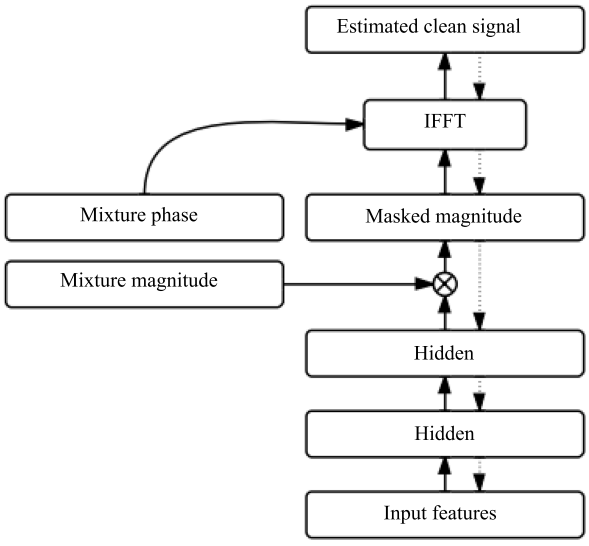

图 3 Wang 等提出的语音分离系统的网络结构[21]

Fig. 3 The network structure of the proposed speech separation system by Wang et al. for speech separation[21]

Narayanan 等[60] 提出将时频掩蔽融入到语音识别的声学模型中, 时频掩蔽作为神经网络的中间处理层, 从带噪的梅尔谱特征中掩蔽出语音的梅尔谱特征, 然后输入到下层网络中进行状态概率估计, 如图 4 所示. 注意时频掩蔽仅仅是神经网络的中间处理层的输出, 并不是以理想时频掩蔽作为目标学习而来的, 确切地说是根据语音识别的状态目标学习而来的, 实验结果显示, 时频掩蔽输出具有明显的降噪效果, 这从侧面显示了语音识别与语音分离之间存在密切联系.

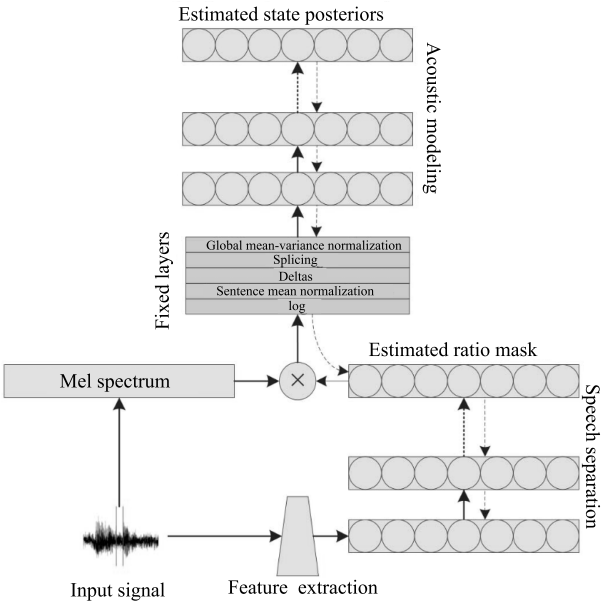

图 4 Narayanan 等提出的神经网络的结构[60]

Fig. 4 The structure of the proposed network by Narayanan et al.[60]

以上介绍的是监督性语音分离的主要目标. 在时频掩蔽目标中, 理想二值掩蔽具有最为简单的形式, 而且具有听觉感知掩蔽的心理学依据, 是早期监督性语音分离最常用的分离目标, 其分离的语音能够极大地提高语音的可懂度, 然而语音的听觉感知质量往往得不到提高. 理想浮值掩蔽具有 0 到 1 的平滑形式, 近似于维纳滤波, 不仅能够提高分离语音的可懂度而且能够显著地提高语音的感知质量, 在语音和噪音独立的情况下能够取得最优信噪比增益. 相比于理想二值掩蔽, 最优浮值掩蔽是更为一般意义下的最优信噪比增益目标, 它考虑了语音和噪音之间的相关性, 但是最优浮值掩蔽的学习难度也相对较大, 目前尚未应用到监督性语音分离中. 之前大部分时频掩蔽, 都没有考虑语音的相位信息, 而研究表明相位信息对于提高语音的感知质量具有重要作用. 复数域的理想浮值掩蔽考虑了语音的相位信息, 被应用到监督性语音分离中, 取得了显著的性能提

高. 语音幅度谱估计的目标直接估计目标语音的幅度谱, 相对于时频掩蔽目标, 更为直接也更为灵活, 然而其学习难度也更大, 常用的语音幅度谱估计的目标是短时傅里叶变换域的语音幅度谱. 隐式时频掩蔽目标并没有直接估计理想时频掩蔽, 而是将时频掩蔽融入到实际应用的模型中, 直接估计最终的目标, 例如目标语音的波形或者语音识别的状态概率, 这种方式学习到的时频掩蔽和实际目标最相关, 目前, 正广泛应用于语音分离系统中.

5 模型

语音分离能够很自然地表达成一个监督性学习问题, 一个典型的监督性语音分离系统利用学习模型学习一个从带噪特征到分离目标的映射函数. 目前已有许多学习模型应用到语音分离中, 常用的模型大致可以分为两类: 浅层模型和深层模型. 在早期的监督性语音分离中[1, 14, 31], 浅层模型通常直接对输入的带噪时频单元的分布进行概率建模或区分性建模, 例如 GMM[1] 和 SVM[31], 或者直接对输入的带噪特征数据进行矩阵分解, 以推断混合数据中语音和噪音的成分, 例如 NMF[14]. 由于浅层模型没有从数据中自动抽取有用特征的能力, 因此, 它们严重依赖于人工设计的特征, 另外, 浅层模型对高维数据处理的能力通常比较有限, 很难通过扩展上下文帧来挖掘语音信号中的时频相关性. 深层模型是近几年来受到极大关注的学习模型, 在语音和图像等领域都取得了巨大的成功. 由于深层模型层次化的非线性处理, 使得它能够自动抽取输入数据中对目标最有力的特征表示, 相比于浅层模型, 深层模型能够处理更原始的高维数据, 对特征设计的知识要求相对较低, 而且深层模型擅长于挖掘数据中的结构化特性和结构化输出预测. 由于语音的产生机制, 语音分离的输入特征和输出目标都呈现了明显的时空结构, 这些特性非常适合用深层模型来进行建模. 许多深层模型广泛应用到语音分离中, 包括 DNN[18]、DSN[33−34]、CNN[22]、RNN[19−20, 28]、Deep NMF[23] 和 LSTM[39].

5.1 浅层模型

1) 高斯混合模型 (GMM). 高斯混合模型能够刻画任意复杂的分布, Kim 等[1] 利用 GMM 分别对每一个频带被目标语音主导的时频单元和被噪音主导的时频单元进行建模, 这里各个频带是独立建模的, 在测试阶段, 给定时频单元的输入特征, 计算被目标语音主导和被噪音主导的概率, 然后进行贝叶斯推断, 判断时频单元是被目标语音主导还是被噪音主导, 如果被语音主导标记为 1, 否则标记为 0, 当所有的时频单元被判断出来, 则二值掩蔽被估计出来. 最后, 利用估计的二值掩蔽和混合语音的 Gammtone 滤波输出合成目标语音的时域波形.

高斯混合模型是一种生成式的模型, 目标语音主导的时频单元的概率分布和噪音主导的时频单元的概率分布有很多重叠部分, 并且它不能挖掘特征中的区分信息, 不能进行区分性训练. 孤立地对每一个频带建模, 无法利用频带间的相关性, 同时会导致训练和测试代价过大, 很难具有实用性.

2) 支持向量机 (SVM). 支持向量机能够学习数据中的最优分类面, 以区分不同类别的数据. Han 等[32] 提出用 SVM 对每一个频带的时频单元进行建模, 学习被目标语音主导的时频单元和被噪音主导的时频单元最优区分面. 在测试阶段, 输入时频单元的特征, 通过计算到分类面的距离实现时频单元的分类.

相比于 GMM, SVM 取得了更好的分类准确性和泛化性能. 这主要得益于 SVM 的区分性训练. 但是 SVM 仍然是对每一个时频单元进行单独建模, 忽略了它们之间的相关性和语音的时空结构特性, 同时 SVM 是浅层模型, 并没有特征抽象和层次化学习的能力.

3) 非负矩阵分解 (NMF). 非负矩阵分解是著名的表示学习方法, 它能挖掘隐含在非负数据中的局部表示. 给定非负矩阵 $X \in \mathbf{R}_+$, 非负矩阵分解将 X 近似分解成两个非负矩阵的乘积, $X \approx WH$, 其中 W 是非负基矩阵, H 是对应的激活系数矩阵. 当非负矩阵分解应用到纯净语音或者噪音的幅度谱时, NMF 能挖掘出语音或者噪音的基本谱模式. 在语音分离中, 首先在训练阶段, 在纯净的语音和噪音上分别训练 NMF 模型, 得到语音和噪音的基矩阵. 然后, 在测试阶段, 联合语音和噪音的基矩阵得到一个既包含语音成分又包含噪音成分的更大的基矩阵, 利用得到的基矩阵, 通过非负线性组合重构混合语音幅度谱, 当重构误差收敛时, 语音基和噪音基对应的激活矩阵被计算出来, 然后利用非负线性组合即可分离出混合语音中的语音和噪音.

NMF 是单层线性模型, 很难刻画语音数据中的非线性关系, 另外, 在语音分离过程中, NMF 的推断过程非常费时, 很难达到实时性要求, 大大限制了 NMF 在语音分离中的实际应用.

5.2 深层模型

1) 深度神经网络 (DNN). DNN 是最常见的深层模型, 一个典型的 DNN 通常由一个输入层, 若干个非线性隐含层和一个输出层组成, 各个层依次堆叠, 上层的输出输入到下一层中, 形成一个深度的网络. 层次化的非线性处理使得 DNN 具有强大的表示学习的能力, 能够从原始数据中自动学习对目标

最有用的特征表示, 抓住数据中的时空结构. 然而, 深度网络的多个非线性隐含层使得它的优化非常困难, 往往陷入性能较差的局部最优点. 为解决这个问题, 2006 年 Hinton 等提出了一种无监督的预训练方式, 极大地改善了深度神经网络的优化问题[65]. 自此, 深度神经网络得到广泛的研究, 在语音和图像等领域取得巨大的成功. Xu 等将 DNN 应用于语音分离中, 取得了显著的性能提升. 在文献 [18] 中, DNN 被用来学习一个从带噪特征到目标语音的对数能量幅度谱的映射函数, 如图 5 所示. 上下文帧的对数幅度谱作为输入特征, 通过两个隐层的非线性变换和输出层的线性变换, 估计得到对应帧的目标语音的对数幅度谱. 最后, 使用混合语音的相位, 利用 ISTFT 得到目标语音的时域波形信号.

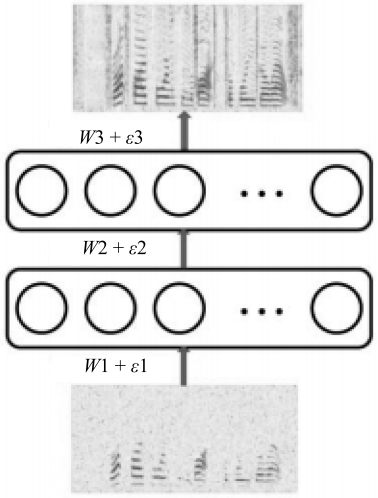

图 5 Xu 等提出的基于 DNN 的语音分离系统的网络结构[18]

Fig. 5 The structure of the proposed DNN-based speech separation system by Xu et al.[18]

实验结果显示 Xu 等提出的方法在大规模训练数据上取得了优异的语音分离性能. 然而, 直接估计目标语音的对数幅度谱是一个非常困难的任务, 需要大量的训练数据才能有效地训练模型, 同时其泛化性能也是一个重要的问题.

2) 深度堆叠网络 (DSN). 语音信号具有很强的时序相关性, 探究这些特性能够提高语音分离的性能, 为此, Nie 等[33] 利用 DSN 的层次化模块结构对时频单元的时序相关性进行建模, 定义为 DSN-TS, 如图 6 所示. DSN 的基本模块是一个由一个输入层, 一个隐层和一个线性输出层组成的前向传播网络. 模块之间相互堆叠, 每一个模块依次对应一个时刻帧, 前一个模块的输出连接上下一个时刻的输入特征作为下一个模块的输入, 如此类推, 便可估计所有时刻的时频单元的掩蔽.

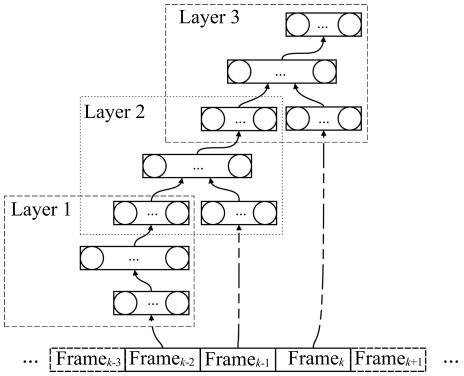

图 6 Nie 等提出的基于 DSN-TS 的语音分离系统的网络结构[33]

Fig. 6 The structure of the proposed DSN-TS-based speech separation system by Nie et al.[33]

相比于之前的模型, DSN-TS 考虑了时频单元之间的时序关系, 在分离性能上取得进一步提高. 然而, DSN-TS 对每一个频带单独建模, 忽略了频带之间的相关性.

基音是语音的一个显著的特征, 在传统的计算听觉场景分析中常被用作语音分离的组织线索. 基于基音的特征也常被用于语音分离, 然而, 噪音环境下, 基音的提取是一个挑战性的工作, Zhang 等[34] 巧妙地将噪声环境下的基音提取和语音分离融合到 DSN 中, 定义为 DSN-Pitch, 如图 7 所示. 在 DSN-Pitch 中, 基音提取和语音分离交替进行, 相互促进. DSN-Pitch 同时提高了语音分离的性能和基音提取的准确性, 然而, DSN-Pitch 依然对每一个时频单元单独建模, 严重忽略了它们之间的时空相关性.

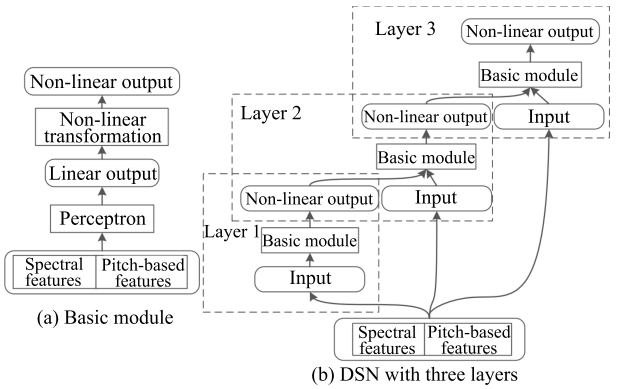

图 7 Zhang 等提出的基于 DSN 的语音分离系统的网络结构[34]

Fig. 7 The structure of the proposed DSN-based speech separation system by Zhang et al.[34]

3) 深度循环神经网络 (Deep recurrent neural network, DRNN). 由于语音的产生机制, 语音具有明显的长短时谱依赖性, 这些特性能够被用来帮助

语音分离. 尽管 DNN 具有强大的学习能力, 但是 DNN 仅能通过上下文或者差分特征对数据中的时序相关性进行有限的建模, 而且会极大地增加输入数据的维度, 大大地增加了学习的难度. RNN 是非常常用的时序模型, 利用其循环连接能够对时序数据中长短时依赖性进行建模. Huang 等将 RNN 应用到语音分离中, 取得比较好的语音分离性能[29]. 标准的 RNN 仅有一个隐层, 为了对语音数据进行层次化抽象, Huang 等[29] 使用深度的 RNN 作为最终的分离模型, 如图 8 所示.

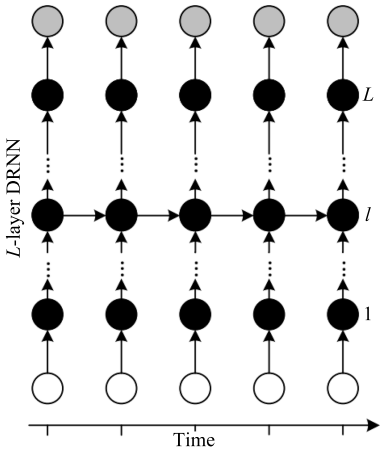

图 8　Huang 等提出的基于 DRNN 的语音分离系统的网络结构[29]

Fig. 8　The structure of the proposed DRNN-based speech separation system by Huang et al.[29]

相对于 DNN, DRNN 能够抓住数据中时序相关性, 但是由于梯度消失的问题, DRNN 不容易训练, 对长时依赖的建模能力有限. 实验结果表明, 相对于 DNN, DRNN 在语音分离中的性能提升比较有限.

4) 长短时记忆网络 (Long short-term memory, LSTM). 作为 RNN 的升级版本, 在网络结构上, LSTM 增加了记忆单元、输入门、遗忘门和输出门, 这些结构单元使得 LSTM 相比于 RNN 在时序建模能力上得到巨大的提升, 能够记忆更多的信息, 并能有效抓住数据中的长时依赖. 语音信号具有明显的长短时依赖性, Weninger 等将 LSTM 应用到语音分离中, 取得了显著的性能提升[5, 39].

5) 卷积神经网络 (Convolutional neural network, CNN). CNN 在二维信号处理上具有天然的优势, 其强大的建模能力在图像识别等任务已得到验证. 在语音分离中, 一维时域信号经过时频分解技术变成二维的时频信号, 各个时频单元在时域和频域上具有很强的相关性, 呈现了明显的时空结构. CNN 擅长于挖掘输入信号中的时空结构, 具有权值共享, 形变鲁棒性特性, 直观地看, CNN 适合于语音分离任务. 目前 CNN 已应用到语音分离中, 在相同的条件下取得了最好的分离性能, 超过了基于 DNN 的语音分离系统[22, 66].

6) 深度非负矩阵分解 (Deep nonnegative matrix factorization, Deep NMF). 尽管 NMF 能抓住隐含在非负数据中的局部基表示, 但是, NMF 是一个浅层线性模型, 很难对非负数据中的结构特性进行层次化的抽象, 也无法处理数据中的非线性关系. 而语音数据中存在丰富的时空结构和非线性关系, 挖掘这些信息能够提高语音分离的性能. 为此, Le Roux 等将 NMF 扩展成深度结构, 在语音分离应用上取得巨大的性能提升[23, 40−41].

总结以上介绍的语音分离模型, 可以看到: 浅层模型, 复杂度低但泛化性能好, 无法自动学习数据中的特征表示, 其性能严重依赖于人工设计的特征, 基于浅层模型的语音分离系统其性能和实用性有限; 深度模型, 复杂度高建模能力强, 其泛化性能可以通过扩大训练数据量来保证, 另外, 深度模型能够自动学习数据中有用的特征表示, 因此对特征设计的要求不高, 而且, 能够处理复杂的高维数据, 可以通过将上下文帧级联输入到深度模型中, 以便为语音分离提供更多的信息. 同时, 深度模型具有丰富的结构, 能够抓住语音数据的很多特性, 例如时序性、时空相关性、长短时谱依赖性和自回归性等. 深度模型能够处理复杂的结构映射, 因此, 能够为基于深度模型的语音分离设计更加复杂的目标来提高语音分离的性能. 目前, 监督性语音分离的主流模型是深度模型, 并开始将浅层模型扩展成深度模型.

6　总结与展望

6.1　总结

本文从时频分解、特征、目标和模型四个方面对基于深度学习的语音分离技术的整体框架和主要流程进行综合概述和分析比较. 对于时频分解, 听觉滤波器组和傅里叶变换是最常用的技术, 它们在分离性能上没有太大的差异. 听觉滤波器组在低频具有更高的分辨率, 但是计算复杂度比较高, 傅里叶变换具有快速算法而且是一个可逆变换, 在监督性语音分离中日益成为主流. 对于特征, 目前帧级别的 MRCG 和 FFT-magnitude 分别是 Gammtone 域下和傅里叶变换域下最主流的特征, 已经在许多研究中得到验证, 是目前语音分离使用最多的特征. 时频掩蔽是监督性语音分离最主要的分离目标, 浮值掩蔽在可懂度和感知质量上都优于二值掩蔽, 目前是语音分离的主流目标. 然而, 时频掩蔽并不能优化实际的分离目标, 傅里叶幅度谱是一种更接近实际

目标的分离目标, 相对于时频掩蔽具有更好的灵活性, 能够进行区分性学习, 日益得到研究者的重视, 但是, 其学习难度更大. 目前结合时频掩蔽和傅里叶幅度谱的隐式掩蔽目标显示了光明的研究前景, 正被日益广泛地应用到语音分离中, 而且各种变形的隐式掩蔽目标正不断地被提出. 自从语音分离被表达成监督性学习问题, 各种学习模型被尝试着应用到语音分离中, 在各种模型中, 深度模型在语音分离任务中显示了强大的建模能力, 取得了巨大的成功, 目前已经成为语音分离的主流方法.

语音分离在近几年来得到研究者的广泛关注, 针对时频分解, 特征、目标和模型都有许多方法被提出. 对现存的方法, 我们从建模单元、分离目标和学习模型三个方面进行简单的分类总结:

1) 建模单元. 语音分离的建模单元主要有两类: a) 时频单元的建模; b) 帧级别的建模. 时频单元的建模对每一个时频单元单独建模, 通过分类模型估计每一个时频单元是被语音主导还是被噪音主导, 早期, 基于二值分类的监督性语音分离通常是基于时频单元建模的. 帧级别的建模将一帧或者若干帧语音的所有时频单元看成一个整体, 同时对它们进行建模, 估计的目标也是帧级别的. 相对于时频单元的建模, 帧级别的建模能够抓住语音的时频相关性, 一般认为帧级别的建模能取得更好的语音分离性能. 目前, 帧级别的建模已成为监督性语音分离的主流建模方法, 从时频单元到帧级别的建模单元转变是一个巨大的进步, 这主要得益于深度学习强大的表示与学习能力, 能够进行结构性输出学习. 传统的浅层学习能力很难学习结构复杂的高维度的输出.

2) 分离目标. 语音分离的目标主要分为三类: a) 时频掩蔽; b) 目标语音的幅度谱估计; c) 隐式时频掩蔽. 时频掩蔽是语音分离的主要目标, 许多监督性语音分离系统从带噪特征中估计目标语音的二值掩蔽或者浮值掩蔽, 然而, 时频掩蔽是一个中间目标, 并没有直接优化实际的语音分离目标. 相对于时频掩蔽, 估计目标语音的幅度谱更接近实际的语音分离目标, 而且更为灵活, 能够更充分利用语音和噪音的特性构造区分性的训练目标. 但是, 由于幅度谱的变化范围比较大, 在学习难度上要比时频掩蔽目标的学习难度大. 深度学习具有强大的学习能力, 现在已有许多基于深度学习的语音分离方法直接估计目标语音的幅度谱. 隐式的时频掩蔽方法将时频掩蔽融合到深度神经网路中, 并不直接估计理想时频掩蔽, 而是作为中间处理层帮助实际目标的估计. 在这里, 时频掩蔽并不是神经网络学习的目标, 它是通过实际应用的目标的估计误差隐式地得到的. 相对于时频掩蔽和目标语音幅度谱估计的方法, 隐式时频掩蔽方法融合了时频掩蔽和目标语音幅度谱估计方法的优势, 能够取得更好的语音分离性能, 目前已有几个基于隐式的时频掩蔽方法的监督性语音分离方法被提出, 并且取得了显著的性能提升.

3) 学习模型. 监督性语音分离的学习模型主要分为两类: a) 浅层模型; b) 深层模型. 早期的监督性语音分离主要使用浅层模型, 一般对每一个频带的时频单元单独建模, 并没有考虑时频单元之间的时空相关性. 基于浅层模型的语音分离系统分离的语音的感知质量通常比较差. 随着深度学习的兴起, 深层模型开始广泛应用到语音分离中, 目前已经成为监督性语音分离最主流的学习模型. 深层模型具有强大的建模能力, 能够挖掘数据中的深层结构, 相对于浅层模型, 深层模型分离的语音不仅在感知质量和可懂度方便都得到巨大的提升, 而且随着数据的增大, 其泛化性能和分离性能得到不断的提高.

6.2 展望

最近几年, 在全世界研究者的共同努力下, 监督性语音分离得到巨大的发展. 针对监督性语音分离的特征、目标和模型三个主要方面都进行了深入细致的研究, 取得许多一致性的共识. 目前, 监督性语音分离的框架基本成熟, 即利用深度模型学习一个从带噪特征到分离目标的映射函数. 很难在框架层面进行重大的改进, 针对现有的框架, 特别是基于深度学习的语音分离框架, 我们认为, 未来监督性语音分离可能在下面几个方面.

1) 泛化性. 尽管监督性语音分离取得了很好的分离性能, 特别是深度学习的应用, 极大地促进了监督性语音分离的发展. 但在听觉条件或者训练数据不匹配的情况下, 例如噪音不匹配和信噪比不匹配的情况下, 分离性能会急剧下降. 目前解决这个问题最有效方法是扩大数据的覆盖面, 但现实情况很难做到覆盖大部分的听觉环境, 同时训练数据增大又带来训练时间的增加, 不利于模型更新. 我们认为解决这个问题的两个可行的方向是: a) 挖掘人耳的听觉心理学知识, 例如将计算听觉场景分析的知识融入到监督性语音分离模型中. 人耳对于声音的处理具有很好的鲁棒性. 早期计算听觉场景分析的许多研究表明基音和听觉掩蔽等对噪音是非常鲁棒的, 将这些积累的人工知识和计算听觉场景分析的处理过程有效融入到监督性语音分离中可能会提高监督性语音分离的泛化性能. b) 更多地关注和挖掘语音的固有特性. 由于人声产生的机理, 人声具有很多明显的特性, 例如稀疏性、时空连续性、明显的谐波结构、自回归性、长短时依赖性等, 对于噪音, 这些特性具有明显的区分性, 而且, 对于不同的人发出的人声, 不同的语言和内容, 人声都具有这些特性. 而噪音的

变化却多种多样, 很难找到共有的模式. 因此应该将更多的精力放到对语音固有特性的研究和挖掘上而不是只关注噪音. 将语音固有的特性融入到监督性语音分离模型中可能会提高语音分离的性能和泛化能力.

2) 生成式模型和监督性模型联合. 人耳对声音的处理过程可能是模式驱动的, 即在大脑高层可能存储有许多关于语音的基本模式, 当我们听到带噪语音的时候, 带噪语音会激发大脑中相似的语音模式响应, 这些先后被激活的语音模式组合起来即可形成大脑可理解的语义单元, 使得人们能从带噪语音中听清语音. 而这些语音模式一方面是从父母继承而来的, 一方面是后天学习来的. 我们可以利用生成式模型从大量纯净的语音中学习语音的基本谱模式, 然后利用监督性学习模型来估计语音基本谱模式的激活量, 利用这些激活的基本谱模式可以重构纯净的语音. 基本谱模式的学习可以利用很多生成式模型, 而它们之间的融合也可以有多种形式, 这方面有许多内容值得进一步探讨.

References

1 Kim G, Lu Y, Hu Y, Loizou P C. An algorithm that improves speech intelligibility in noise for normal-hearing listeners. *The Journal of the Acoustical Society of America*, 2009, **126**(3): 1486−1494

2 Dillon H. *Hearing Aids*. New York: Thieme, 2001.

3 Allen J B. Articulation and intelligibility. *Synthesis Lectures on Speech and Audio Processing*, 2005, **1**(1): 1−124

4 Seltzer M L, Raj B, Stern R M. A Bayesian classifier for spectrographic mask estimation for missing feature speech recognition. *Speech Communication*, 2004, **43**(4): 379−393

5 Weninger F, Erdogan H, Watanabe S, Vincent E, Le Roux J, Hershey J R, Schuller B. Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR. In: Proceedings of the 12th International Conference on Latent Variable Analysis and Signal Separation. Liberec, Czech Republic: Springer International Publishing, 2015. 91−99

6 Weng C, Yu D, Seltzer M L, Droppo J. Deep neural networks for single-channel multi-talker speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2015, **23**(10): 1670−1679

7 Boll S F. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1979, **27**(2): 113−120

8 Chen J D, Benesty J, Huang Y T, Doclo S. New insights into the noise reduction wiener filter. *IEEE Transactions on Audio, Speech, and Language Processing*, 2006, **14**(4): 1218−1234

9 Loizou P C. *Speech Enhancement: Theory and Practice*. New York: CRC Press, 2007.

10 Liang S, Liu W J, Jiang W. A new Bayesian method incorporating with local correlation for IBM estimation. *IEEE Transactions on Audio, Speech, and Language Processing*, 2013, **21**(3): 476−487

11 Roweis S T. One microphone source separation. In: Proceedings of the 2000 Advances in Neural Information Processing Systems. Cambridge, MA: The MIT Press, 2000. 793−799

12 Ozerov A, Vincent E, Bimbot F. A general flexible framework for the handling of prior information in audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 2012, **20**(4): 1118−1133

13 Reddy A M, Raj B. Soft mask methods for single-channel speaker separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 2007, **15**(6): 1766−1776

14 Mohammadiha N, Smaragdis P, Leijon A. Supervised and unsupervised speech enhancement using nonnegative matrix factorization. *IEEE Transactions on Audio, Speech, and Language Processing*, 2013, **21**(10): 2140−2151

15 Virtanen T. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE Transactions on Audio, Speech, and Language Processing*, 2007, **15**(3): 1066−1074

16 Wang D L, Brown G J. *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Piscataway: IEEE Press, 2006.

17 Wang Y X, Narayanan A, Wang D L. On training targets for supervised speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2014, **22**(12): 1849−1858

18 Xu Y, Du J, Dai L R, Lee C H. An experimental study on speech enhancement based on deep neural networks. *IEEE Signal Processing Letters*, 2014, **21**(1): 65−68

19 Huang P S, Kim M, Hasegawa-Johnson M, Smaragdis P. Deep learning for monaural speech separation. In: Proceedings of the 2014 IEEE International Conference on Acoustics, Speech, and Signal Processing. Florence: IEEE, 2014. 1562−1566

20 Weninger F, Hershey J R, Le Roux J, Schuller B. Discriminatively trained recurrent neural networks for single-channel speech separation. In: Proceedings of the 2014 IEEE Global Conference on Signal and Information Processing. Atlanta, GA: IEEE, 2014. 577−581

21 Wang Y X, Wang D L. A deep neural network for time-domain signal reconstruction. In: Proceedings of the 2015 IEEE International Conference on Acoustics, Speech, and Signal Processing. South Brisbane: IEEE, 2015. 4390−4394

22 Simpson A J, Roma G, Plumbley M D. Deep karaoke: extracting vocals from musical mixtures using a convolutional deep neural network. In: Proceedings of the 12th International Conference on Latent Variable Analysis and Signal Separation. Liberec, Czech Republic: Springer International Publishing, 2015. 429−436

23 Le Roux J, Hershey J R, Weninger F. Deep NMF for speech separation. In: Proceedings of the 2015 IEEE International Conference on Acoustics, Speech, and Signal Processing. South Brisbane: IEEE, 2015. 66−70

24 Gabor D. Theory of communication. Part 1: the analysis of information. *Journal of the Institution of Electrical Engineers — Part III: Radio and Communication Engineering*, 1946, **93**(26): 429−441

25 Patterson R, Nimmo-Smith I, Holdsworth J, Rice P. An efficient auditory filterbank based on the gammatone function. In: Proceedings of the 1987 Speech-Group Meeting of the Institute of Acoustics on Auditory Modelling. RSRE, Malvern, 1987. 2−18

26 Wang Y X, Han K, Wang D L. Exploring monaural features for classification-based speech segregation. *IEEE Transactions on Audio, Speech, and Language Processing*, 2013, **21**(2): 270−279

27 Chen J T, Wang Y X, Wang D L. 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

28 Huang P S, Kim M, Hasegawa-Johnson M, Smaragdis P. Singing-voice separation from monaural recordings using deep recurrent neural networks. In: Proceedings of the 15th International Society for Music Information Retrieval. Taipei, China, 2014.

29 Huang P S, Kim M, Hasegawa-Johnson M, Smaragdis P. Joint optimization of masks and deep recurrent neural networks for monaural source separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2015, **23**(12): 2136−2147

30 Wang Y X, Wang D L. Towards scaling up classification-based speech separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 2013, **21**(7): 1381−1390

31 Han K, Wang D L. A classification based approach to speech segregation. *The Journal of the Acoustical Society of America*, 2012, **132**(5): 3475−3483

32 Han K, Wang D L. Towards generalizing classification based speech separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 2013, **21**(1): 168−177

33 Nie S, Zhang H, Zhang X L, Liu W J. Deep stacking networks with time series for speech separation. In: Proceedings of the 2014 IEEE International Conference on Acoustics, Speech, and Signal Processing. Florence: IEEE, 2014. 6667−6671

34 Zhang H, Zhang X L, Nie S, Gao G L, Liu W J. A pairwise algorithm for pitch estimation and speech separation using deep stacking network. In: Proceedings of the 2015 IEEE International Conference on Acoustics, Speech, and Signal Processing. South Brisbane: IEEE, 2015. 246−250

35 Han K, Wang Y X, Wang D L, Woods W S, Merks I, Zhang T. Learning spectral mapping for speech dereverberation and denoising. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2015, **23**(6): 982−992

36 Nie S, Xue W, Liang S, Zhang X L, Liu W J, Qiao L W, Li J P. Joint optimization of recurrent networks exploiting source auto-regression for source separation. In: Proceedings of the 16th Annual Conference of the International Speech Communication Association. Dresden, Germany, 2015.

37 Dahl G E, Yu D, Deng L, Acero A. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 2012, **20**(1): 30−42

38 Wang Y X. Supervised Speech Separation Using Deep Neural Networks [Ph. D. dissertation], The Ohio State University, USA, 2015.

39 Weninger F, Eyben F, Schuller B. Single-channel speech separation with memory-enhanced recurrent neural networks. In: Proceedings of the 2014 IEEE International Conference on Acoustics, Speech, and Signal Processing. Florence: IEEE, 2014. 3709−3713

40 Hershey J R, Le Roux J, Weninger F. Deep unfolding: model-based inspiration of novel deep architectures. arXiv: 1409.2574, 2014.

41 Hsu C C, Chien J T, Chi T S. Layered nonnegative matrix factorization for speech separation. In: Proceedings of the 16th Annual Conference of the International Speech Communication Association. Dresden, Germany: ICSA, 2015. 628−632

42 Liang S, Liu W J, Jiang W, Xue W. The optimal ratio time-frequency mask for speech separation in terms of the signal-to-noise ratio. *The Journal of the Acoustical Society of America*, 2013, **134**(5): EL452−EL458

43 Liang S, Liu W J, Jiang W, Xue W. The analysis of the simplification from the ideal ratio to binary mask in signal-to-noise ratio sense. *Speech Communication*, 2014, **59**: 22−30

44 Anzalone M C, Calandruccio L, Doherty K A, Carney L H. Determination of the potential benefit of time-frequency gain manipulation. *Ear and Hearing*, 2006, **27**(5): 480−492

45 Brungart D S, Chang P S, Simpson B D, Wang D L. Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation. *The Journal of the Acoustical Society of America*, 2006, **120**(6): 4007−4018

46 Li N, Loizou P C. Factors influencing intelligibility of ideal binary-masked speech: implications for noise reduction. *The Journal of the Acoustical Society of America*, 2008, **123**(3): 1673−1682

47 Wang D L, Kjems U, Pedersen M S, Boldt J B, Lunner T. Speech intelligibility in background noise with ideal binary time-frequency masking. *The Journal of the Acoustical Society of America*, 2009, **125**(4): 2336−2347

48 Hartmann W, Fosler-Lussier E. Investigations into the incorporation of the ideal binary mask in ASR. In: Proceedings of the 2011 IEEE International Conference on Acoustics, Speech, and Signal Processing. Prague: IEEE, 2011. 4804−4807

49 Narayanan A, Wang D L. The role of binary mask patterns in automatic speech recognition in background noise. *The Journal of the Acoustical Society of America*, 2013, **133**(5): 3083−3093

50 Paliwal K, Wójcicki K, Shannon B. The importance of phase in speech enhancement. *Speech Communication*, 2011, **53**(4): 465−494

51 Mowlaee P, Saiedi R, Martin R. Phase estimation for signal reconstruction in single-channel speech separation. In: Proceedings of the 2012 International Conference on Spoken Language Processing. Portland, USA: ISCA, 2012. 1−4

52 Krawczyk M, Gerkmann T. STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2014, **22**(12): 1931−1940

53 Williamson D S, Wang Y X, Wang D L. Complex ratio masking for monaural speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2016, **24**(3): 483−492

54 Mallat S. *A Wavelet Tour of Signal Processing*. Burlington: Academic Press, 1999.

55 Hermansky H, Morgan N. Rasta processing of speech. *IEEE Transactions on Speech and Audio Processing*, 1994, **2**(4): 578−589

56 Shao Y, Jin Z Z, Wang D L, Srinivasan S. An auditory-based feature for robust speech recognition. In: Proceedings of the 2009 IEEE International Conference on Acoustics, Speech, and Signal Processing. Taipei, China: IEEE, 2009. 4625−4628

57 Hu G N, Wang D L. A tandem algorithm for pitch estimation and voiced speech segregation. *IEEE Transactions on Audio, Speech, and Language Processing*, 2010, **18**(8): 2067−2079

58 Han K, Wang D L. An SVM based classification approach to speech separation. In: Proceedings of the 2011 IEEE International Conference on Acoustics, Speech, and Signal Processing. Prague: IEEE, 2011. 4632−4635

59 Narayanan A, Wang D L. Investigation of speech separation as a front-end for noise robust speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2014, **22**(4): 826−835

60 Narayanan A, Wang D L. Improving robustness of deep neural network acoustic models via speech separation and joint adaptive training. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2015, **23**(1): 92−101

61 Wang D L. On ideal binary mask as the computational goal of auditory scene analysis. *Speech Separation by Humans and Machines*. US: Springer, 2005. 181−197

62 Healy E W, Yoho S E, Wang Y X, Wang D L. An algorithm to improve speech recognition in noise for hearing-impaired listeners. *The Journal of the Acoustical Society of America*, 2013, **134**(4): 3029−3038

63 Kjems U, Boldt J B, Pedersen M S, Lunner T, Wang D L. Role of mask pattern in intelligibility of ideal binary-masked noisy speech. *The Journal of the Acoustical Society of America*, 2009, **126**(3): 1415−1426

64 Srinivasan S, Roman N, Wang D L. Binary and ratio time-frequency masks for robust speech recognition. *Speech Communication*, 2006, **48**(11): 1486−1501

65 Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks. *Science*, 2006, **313**(5786): 504−507

66 Sprechmann P, Bruna J, LeCun Y. Audio source separation with discriminative scattering networks. In: Proceedings of the 12th International Conference on Latent Variable Analysis and Signal Separation. Liberec, Czech Republic: Springer International Publishing, 2015. 259−267

刘文举 中国科学院自动化研究所研究员. 主要研究方向为计算听觉场景分析, 语音增强, 语音识别, 声纹识别, 声源定位和声音事件检测. 本文通信作者.
E-mail: lwj@nlpr.ia.ac.cn
(**LIU Wen-Ju** Professor at the Institute of Automation, Chinese Academy of Sciences. His research interest covers computational auditory scene analysis, speech enhancement, speech recognition, speaker recognition, source location, and voice event detection. Corresponding author of this paper.)

聂 帅 中国科学院自动化研究所博士研究生. 2013 年获得内蒙古大学学士学位. 主要研究方向为语音信号处理技术, 深度学习, 语音分离, 计算听觉场景分析.
E-mail: shuai.nie@nlpr.ia.ac.cn
(**NIE Shuai** Ph. D. candidate at the Institute of Automation, Chinese Academy of Sciences. He received his bachelor degree from Inner Mongolia University in 2013. His research interest covers acoustic and speech signal processing, deep learning, speech separation, and computational auditory scene analysis.)

梁 山 中国科学院自动化研究所助理研究员. 2008 年获得西安电子科技大学学士学位, 2014 年获得中国科学院自动化研究所博士学位. 主要研究方向为语音信号处理技术, 语音分离, 计算听觉场景分析, 语音识别.
E-mail: sliang@nlpr.ia.ac.cn
(**LIANG Shan** Assistant professor at the Institute of Automation, Chinese Academy of Sciences. He received his bachelor degree from Xidian University in 2008, and Ph. D. degree from the Institute of Automation, Chinese Academy of Sciences in 2014. His research interest covers acoustic and speech signal processing, speech separation, computational auditory scene analysis, and speech recognition.)

张学良 内蒙古大学副教授. 2003 年获得内蒙古大学学士学位, 2005 年获得哈尔滨工业大学硕士学位, 2010 年获得中国科学院自动化研究所博士学位. 主要研究方向为语音分离, 计算听觉场景分析, 语音信号处理.
E-mail: cszxl@imu.edu.cn
(**ZHANG Xue-Liang** Associate professor at Inner Mongolia University. He received his bachelor degree from Inner Mongolia University in 2003, master degree from Harbin Institute of Technology in 2005, and Ph. D. degree from the Institute of Automation, Chinese Academy of Sciences in 2010, respectively. His research interest covers speech separation, computational auditory scene analysis, and speech signal processing.)