# 鸡尾酒会问题与相关听觉模型的研究现状与展望

黄雅婷 [1, 2] 石 晶 [1, 2] 许家铭 [1] 徐 波 [1, 2, 3]

**摘 要** 近些年, 随着电子设备和人工智能技术的飞速发展, 人机语音交互的重要性日益凸显. 然而, 由于干扰声源的存在, 在鸡尾酒会等复杂开放环境下的语音交互技术远没有达到令人满意的程度. 现阶段, 开发一个具备较强自适应性和鲁棒性的听觉计算系统仍然是一件极具挑战性的任务. 因此, 鸡尾酒会问题的深入探索对智能语音处理领域中的说话人识别、语音识别、关键词唤醒等一系列重要任务都具有非常重要的研究意义和应用价值. 本文综述了鸡尾酒会问题相关听觉模型研究的现状与展望. 在简要介绍了听觉机理的相关研究, 并概括了解决鸡尾酒会问题的多说话人语音分离相关计算模型之后, 本文还讨论了受听觉认知机理启发的听觉注意建模方法, 认为融入声纹记忆和注意选择的听觉模型在复杂的听觉环境下具有更好的适应性. 之后, 本文简单回顾了近期的多说话人语音识别模型. 最后, 本文讨论了目前各类计算模型用于处理鸡尾酒会问题时遇到的困难和挑战, 并对未来的研究方向进行了展望.

**关键词** 鸡尾酒会问题, 听觉模型, 语音分离, 听觉注意, 语音识别

**引用格式** 黄雅婷, 石晶, 许家铭, 徐波. 鸡尾酒会问题与相关听觉模型的研究现状与展望. 自动化学报, 2019, **45**(2): 234−251

**DOI** 10.16383/j.aas.c180674

# Research Advances and Perspectives on the Cocktail Party Problem and Related Auditory Models

HUANG Ya-Ting[1, 2] SHI Jing[1, 2] XU Jia-Ming[1] XU Bo[1, 2, 3]

**Abstract** With the rapid development of electronic devices and artificial intelligence technologies, speech-based human-machine interaction has become increasingly prominent in recent years. However, the performance of these technologies in open complex environments, such as in the cocktail parties, is far from satisfactory. It is still a very challenging task to develop a computational auditory system with strong adaptivity and robustness at present. Therefore, the in-depth exploration of cocktail party problem plays an important role in the tasks of the intellectual speech processing field, such as speaker recognition, speech recognition, keyword spotting and so on. This paper reviews the auditory models related to the cocktail party problem and their developments. We first briefly introduce some relevant hearing research and computational models attacking the multi-speaker speech separation task for solving the cocktail party problem. Then we discuss the auditory attention modeling method inspired by cognitive science. We believe that the auditory model integrated with the memory of voiceprint information and selective attention is more suitable for complex auditory environments. Afterwards, we briefly review current works of multi-speaker speech recognition. Finally, the difficulties and challenges that the current computational models are confronted with are discussed and we give some views on the future research.

**Key words** Cocktail party problem, auditory model, speech separation, auditory attention, speech recognition

**Citation** Huang Ya-Ting, Shi Jing, Xu Jia-Ming, Xu Bo. Research advances and perspectives on the cocktail party problem and related auditory models. *Acta Automatica Sinica*, 2019, **45**(2): 234−251

收稿日期 2018-10-18 录用日期 2019-01-08
Manuscript received October 18, 2018; accepted January 8, 2019
国家自然科学基金 (61602479), 中国科学院战略性先导科技专项 (XDBS01070000), 北京市科技重大专项 (Z181100001518006) 资助
Supported by National Natural Science Foundation of China (61602479), the Strategic Priority Research Program of Chinese Academy of Sciences (XDBS01070000), and the Beijing Brain Science Project (Z181100001518006)
本文责任编委 党建武
Recommended by Associate Editor DANG Jian-Wu
1. 中国科学院自动化研究所 北京 100190 2. 中国科学院大学 北京 100049 3. 中国科学院脑科学与智能技术卓越创新中心 上海 200031
1. Institute of Automation, Chinese Academy of Sciences, Beijing 100190 2. University of Chinese Academy of Sciences, Beijing 100049 3. Center for Excellence in Brain Science and Intelligence Technology, CAS, Shanghai 200031

鸡尾酒会问题 (Cocktail party problem) 最早是由英国认知科学家 Cherry 于 1953 年[1] 在研究选择注意 (Selective attention) 机制时提出的一个著名问题, 该问题尝试探明在受到其他说话人或者噪音干扰的情况下人类理解目标说话人言语这一过程背后的逻辑基础, 从而建模出能够过滤出目标说话人信号的智能机器. 通俗地描述, 鸡尾酒会问题关注人类在复杂听觉环境下的一种听觉选择能力. 在这种情况下, 人可以很容易地将注意力集中在某一个感兴趣的声音刺激上并忽略其他背景声音, 而

计算听觉模型却受噪音影响严重. 如何设计一个能够灵活适应鸡尾酒会环境的听觉模型是计算听觉领域的一个重要问题, 在语音识别, 语音增强, 说话人识别, 语音分离等一系列重要任务上都具有非常重要的研究意义和应用价值. 尤其在近些年, 随着智能设备和便携式计算设备的爆炸式发展, 语音已经成为了人类接入智能计算设备和平台的最重要的入口之一. 基于此, 面对日常生活中最典型和常见的复杂听觉场景, 如何有效地处理鸡尾酒会问题就显得意义重大. 换句话说, 针对鸡尾酒会问题的计算模型, 也即针对复杂听觉场景的建模方法的好坏直接影响着输入信息的解析是否完备, 关键信息是否被有效筛选, 干扰信息是否被忽略, 以致从源头上影响了后续任务能否成功完成, 其重要性自然不言而喻. 如上所述, 智能设备的广泛普及为鸡尾酒会问题带来了前所未有的挑战和需求, 但同时人工智能方法和学科的高速发展也为解决鸡尾酒会问题带来了更好的机遇.

事实上, 面对复杂环境的听觉选择注意能力是人类在进化过程中听觉系统形成的一项惊人天赋[2]. 鸡尾酒会效应的产生机制虽然十分复杂, 但对于人类来说, 在多个声源之间转换注意是一件非常轻松的事, 以至于我们甚至感受不到这个过程的存在. 遗憾的是, 目前, 智能机器却难以取得跟人类一样理想的表现. 不过, 经过半个多世纪的不断探索, 隐藏在鸡尾酒会问题背后的神经机制虽然尚未明朗, 但是相关研究依旧取得了一定的成果. 例如, 研究人员们对于人类听觉通路的形成过程, 听觉信号在神经传递过程中的编码方式等, 已经有了比较清晰的认识. 在另一方面, 对于当今的人工智能方法和建模方式而言, 尤其对于神经网络和深度学习的方法, 借鉴人脑过程中的相关机制来构建类脑的, 脑启发式的模型已经成为一种非常有效的手段. 比如, 卷积神经网络 (Convolutional neural networks, CNN) 的设计过程正是借鉴了人类视觉通路中感受野和层次化纹理感应的相关机制, 有效地构建了类似的计算模型框架, 从而在图像处理领域取得了非常卓越的进步. 类似地, 我们认为, 要真正解决鸡尾酒会问题, 需要从听觉研究取得的相关成果中有所借鉴. 因此, 本文从人类处理鸡尾酒会问题的相关听觉机制出发, 总结了听觉选择过程中的一些关键机制, 并在之后详细对比了目前就鸡尾酒会问题建模的一些计算模型.

本文具体章节安排如下: 第 1 节介绍与鸡尾酒会问题相关的听觉机制; 第 2 节介绍之前就鸡尾酒会问题进行建模的多说话人语音分离计算模型; 第 3 节介绍基于听觉注意的多说话人语音分离计算模型; 第 4 节介绍近期就鸡尾酒会问题进行建模的多说话人语音识别计算模型; 第 5 节就目前研究存在的问题进行总结并对未来的研究方向进行展望.

## 1 听觉机理的相关研究

本节将就至今为止听觉通路的相关神经学和心理学机制进行一个简单的介绍, 以期对鸡尾酒会问题相关的计算模型的建立提供基础的生理学背景知识, 并能够起一定启发作用.

### 1.1 听觉神经编码与听觉通路

人类听觉系统能够在复杂的听觉环境下, 鲁棒地对外界的各种声音进行编码, 加工和处理. 尽管对听觉通路的研究并没有对视觉通路的研究那般透彻, 到目前为止, 相关的研究对听觉通路中的早期皮层下处理过程已较为清楚, 并开始对听觉通路中的后期初级听觉皮层及之后的皮层结构的功能进行深入研究[3]. 听觉通路从耳蜗开始, 通过听神经中的神经元以短电脉冲即动作电位的形式传递信息[4−5], 经过 4 ∼ 7 个核团传输到听觉皮层. 相比视觉通路, 听觉通路经过了数量更多的神经核团的处理. 虽然听觉通路中处理复杂声音的具体神经编码方式尚未明确, 不过已发现了以下三种主要的编码方式: 频率编码 (Rate coding), 时间编码 (Temporal coding) 和群体编码 (Population coding). 频率编码即神经元通过动作电位的发放频率来编码刺激信息. 在理论和实践层面, 无论在神经科学还是计算建模方面, 神经元的发放频率都被广泛使用来描述神经元的活动. 当神经元发放动作电位的时间携带与刺激有关的信息时, 我们称这种编码方式为时间编码. 相比频率编码只考虑一段时间内的脉冲发放频率, 时间编码多了时间这个维度, 比频率编码更为有效. 而群体编码则指一个神经元群组共同编码刺激的编码方式, 例如耳蜗中的毛细胞对声音频谱的编码就属于群体编码. 每个毛细胞对应一条具有一个最佳响应频率的频率响应曲线, 各个频率的毛细胞在耳蜗中按照一定的空间位置形成一个拓扑分布 (Tonotopy). 当某个频率的刺激出现时, 就会激活最佳频率与该刺激相近的一组毛细胞的活动, 因此单个频率是由一组神经元来编码的.

听觉通路中各部分的连接非常复杂. 类比视觉系统的腹侧通路和背侧通路, 一般认为听觉通路中也存在腹侧通路用来处理声音的非空间属性和背侧通路处理声音的空间属性. 但也有研究表明分布式的自适应网络可能比上述的两条并行通路更适合解释听觉认知, 在这种理论下, 脑区之间的反馈连接有助于促进听觉物体选择[6]. 除了串行, 并行和反馈连接之外, 听觉通路中还存在汇合连接, 即某

个区域整合从另外几个区域得到的信息, 例如下丘 (Inferior colliculus, IC); 发散连接, 即某个区域的信息传递到其他几个区域进行处理, 例如内侧膝状体 (Medial geniculate body, MGB); 短路连接, 比如从蜗核 (Cochlear nucleus, CN) 直接连到内侧膝状体[3].

### 1.2 听觉处理与鸡尾酒会问题的相关脑机制

回顾鸡尾酒会问题被提出的场景, 当时英国认知科学家 Cherry 正是在研究人类选择注意机制时阐述了这一著名问题. 在人类进化过程中, 由于大脑中央处理部 (Central processor) 的能力有限, 继而形成了选择注意机制来对需要更详细加工的部分进行进一步加工[7]. 事实上, 人类对复杂听觉环境认知时, 听觉注意 (Auditory attention) 往往起到非常重要的作用. 有实验研究发现, 人类不可能听到或者记住两个同时发生的语音. 相反, 人类却可以精准地从被混合的复杂语音中选择出来其注意到的语音, 以及同时忽略掉其他语音或者噪音等背景音[8]. 以上种种研究表明, 听觉注意在人类处理复杂听觉场景中是非常重要且必不可少的一个机制. 听觉系统处理外界刺激一般可以分为自下而上 (Bottom-up) 的刺激驱动的过程和自上而下 (Top-down) 的任务驱动的过程. 自下而上的处理过程是指从输入的刺激进行处理, 继而完成相应的任务. 自上而下的处理过程是指在高层的抽象概念或信息的指引下完成特定的任务, 其过程通常涉及长期记忆和学习机制. 传统听觉研究认为, 在自下而上的过程中, 在处理较为简单的刺激时, 听觉系统遵从 Old-plus-new 原则, 即信号中的突然改变可以认为是源自单一声源的改变, 而频谱中若只有能量增加则可以认为原声源不变而有新声源出现. 但是自然界中充满了各式各样复杂的声音, 很难出现像传统听觉研究中的单一频率的纯音刺激, 因而 Old-plus-new 原则往往难以解释复杂声音. 近年来, 时间相干性 (Temporal coherence) 的提出较为有效地解释了复杂声音的处理[9]. 时间相干性理论主要基于以下基本假设: 来自同一听觉流 (Auditory stream) 的各个特征通道在时间上的变化是高度相关的, 而来自不同听觉流的各个特征通道在同一时间同时变化的可能性很低, 从而根据时间相干性, 我们可以将各个听觉流分离开来. 尽管听觉注意的参与对于听觉流的分离并非必须, 但是其参与对于听觉流的形成依然有十分深刻的影响. 当新奇刺激呈现的时候, 比如不熟悉的说话人的语音, 由于没有先验知识, 时间相干性在驱使注意绑定属于同一个声源的特征时起重要作用. 时间相干性在绑定跨模态特征方面也起到一定作用.

另外, 听觉系统对新奇的刺激高度敏感. 刺激特异性适应 (Stimulus-specific adaptation, SSA) 是指听觉上行通路中神经细胞对普遍或者重复性的声音的响应有所降低, 而对新奇, 稀有的声音维持高度敏感性的一种现象[10]. 刺激特异性现象跟大脑中用以维持和更新听觉表示的基于规律 (Regularity) 的改变机制有关, 并涉及感知记忆的加工, 即涉及自上而下的先验知识. 这里听觉规律是指声音序列中的重复可预测的模式. 研究表明 SSA 现象是由以下两种因素共同决定的: “局部效应” 和 “局部加整体”. 局部效应是指对当前刺激的响应仅由过去短期的刺激历史决定. 局部加整体附加考虑了每个刺激的整体出现概率[11]. Winkler 等认为基于规律的表示具有预测性, 是感知物体 – 即感知的基本单元[12]. 同时, 听觉系统还能在嘈杂的环境中根据上下文信息补全被噪音掩盖的缺失的音素或音节, 这种现象称为音素恢复 (Phonemic restoration). 最近有研究显示, 听觉中枢的一个区域能够实时补充和恢复缺失的音素或音节, 而且大脑中更高级的认知区域的神经活动能够在噪声开始之前就预测被试 (心理学实验或心理测验中接受实验或测试的对象) 要报告的单词[13]. 这些研究给预测加工 (Predictive processing) 理论提供了有力支持, 表明预测加工在感知中起到至关重要的作用[14−16]. 预测加工是近年来认知神经科学中愈发受到关注的前沿理论, 认为大脑是一个具有预测能力的层次化结构, 持续地对未来的内部状态进行预测, 目的在于最小化内部状态和外部感知输入的预测误差以对未来的刺激进行近似. 当刺激以一定的非随机的方式呈现出来的时候, 大脑会将外在刺激与已存储的规律进行匹配, 并根据预测误差对存储的规律进行一定的调整; 即使对于随机刺激或者新奇刺激, 大脑依然采取预测加工的策略来处理[17]. 而预测加工机制和选择注意机制的联系, 在神经科学中甚至提出了两种看似会得到完全相反结果的理论. Pearce-Hall 理论认为由于人脑处理资源有限, 预测加工是误差驱动的, 因而为了最大限度利用有限的计算资源, 应该将更多的选择注意关注到预测误差较大的刺激[18]. 而 Mackintosh-Kruschke 理论则认为选择注意是在特征层次的, 应该更多关注到那些能得到更好的预测的特征上[19−20]. 事实上, 这两种理论关注的层次并不相同, 前者是在刺激层次上来进行讨论的, 而后者是在特征层次上进行讨论的, 因此可以视为互补[21].

此外, 各个模态之间的信息处理不是相互独立的. 多感知整合 (Multisensory integration) 通过组织不同模态的输入, 在多模态脑区 (Heteromodal brain areas) 中进行处理, 得到噪音更少的, 更鲁棒

的目标信号, 从而使背景噪音和目标之间的分离, 连续时间之间的分割更加容易[22]. 研究表明, 视觉输入对其他模态的信息处理具有非常强的影响[23]. 其中, 麦格克效应 (McGurk effect) 显示嘴唇及其周围区域的动作对言语处理起到关键作用. 将一个音节 “ga” 在配合发作 “ba” 的唇部动作的视频呈现给被试看, 被试称听到的音节既不是 “ga” 也不是 “ba”, 而是 “da”[24]. 而且嘴唇和下颚的动作跟言语的声学包络相关, 通过观看说话人的正在说话的脸, 能够增强听觉皮层对言语的跟踪和对目标说话人的注意选择[25]. 关于多感知整合发生在哪个阶段目前未有定论, 有三种可能: 一是早期整合 (Early integration), 在相当早的处理阶段就进行融合, 是一个前注意 (Pre-attentive) 加工过程, 即感知驱动注意[26−28], 框架图如图 1 (a); 二是晚期整合 (Late integration), 在整合过程中需要注意的参与[29−30], 框架图如图 1 (b); 三是并行整合 (Parallel integration), 即发生早期整合还是晚期整合取决于手头上任务可获得的资源[31], 框架图如图 1 (c).

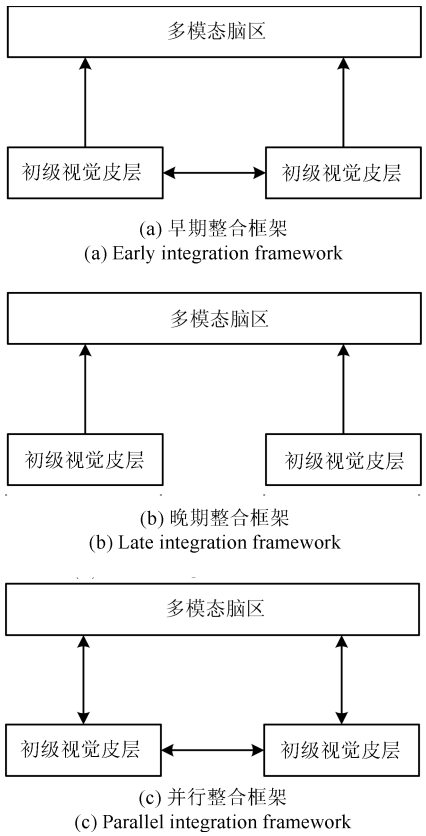



图 1 多感知整合框架[22]

Fig. 1 Multisensory integration framework[22]

## 2 多说话人语音分离计算模型回顾

语音分离是解决鸡尾酒会问题的第一步. 在过去的数十年, 研究人员尝试用各种方法解决多说话人语音分离问题, 但是机器在语音分离上的表现与人类相比, 效果不甚理想. 根据麦克风的个数, 语音分离算法可以分为单通道语音分离算法和多通道语音分离算法. 多通道语音分离相比单通道语音分离多了空间信息. 常见的多通道语音分离算法有基于麦克风阵列的波束成形 (Beamforming) 算法[32−33] 和多通道盲信号分离 (Blind signal separation, BSS) 算法[34]. 基于麦克风阵列的波束成形算法通过麦克风阵列的恰当配置进行空间滤波, 根据空间位置来削弱干扰信号而增强来自期望声源的各通道信号的加和, 通常可以分为可控波束成形技术和自适应波束成形技术. 可控波束成形技术的滤波器参数的确定依赖于声源信号的频谱特性的先验知识, 主要有延迟−累加 (Delay-and-sum) 波束成形算法和滤波−累加 (Filter-and-sum) 波束成形算法. 自适应波束成形技术的滤波器参数的确定则基于期望信号和干扰信号的统计属性, 通过优化一定的准则来确定滤波器参数, 常用的准则有最大信噪比 (Maximum signal-to-noise ratio, MSNR) 准则, 最小均方差 (Minimum mean-squared error, MMSE) 准则, 最小方差无失真响应 (Minimum variance distortionless response, MVDR) 准则和线性约束最小方差 (Linear constraint minimum variance, LCMV) 准则. 由于波束成形算法利用空间信息来分离语音, 因此当目标语音和噪声源位置相近的时候, 算法就会失效. 除此以外, 当声学环境的混响时间很大或者麦克风数少于信号源数的时候, 表现也会大大下降. 而多通道盲信号分离算法主要分为线性混合模型和卷积混合模型. 通过多个麦克风获取多个信号源也就是声源 $S(t)$ 混合后的信号 $X(t)$, 其混合过程 $H$ 未知, 见式 (1).

$$X(t) = H(S(t)) \tag{1}$$

通过假设信号源之间统计上相互独立, 可以克服对信号源和混合过程缺乏先验知识这个问题. 因此可以使用独立成分分析 (Independent component analysis, ICA) 来对分离过程 $G$ 进行建模获得重建信号, 见式 (2). 当麦克风数少于声源数时, 传统盲信号分离就会变得困难.

$$\hat{S}(t) = G(X(t)) \tag{2}$$

一般来说, 单通道语音分离比多通道语音分离更具挑战性. 在后文中, 除非明确指明多通道语音分离算法, 约定语音分离指的是单通道语音分离. 下面根据输入的不同, 将语音分离算法分为只利用听觉信息的语音分离算法和引入视觉信息的语音分离算法进

行介绍.

### 2.1 只运用听觉信息的多说话人语音分离算法

#### 2.1.1 传统语音分离算法

根据算法原理的不同形式, 传统的语音分离算法可以分为基于信号处理的算法, 基于分解的算法和基于规则的算法[35]. 基于信号处理的方法从信号处理的角度估计噪音的功率谱或者理想维纳滤波器, 通常用在语音增强中, 比如谱减法[36], 维纳滤波器[37−38]. 该类算法假定语音服从一定的分布, 而噪音是平稳或慢变的. 但是真实环境中的语音很难满足这些假设条件, 此时这类算法就会失效[39]. 基于分解的算法的基本假设是声音的频谱具有低秩结构, 因此可以用一个数量比较小的基来进行表示, 如式 (3) 所示,

$$X = WH \tag{3}$$

这里, 频谱 $X \in \mathbf{R}^{F\times T}$ 被分解成基矩阵 $W \in \mathbf{R}^{F\times K}$ 和激活矩阵 $H \in \mathbf{R}^{K\times T}$ 的矩阵乘积, 其中 $K$ 是超参数, 通常比 $F$ 和 $T$ 小很多. 在语音处理中, 一种最广泛的做法是令 $W$ 和 $H$ 非负, 从而得到非负矩阵分解 (Non-negative matrix factorization, NMF)[40]. NMF 能够挖掘到语音或噪音中非负数据的基本谱模式[39]. 在 NMF 的基础上引入其他约束, 则可以得到 NMF 的不同变种. 稀疏 NMF[41−43] 对 NMF 加入稀疏约束, 来提高分解的鲁棒性. 卷积 NMF 则将频谱 $X$ 分解成矩阵卷积的形式来对时间依赖进行建模, 此时基矩阵随时间变化, 每个时刻的基矩阵编码了该时刻的频谱, 激活矩阵也对应变化. RNMF (Robust non-negative matrix factorization) 则将 NMF 与 RPCA (Robust principle component analysis) 结合起来, 将频谱分解成重建信息与低秩残差, 其中重建信息代表语音, 而低秩残差代表噪音[44−45]. 但是基于分解的方法属于浅层模型, 假定频谱可以表示成基的线性组合, 而声音本身却是高度非线性的, 因此这种假设过于简单, 不能对声音的长时依赖等建模. 为了挖掘语音中丰富的时空结构和非线性关系, 后续有工作将 NMF 拓展成深层结构, 提出 DNMF[46], L-NMF[47] 等结构, 从而获得了性能的巨大提升. 另一方面, 从计算角度来看, 基于分解的方法计算代价昂贵, 学习到的表示所需要的参数会随着数据变化的增长而呈现线性增长, 一般采用迭代算法求解, 计算复杂度高, 难以满足实时应用要求[35]. 基于规则的算法, 也指计算听觉场景分析 (Computational auditory scene analysis, CASA), 旨在建立像人类一样处理鸡尾酒会问题的智能系统用以分离混合的声音[48]. 这类系统一般根据听觉场景分析研究中发现的一些规则或机制来对鸡尾酒会问题进行建模.

CASA 系统一般分为两个阶段: 特征提取和特征绑定的阶段, 分组之间的竞争阶段[49]. 特征提取阶段会经过一个听觉外周模型提取出声音的特征属性, 然后根据这些特征属性来进行分组得到不同的听觉流 (Auditory stream). 常用来分组的声音属性, 也就是分组线索 (Grouping cues) 有声音的开始和/或结束时间, 谐波结构, 基音, 音色和位置等[6]. 根据建模遵循的规则不同, CASA 模型主要可以分为三种: 基于贝叶斯推断规则的模型, 基于神经计算的模型和基于时间相干性的模型[49]. 这几类模型主要在处理分组之间的竞争和对预测机制的建模上有所不同. 基于贝叶斯推断的模型中, 预测与分组之间的竞争密切相关, 通过调整各分组之间的先验概率来实现竞争机制, 同时用先验概率来得到预测结果; 而分组的数量可固定也可不作限制. Barniv 等在其模型中不对分组的数量也就是分类数作限制, 当有输入的条件似然低于一个阈值的时候, 定义一个新类, 此时原有类的先验概率会降低并和条件似然成比例, 但是这个输入不一定会被分到新的类, 只有当更多的输入映射到新的类对应的特征区域后, 新类的先验概率有所增加, 才会参与分类[50]. 基于神经计算的模型则以神经元为单位来表示听觉流, 听觉流之间的竞争则由神经元之间的抑制连接来实现. 这类方法主要基于神经科学中发现的神经振荡 (Neural oscillation)[51] 机制, 采用振荡脉冲网络来对分离过程进行建模. Wang 等在其两层振荡脉冲神经网络模型中采用局部兴奋 (Local excitatory) 和全局抑制 (Global inhibitory) 的动态机制, 根据振荡子之间的同步性来分离不同的听觉流[52]. 基于神经计算的模型并不像基于贝叶斯推断的模型一样本身就对预测机制进行建模, 不过 Mill 等在他们的模型中额外加入了预测机制, 即对接下来的声音预测误差的表示进行修正和通过相同声音的听觉流之间的抑制来促进其竞争[53]. 基于时间相干性的模型则是根据上一节提到的时间相干性来对分离过程进行建模, 注意和记忆可加入到模型中, 也可以额外加入预测机制[54−56]. 对比这三类模型, 基于贝叶斯的模型本身就具有预测机制, 而神经计算模型和基于时间相干性的模型则不具有这种特性. 不过基于贝叶斯推断的模型的竞争机制基于对先验概率之间的调整, 相对其他模型来讲比较抽象, 而神经计算模型则更为直观, 而且更容易拓展. 基于贝叶斯推断的模型和神经计算模型假设, 特征提取和特征绑定的过程和分组之间的竞争过程是相互独立的, 但实际上两者相互影响; 相比之下, 基于时间相干性的模型则直接提供了一个一步解决的方案, 用时间相干

性来建立特征绑定和听觉流形成. 不过基于时间相干性的模型并没有提供捕捉声音中的高阶规律的途径, 而已有研究表明高阶规律对听觉流的分离有帮助作用. 可以注意到这三类模型适用于解决听觉处理的不同问题, 基于贝叶斯推理的模型提供了使用先验知识的预测框架; 神经计算模型里的竞争机制更为直观; 基于时间相干性的的模型则对特征绑定和听觉物体形成问题提供了较好的解决方案[49]. 但是 CASA 模型也有一些缺陷, 由于这些模型基本上是基于听觉场景分析研究得到的一些规则来进行建模, 而听觉场景分析的研究一般采用较为简单的刺激, 得到的规则在复杂听觉环境下并不一定适用, 大多数 CASA 模型的任务目标是为了重现听觉场景分析中的实验结果, 很少有能应用到实际中的大规模数据集上的模型; 而且, 大部分 CASA 模型严重依赖于分组线索, 尤其是基音提取的准确性, 而这在复杂听觉环境下又难以保证, 因此语音分离效果并不理想[35].

### 2.1.2 基于深度学习的语音分离算法

近年来, 随着计算成本的降低与计算速度的提高, 语音分离任务的表现越来越得益于数据驱动型方法, 尤其是深度学习方法. CASA 模型的一个主要目标是学习一个理想二值掩蔽 (Ideal binary mask, IBM), 来决定频谱中的目标信号在哪些时频单元 (Time-frequency units) 中做主导, Wang 等将时频单元级别的特征作为深度神经网络 (Deep neural networks, DNN) 的输入, 将学习到的特征和原始特征拼接在一起作为输入, 利用线性 SVM 进行二分类并得到 IBM, 在一定程度上缓解了传统语音分离问题难以在大数据集上进行训练的问题[57]. 一方面, 时频单元级别的特征能够关注到更加微小的细节, 却缺乏对语音的全局性和整体性的描述, 无法获得语音的时空结构和时序相关性[39]; 另一方面, IBM 的估计若出错, 则会导致信息丢失过大[58]. 在后续工作中, Narayanan 等将相邻子带的输出作为最后的分类器的输入, 将理想比值掩蔽 (Ideal ratio mask, IRM) 作为 DNN 的训练目标, 做语音增强任务[59]. 上述两项工作需要对每个滤波器组通道 (Filterbank channel) 训练一个神经网络, 当滤波器组通道数太大的时候, 训练如此多的神经网络非常不实际, 难以达到拓展性要求. 为解决这个问题, Huang 等提出用一个神经网络直接同时训练所有特征通道和掩蔽函数应用到两个说话人的语音分离任务上. 在每一个时刻, 将落在以该时刻为中心的时间窗口内的特征拼接起来作为深度神经网络或递归神经网络 (Recurrent neural networks, RNN) 的输入, 学习得到两个声源的频谱, 并在神经网络之后额外加入一个掩蔽层将 IRM 整合到网络中, 从而联合地训练优化整个网络, 见式 (4) 和式 (5), 其中 $X_t$ 表示在 $t$ 时刻混合语音的频谱, $\hat{\boldsymbol{y}}_{1_t}$ 和 $\hat{\boldsymbol{y}}_{2_t}$ 表示神经网络的预测, $\widetilde{\boldsymbol{y}}_{1_t}$ 和 $\widetilde{\boldsymbol{y}}_{2_t}$ 表示最后经过掩蔽层得到的输出[60], $\odot$ 为逐个元素依次相乘 (Element-wise multiplication), 系统框架如图 2.

$$\widetilde{\boldsymbol{y}}_{1_t} = \frac{|\hat{\boldsymbol{y}}_{1_t}|}{|\hat{\boldsymbol{y}}_{1_t}| + |\hat{\boldsymbol{y}}_{2_t}|} \odot X_t \tag{4}$$

$$\widetilde{\boldsymbol{y}}_{2_t} = \frac{|\hat{\boldsymbol{y}}_{2_t}|}{|\hat{\boldsymbol{y}}_{1_t}| + |\hat{\boldsymbol{y}}_{2_t}|} \odot X_t \tag{5}$$

文献 [60] 还提出一个区分性的训练目标使得在考虑源信号与预测信号的相似性的同时, 还考虑预测信号与其他源信号的相似性, 见式 (6), 其中 $\gamma$ 是超参数.

$$\begin{aligned} &\|\widetilde{\boldsymbol{y}}_{1_t} - \boldsymbol{y}_{1_t}\|_2^2 - \gamma\|\widetilde{\boldsymbol{y}}_{1_t} - \boldsymbol{y}_{2_t}\|_2^2 + \\ &\quad \|\widetilde{\boldsymbol{y}}_{2_t} - \boldsymbol{y}_{2_t}\|_2^2 - \gamma\|\widetilde{\boldsymbol{y}}_{2_t} - \boldsymbol{y}_{1_t}\|_2^2 \end{aligned} \tag{6}$$

在其后续工作中, 文献 [61] 进一步拓展该框架为一个应用更为广泛的通用框架, 将深度递归神经网络 (Deep recurrent neural networks, DRNN) 和堆叠递归神经网络 (Stacked RNN) 应用到模型建模中, 并通过实验结果验证了额外的掩蔽层和区分性训练的有效性. Du 等则应用深度神经网络模型作为回归模型, 利用其高度非线性特性对混合语音与纯净语音之间的映射关系进行建模[62−63], 作者将对数功率谱作为 DNN 的输入, 用 DNN 直接学习输出

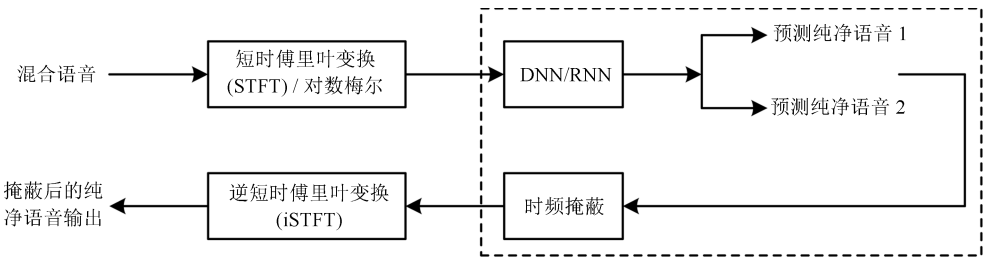



图 2 Huang 等提出的基于深度学习的语音分离系统的结构[60]

Fig. 2 The structure of the proposed deep learning based speech separation system by Huang et al.[60]

目标说话人或者目标说话人和干扰说话人的对数功率谱, 再对得到的信号进行重建. 在其后续工作中, 作者将这个模型加以拓展, 提出 SND-DNN (Signal-noise-dependent DNN) 系统利用一个正信噪比的 DNN 和一个负信噪比的 DNN 克服单个 DNN 难以学习不同信噪比下混合语音的变化特性的问题, 并联合说话人识别和语音分离采取多遍算法获得了比单个 DNN 更好的分离表现[64]. Weninger 等将信号估计 (Signal approximation, SA) 作为目标, 并将长短时记忆网络 (Long-short term memory networks, LSTM) 应用到语音分离问题中, 其实验结果显示 LSTM 比 DNN 在分离性能上更优[65]. 其中, SA 通过训练一个比值掩蔽估计器 (Ratio mask estimator) 来最小化纯净语音和预测语音之间的频谱幅度之间的差距, 见式 (7) 和式 (8), 其中 $X(t,f)$ 和 $\hat{S}(t,f)$ 分别是混合语音和预测的语音的频谱幅度, SMM (Spectral magnitude mask) 是傅里叶变换掩蔽, $RM$ 是对 $SMM$ 的一个估计. 信号近似可以看作是比值掩蔽方法和频谱映射方法的结合[66].

$$SMM(t,f)=\frac{|\hat{S}(t,f)|}{|X(t,f)|} \tag{7}$$

$$SA(t,f)=[RM(t,f)|X(t,f)|-|\hat{S}(t,f)|]^2 \tag{8}$$

总的来说, 早期利用基于深度学习的语音分离算法在模型网络架构上较为简单, 根据训练目标主要可以分为三类: 基于时频掩蔽的算法[57−60], 基于频谱映射的算法[62−64] 和基于信号近似的算法[65]. 基于时频掩蔽的算法相比基于频谱映射的算法, 能够更好地发现目标和干扰说话人之间的互信息, 运用在数据驱动的深度学习算法中能够更好地利用训练集中大量的训练数据. 而基于频谱映射的算法相比于基于时频掩蔽的算法, 则对数据中的 SNR 变化更不敏感, 在低 SNR 的时候表现比基于时频掩蔽的算法表现会更好[67]. 后续的工作多数在这三类模型基础上进行扩展, 对网络架构及训练方法等进行改进.

上述基于深度学习的模型在利用上下文信息的时候时间分辨率固定, 而难以发现尺度较大的时序依赖性并进行建模, 引入不同时间分辨率的表示或多尺度上下文时间窗口的模型有助于整合更长时间内的上下文信息. 对比不同分辨率的表示或基于单一尺度上下文时间窗口的模型, 在语音分离任务上更胜一筹[67−68]. Sprechmann 等[68] 提出每层产生不同时间分辨率的特征图的 Wavelet pyramid scattering transform 网络, 并将学习到的多时间分辨率特征作为深度神经网络, 卷积神经网络的输入, 实验结果显示使用了多时间分辨率的小波特征作为输入的模型在语音分离各项指标 SDR, SIR 和 SAR[69] 上表现远超使用单一时间分辨率的短时傅里叶变换表示作为输入的模型[68]. Zhang 等利用集成学习的思想提出 Multi-context networks, 对有不同尺度的上下文窗口时间长度的 DNN 的输出作平均 (Multi-context averaging, MCA) 或者堆栈 (Multi-context stacking, MCS), 其中 MCS 模型的模块可以是基于时频掩蔽的模型, 基于频谱映射的模型和基于信号近似的模型; 实验结果显示 Multi-context networks 比单一固定上下文窗口时间长度的 DNN 在语音分离任务上效果更好[67]. 深度神经网络模型训练学习对特定说话人的掩蔽函数或者频谱映射常常面临两个困难, 即排列问题 (Permutation problem) 和输出维度不匹配问题 (Output dimension mismatch problem). 前者通常是由于训练样本的目标标签有序, 而混合语音中各个源的顺序却是顺序无关而导致. 后者一般源自大多数模型都采用的固定源的数目的设置, 导致一些模型不具备适应混合语音中源数目可变特性的灵活性[70]. Hershey 等将深度神经网络模型和谱聚类结合起来, 提出深度聚类 (Deep clustering, DC) 算法来解决这两个问题[71−72]. DC 算法提出一个目标函数使得 DNN 学习到一种单位嵌入表示, 使得同一个源信号占主导地位的时频单元之间距离最小, 而不同源信号占主导地位的时频单元之间距离最大, 用这样的嵌入表示得到的目标函数具有低秩的特性, 从而在实现的时候高效地计算出矩阵的导数, 降低谱聚类的计算复杂度, 同时获得良好的聚类效果. 设 $V=f_\theta(x)\in\mathbf{R}^{N\times D}$ 是通过参数为 $\theta$ 的深度神经网络学习到的 $D$ 维嵌入表示且 $|\boldsymbol{v}_i|^2=1$, 则 $A=VV^{\mathrm{T}}$ 可以用来表示一个估计的 $N\times N$ 的亲和度矩阵. $Y=\{y_{i,c}\}\in\mathbf{R}^{N\times C}$ 是一个指示矩阵, 将每个元素 $i$, 在语音分离的场合 $i$ 指时频单元的索引, 映射到 $C$ 个聚类之中: 即 $y_{i,c}=1$ 表示元素 $i$ 属于聚类 $c$, 因此 $(YY^{\mathrm{T}})_{i,j}=1$ 表示元素 $i$ 和元素 $j$ 同属于一个聚类 $c$, $(YY^{\mathrm{T}})_{i,j}=0$ 表示元素 $i$ 和元素 $j$ 属于不同的聚类, 则 $A^*=YY^{\mathrm{T}}$ 可以表示一个真实的 $N\times N$ 的亲和度矩阵. 因此, DC 定义了一个目标函数, 来使得估计的亲和度矩阵尽可能接近真实的亲和度矩阵, 目标函数见式 (9), 其中 $\|\cdot\|_{\mathrm{F}}^2$ 是 Frobenius 范数.

$$C=\|A-A^*\|_{\mathrm{F}}^2=\|VV^{\mathrm{T}}-YY^{\mathrm{T}}\|_{\mathrm{F}}^2=\sum_{i,j:\boldsymbol{y}_i=\boldsymbol{y}_j}(\langle\boldsymbol{v}_i,\boldsymbol{v}_j\rangle-1)^2+\sum_{i,j:\boldsymbol{y}_i\neq\boldsymbol{y}_j}\langle\boldsymbol{v}_i,\boldsymbol{v}_j\rangle^2 \tag{9}$$

DC 有较好的泛化能力, 直接将只用两个说话人混合语音进行训练得到的模型应用到分离三个说话人混合语音的任务上, 依旧能够获得较好的分离表

现. 但由于 DC 优化的目标函数是映射到嵌入空间的源的亲和度矩阵而非信号本身, 后续需要另外用聚类算法来进行聚类, 因而不是一个端到端的系统. Yu 等提出帧级别的具有排列不变性的训练方法 (Permutation invariant training, PIT) 来解决排列问题[73], 具体框架如图 3. PIT 方法的关键在于误差回传的时候计算预测输出序列与标注序列各种排列的均方差, 并选择最小均方差用于优化参数. 在后续工作中, 研究者提出语料级别的具有排列不变性的训练方法 (Utterance-level permutation invariant training, uPIT), 解决了 PIT 方法中的说话人跟踪问题 (Speaker tracing problem)[74]. Chen 等[70] 根据人类听觉认知研究中的感知磁效应 (Perceptual magnet effect)[75] 提出深度吸引子网络 (Deep attractor network, DANet), 从而做到端到端训练. 和 DC 类似, DANet 在训练阶段用训练神经网络将语音频谱映射到一个 $D$ 维嵌入空间, 不同的是之后 DANet 会在嵌入空间内根据时频单元的嵌入表示生成各个源的吸引子, 之后通过每个时频单元与每个吸引子的相似性来估计每个源的掩蔽, 见式 (10), 其中 $A$ 是吸引子矩阵, $V$ 代表嵌入空间, $M$ 代表掩蔽.

$$M_{f,t,c} = \frac{1}{1+\exp(\sum_d A_{c,d} \times V_{f,t,d})} \tag{10}$$

在测试阶段, 可以用两种策略来估计吸引子, 第一种是用 $K$-means 算法对时频单元进行估计得到吸引子, 第二种是根据吸引子在嵌入空间的位置相对稳定使用固定的吸引子. DANet 对比 DC, 通过生成吸引子有效地将与源相关的信息整合进来; 当吸引子矩阵变成自由参数的时候, 掩蔽没有有关源的信息.

### 2.2 引入视觉信息的多说话人视听觉语音分离算法

前面提到的模型只运用了听觉信息本身作为输入来源. 然而, 只使用听觉信息的模型在分离相似声音的时候, 比如相同性别的说话人的声音时, 面临困难[76]. 在实际生活中, 人类在进行听觉选择的同时, 通常也会接受其他形式的信息来源. 其中, 视觉信息在处理鸡尾酒会问题中也起到了非常明显的促进作用. 基于这种认识, 近年来, 研究人员开始将视觉信息作为额外的输入信息引入到语音分离和鸡尾酒会问题的建模当中. 根据视觉信息和听觉信息之间具有高度相关性的观测, 早期的研究一般寻找与声学特征高度匹配的视觉特征集合作为语音分离的辅助信息, 比如提取嘴唇及其周围的区域与唇部运动相关的视觉信息, 来区分噪音环境下的静音片段和言语片段[77], 为音频的频谱提供估计信息[78−79]. 另一种思路则是通过一些统计模型, 比

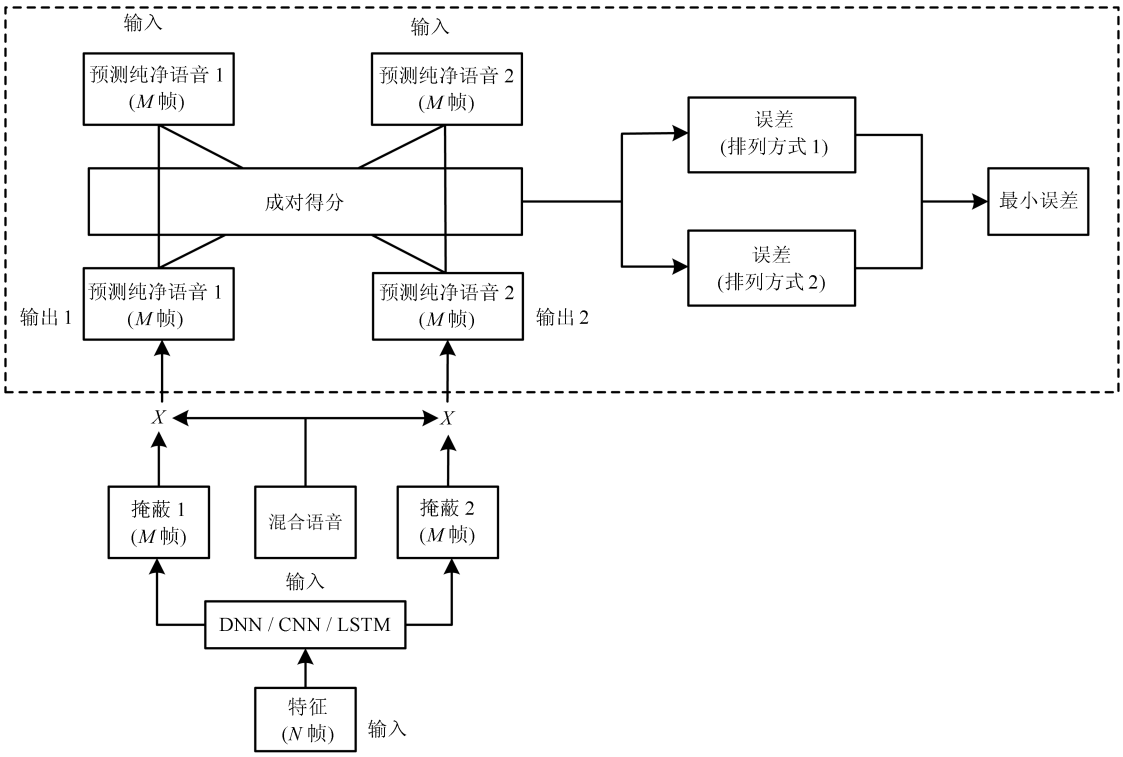



图 3 Yu 等提出的基于排列不变性训练方法的双说话人语音分离系统的结构[73]

Fig. 3 The structure of the proposed PIT-based two-speaker speech separation system by Yu et al.[73]

如隐马尔科夫模型或者高斯混合模型对视听觉信息的分布进行建模, 对视听觉信息的高度非线性相关性进行建模, 通过最大化分离的语音和视频信息之间的相关性[80], 或者采用传统的盲信号分离算法[81] 来得到最后的分离结果. Barzelay 等提出了一个匹配准则, 奖励视听觉信息之间的时间相干性而惩罚不匹配, 寻找各个模态的显著特征之间的对应关系[82]. Casanovas 等提出 BAVSS (Blind audio-visual source separation) 系统, 其中根据视频中检测到的源在音频中存在有且仅有一个声源与其对应的假设和各声源不太可能每时每刻都混合的假设, 用视觉信息确定说话人数目和各自说话的时间片段[83]. 但是, 上述这些工作一般只能在孤立语音或者小数据集上工作. 近几年, 深度神经网络模型的引入使得视听觉语音分离能够在更大规模的数据集上工作. Hou 等提出 AVDCNN (Audio-visual deep CNN) 模型, 分别利用卷积神经网络提取混合语音的信息和唇部图片的信息并将其融合, 在多任务学习的学习目标和编码器–解码器框架下, 恢复出增强语音和重建的唇部图片[84]. 受启于机器唇读的研究进展, Ephrat 等用 Vid2Speech 网络[85] 将已知说话人无声视频作为输入生成音频的频谱, 在分离两个说话人的情况下, 根据生成的两个说话人的频谱强弱, 生成 IBM 或者 IRM 作用于混合语音的频谱得到分离的语音[76]. 与 Hou 的工作类似, Gabbay 等提出相似的基于编码器–解码器的深度学习模型来解决视觉语音增强, 不同的是后者没有恢复输入的唇部图片, 而是在训练集中通过添加同一个说话人的声音作为噪音, 促使网络利用视觉信息[86]. 近期, 同时利用视觉信息和听觉信息来处理鸡尾酒会问题的方法得到了广泛的关注. 其中最主要的一种方法是利用语音和视觉信息的对应性, 完成自监督训练的过程, 以达到利用视觉信息作为刺激来完成语音分离的任务. 例如, Owens 等通过自监督学习[87] 利用视频中视觉信息和听觉信息本身的对齐性, 用神经网络学习视觉和听觉特征是否在时间上对齐, 在较早的阶段就混合视觉和听觉信息而得到不同时间分辨率的多模态整合特征, 用在 On/Off-screen 语音分离任务上[88]. Ephrat 等构建了大规模视听觉数据集 (Audio-visual dataset), 提出基于神经网络的 AVSpeech 模型, 在语音分离任务中利用外部视频分析工具的人脸识别功能引入人脸信息作为刺激, 训练得到了一个说话人无关的模型, 并在真实场景下取得了较好的效果[89]. 最近还有一些鸡尾酒会问题的延伸工作, 对视频中的发声物体的声音进行分离和定位[90−92].

### 2.3 多说话人语音分离算法总结与对比

对第 2.1 节和第 2.2 节的语音分离算法进行一个简单的总结和对比, 如表 1. 基于信号处理的算法, 基于分解的算法和基于规则的算法往往只能在较小规模的数据集上工作, 且难以处理开放数据集的情况. 随着数据的不断积累和计算设备性能的大幅进步, 处理鸡尾酒会问题的模型已经逐渐从基于信号处理, 分解和规则的方法转变成为通过数据驱动形式进行学习的方法. 以深度学习的一系列方法为代表, 当前模型可以在大规模数据集上进行训练, 从而在给定的条件下得到比较好的效果. 然而, 当前对鸡尾酒会问题建模的深度学习方法虽然充分利用了大规模数据集带来的优势, 却也一定程度上过分依赖模型本身的优异性能, 从而忽略了从人类听觉回路中进行借鉴, 造成了可解释性较差, 适用情况较局限等一系列问题. 值得注意的是, 近期出现了一批语音分离的工作, 将人类在鸡尾酒会环境中进行听觉选择的部分机制, 集成到现有的深度学习方法当中来, 从而解决了一些之前工作中存在的问题, 获得了更好的可解释性和比较优秀的性能. 本文将在第 3 节介绍这一类新方法的代表工作.

## 3 基于听觉注意的语音分离计算模型

回顾第 1 节, 我们知道听觉注意在人类处理复杂听觉场景时是非常重要且必不可少的一个机制, 同样, 对于鸡尾酒会问题的语音分离计算模型而言, 听觉注意也应该得到关注. 但从第 2 节回顾的模型可以发现, 现有模型大多数只有自下而上的推断过程, 也就是说, 各类模型往往对复杂的听觉信号进行直接处理, 通过数据驱动的方式进行大量学习, 分离出可能出现的多条语音通道, 而忽略了自上而下的听觉注意过程. 认知心理学研究表明, 自上而下的听觉注意过程有利于更好地利用先验知识, 使人在鸡尾酒会环境中的表现更加高效而鲁棒. 具体而言, Bregman[7] 和 Ciocca[93] 等曾指出, 除了声音在环境当中的物理属性, 听者也会探索他们近期或者长期经验中已经学习到的知识来更好地处理复杂的听觉场景. 事实上, 这种学习到的经验或者说概念中就包含多种不同来源. 例如, 其可能来自于听者对于各类声源的统计特性的熟知, 可能源于对于某个特定声源的短期或者长时记忆, 甚至是能够帮助听者更好地关注目标声源从而忽略其余背景干扰的这种注意状态. 从声源信号处理的角度来看, 这一类自上而下的过程相当于对于可能的最优解施加了一个限定范围, 从而减少了许多无谓的重复处理的过程, 继而在解决鸡尾酒会问题中起到了非常重要的作用[94].

基于以上认识, Xu 等首次将自上而下的任务驱动的听觉注意过程和自下而上的刺激驱动的推断过程整合到一个统一的框架而提出 ASAM (Auditory selection framework with attention and memory)[95], 具体框架如图 4. 在 ASAM 中, 模型设置了一个长期记忆单元, 并在处理过程中对该记忆的各个元素进行更新和提取的操作. 该长期记忆单元类似人脑记忆模型中的长时记忆 (Long-term memory, LTM) 模块, 在整个模型中起到了非常重要的作用. 具体来说, 在模型当中, 长期记忆单元由多个槽组成, 每个槽用以存放并更新学习到的有关说话人的声纹特征. 在自下而上的过程中, 根据刺激对长期记忆进行更新, 长期记忆被建模成一个三元组 $M$, 见式 (11), 其中向量 $K$ 是记忆键值, 矩阵 $V$

表 1 对鸡尾酒会问题建模的单通道语音分离计算模型的回顾总结

Table 1 A review for single-channel speech separation models attacking the cocktail party problem

| 算法分类 | 描述 | 优势 | 劣势 | 代表模型或工作 |
|---|---|---|---|---|
| 基于信号处理的算法 | 假定语音服从一定的分布, 而噪音是平稳或慢变的, 估计噪音的功率谱或者理想维纳滤波器 | 满足条件下能取得较好分离性能 | 现实情况下难以满足假设条件, 因而分离性能大大下降 | 谱减法[36], 维纳滤波器[37−38] |
| 基于分解的算法 | 假设声音的频谱具有低秩结构, 因此可以用一个数量比较小的基来进行表示 | 能够挖掘语音中的基本谱模式 | 1) 线性模型, 难以捕捉语音的高度非线性. 2) 计算代价昂贵, 计算复杂度高, 难以满足实时应用要求 | 1) 浅层模型: NMF[40], 稀疏 NMF[41−43], RNMF[44−45]. 2) 深层模型: D-NMF[46], L-NMF[47]. |
| 基于规则的算法 | 根据听觉场景分析研究中发现的一些规则或机制来对鸡尾酒会问题进行建模 | 以听觉研究得到的规则为支撑, 模型可解释性较强 | 1) 听觉研究一般采用较简单的刺激作为输入, 得到的规律不一定适用于复杂听觉环境. 2) 大部分 CASA 模型严重依赖于分组线索, 尤其是基音提取的准确性, 而这在复杂听觉环境下又难以保证, 因此语音分离效果并不理想. 3) 大多数 CASA 目标是重现 ASA 实验范式中的实验结果, 难以用到实际问题中. | 1) 基于贝叶斯推断的模型: Barniv 等[50]. 2) 基于神经计算的模型: Wang 等[52]. 3) 基于时间相干性的模型: Mill 等[53]. |
| 基于深度学习的算法 | 利用深度神经网络的高度非线性对语音进行建模 | 1) 数据驱动. 2) 能够在大数据集上获得较好性能. | 在真实复杂听觉环境中的表现和人类相比依旧有一定差距: 1) 在开放数据集上的表现逊于封闭数据集. 2) 在区分相似声音时有一定困难. 3) 在处理声源数可变的混合语音时有一定困难. | 1) 只用听觉信息作为输入: Huang 等[60], Du 等[62−63], Weninger 等[65], DC[71−72], PIT[73], DANet[70]. 2) 用视听觉信息作为输入: AVDCNN[84], Gabbay 等[76, 84], Owens 等[88], AVSpeech[89]. |

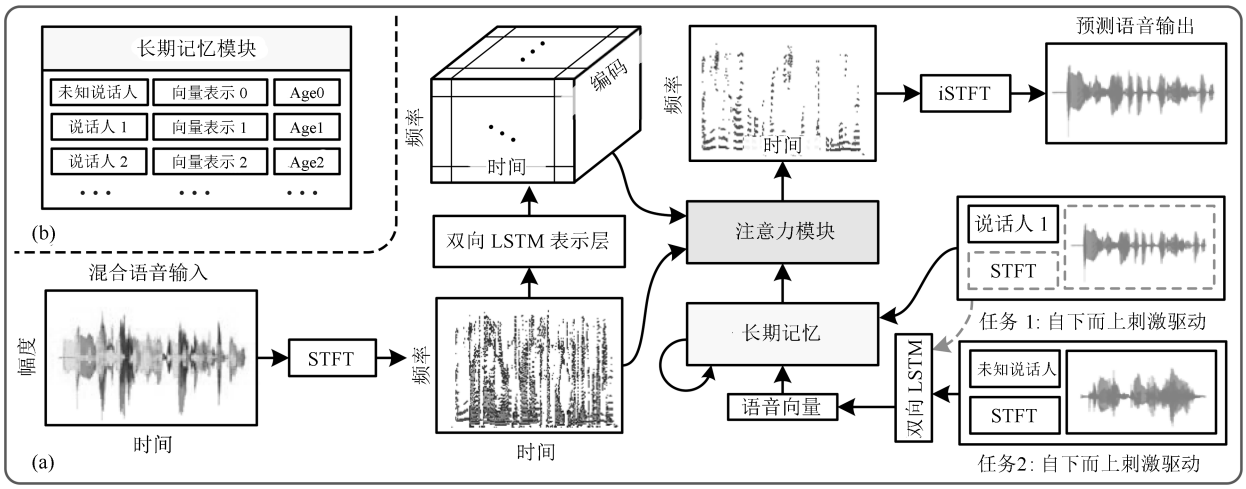



图 4 Xu 等提出的 ASAM 系统的结构[95]

Fig. 4 The structure of the proposed ASAM system by Xu et al.[95]

是记忆的值, 而向量 $A$ 记录长期记忆槽中存放记忆的年龄.

$$M = (K_{\text{memory}}, V_{\text{memory}\times\text{vector}}, A_{\text{memory}}) \quad (11)$$

如果给定说话人 $p$ 在长期记忆中已经存在, 则将其年龄重设为 0 表示最近访问, 并将原记忆值和现有值取平均并归一; 否则, 如果记忆槽未满, 则将说话人 $p$ 写入记忆槽中, 如果记忆槽已满, 则找年龄最大的记忆槽将其覆盖. 每次记忆更新操作, 其他未更新的槽的年龄均加一. 在自上而下的过程中, 模型假定已知需要关注的目标说话人 $p$, 然后从长期记忆中提取该说话人的声纹特征作为先验知识与由当前混合语音经过神经网络得到的隐状态整合到一起, 得到听觉注意掩蔽, 作用于混合语音得到关注目标说话人的言语. 总之, ASAM 模型提出的框架利用类似人脑长时记忆的单元作为关键的信息存储和交互的模块, 将人类听觉过程中的自上而下和自下而上的处理方式统一到一个计算模型中, 突破了之前很多数据驱动模型中单一的自下而上的计算范式, 为鸡尾酒会问题的建模提供了新的思路. 与之前的基于深度学习的方法相比, ASAM 模型明确引入了说话人的声纹信息充当可被学习的先验知识, 为处理鸡尾酒会问题的一项重要线索. 由于先验知识和概念的存在, 自上而下的注意过程变得容易实现. 在复杂的听觉环境下, 这种自上而下的过程可以提升注意的效率, 对于已经建立的概念而言避免了在每个时刻重复且不稳定地推断. 另外, 从 ASAM 模型对于噪音加入之后的性能表现也可以看到, 由于说话人声纹信息的明确性, 其抗干扰能力得到了加强, 避免了在复杂环境下一些无关紧要的各类噪声或背景人声对之前深度学习方法的剧烈影响. 然而, 在 ASAM 模型的设定中, 其对注意目标的形成做了简单的假定, 规定模型一次只能关注一个给定的目标说话人, 这在真实场景中并不现实, 限制了模型在复杂听觉环境中的适用性. 如何从混合语音中自动地抽取多个可能的目标说话人, 并对其各自语音通道进行分离成为了更为关键的问题. 针对这一目标, Shi 等提出 TDAA (Top-down auditory attention) 模型, 使得从混合语音中分离出多个目标说话人成为可能, 而且一定程度上解决了之前的语音分离模型难以处理数目可变说话人的问题[96], 具体框架如图 5. 该模型在设计层面上遵循了模块化的原则, 将原始语音数据驱动的自下而上的过程与目标说话人引导的自上而下的过程串联起来, 更好地模拟了人类听觉通路在鸡尾酒会问题处理过程中的行为. 具体来说, 该模型首先完成自下而上的推理, 预测出候选说话人. 该过程中, TDAA 采用 RNN 分类器一步步地推断出候选说话人, 即每一步从混合语音中推断出最显著的说话人, 然后从混合语音的频谱中减去预测的说话人的频谱作为新的混合语音频谱, 迭代地进行下一步, 预测下一个说话人, 直到最后为空或者满足一定条件为止. 在得到候选说话人之后, 若干个候选说话人各自被用于作为高阶的概念, 引导之后的针对每一个候选说话人的自上而下的语音分离. 在自上而下的过程中, 递归神经网络将输入混合语音的频谱映

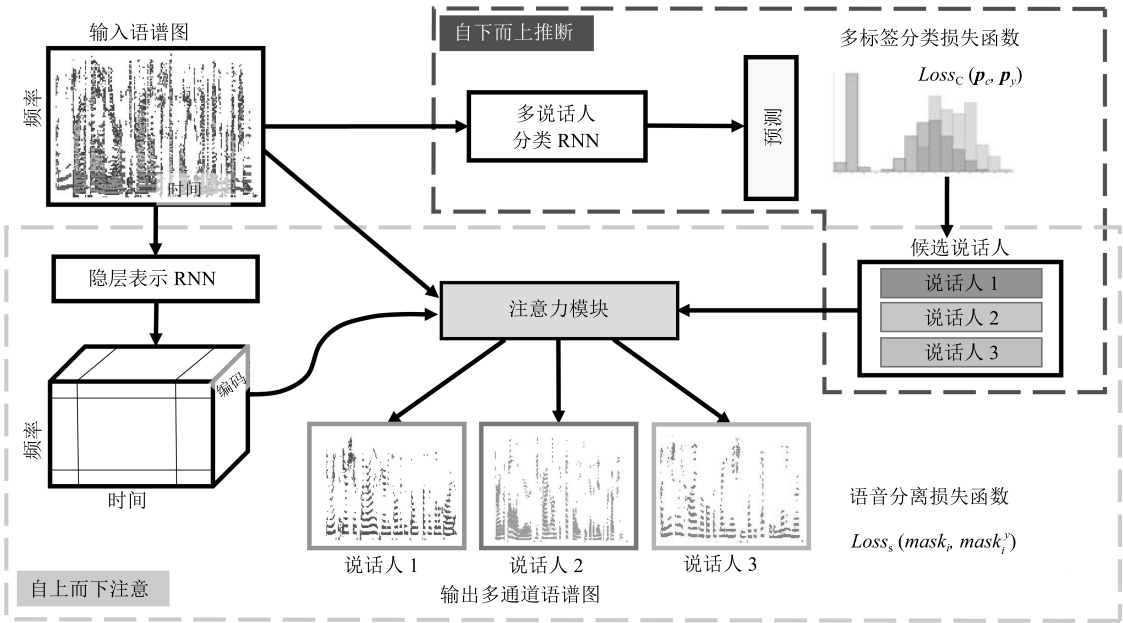



图 5　Shi 等提出的 TDAA 系统的结构[96]

Fig. 5　The structure of the proposed TDAA system by Shi et al.[96]

射到一个 $D$ 维嵌入空间 $V \in \mathbf{R}^{T\times F\times D}$, 其中 $T$ 和 $F$ 表示混合语音频谱的时间维度和频谱维度; 说话人的声纹信息被映射到一个嵌入矩阵 $E \in \mathbf{R}^{N\times D}$, 其中 $N$ 代表训练集中已知的说话人数; 注意模块将 $V$ 和 $E_i$ 作为输入通过注意机制得到候选的第 $i$ 个说话人的 IRM, 作用于混合语音的频谱获得预测的第 $i$ 个说话人的频谱. 对整个 TDAA 模型而言, 其设计的自下而上的过程能够动态地解析出若干个候选说话人作为目标, 解决了之前一大批方法由于多个通道带来的排列问题. 同时, 也使得 TDAA 模型摆脱了说话人数量上的限定, 可以处理说话人数量可变的情况. 从实验结果来看, 该方法在开放数据集上, 即测试集含有未知说话人的数据集上, 也取得了比之前的工作更好的表现. 但是跟封闭数据集比, TDAA 在开放数据集上的表现要差一些, 其表现主要受到自下而上的推断过程的结果影响.

由以上几个工作可以看出, 近期的鸡尾酒会语音分离计算模型依托于深度学习的方法框架, 进行了进一步地升级和探索. 部分工作已经从以深度学习本身的算法为主要关注点, 转变成为结合类脑听觉机制来建模, 从而解决之前遇到的一些问题. 同时形成了以注意为核心, 融合记忆等模块的更加具有解释性的新型架构. 总结而言, 此类结合类脑机制的新型架构更多地借鉴了人类听觉通路过程中的各类机制. 与之前的基于深度学习的方法对比, 该类方法在模型的设计和流程中, 提出了如自上而下的注意过程, 外部长期记忆等有益的模块. 对于目前鸡尾酒会问题中基于数据驱动的, 以自下而上的深度学习方法, 该类计算模型存在受噪声干扰较大, 且通常只能针对于一种设定好的环境 (例如说话人数目固定或者给定) 的局限. 针对这些局限, 结合类脑机制的新方法增强了面对不同情形的鲁棒性, 对于现实环境中的复杂听觉环境无疑具有更好的适应性. 可以预见的是, 这种结合人类听觉工作机制, 脑启发式的模型应该会引起研究人员的关注, 从而进一步促进如听觉注意等机制与鸡尾酒会问题计算模型的更好融合, 并探索出更加合适的建模方式, 为复杂听觉场景的关键问题上提供更好的解决方案.

## 4 多说话人语音识别计算模型回顾

第 2 节和第 3 节描述的计算模型专注于解决复杂听觉环境下的多说话人语音分离问题, 但在鸡尾酒会问题中, 语音分离之后的进一步智能化处理也十分重要. 在深度学习时代之前, 也有不少工作致力于解决多说话人语音识别问题, 其中最有效和著名的一个是 factorial GMM-HMM, 在 2006 年单通道语音分离和识别竞赛中表现超越人类[97−98]. 最近出现一批工作致力于用深度学习的方法, 解决复杂听觉环境下的多说话人语音识别问题. 多说话人语音识别算法目前有两种思路. 第一种思路是构建一个两阶段的模型, 即在语音分离模型之后接语音识别模型对每个分离的语音流进行识别[99−100]. Isik 等在 DC 语音分离模型之后接入一个增强网络进行端到端训练来增强分离的语音流, 再用一个单说话人语音识别系统对分离的语音进行识别[72]. Qian 等提出基于 PIT 的多说话人语音分离−识别系统, 即在基于 PIT 的语音分离系统后接入基于 PIT 的语音识别系统, 对整个系统进行联合训练[98]. Settle 等在 CTC/Attention 混合结构 (Hybrid connectionist temporal classification/attention architecture) 下在改进的 DC 模型后接入一个端到端的语音识别系统并采用 PIT 方法, 对语音分离系统和语音识别系统进行联合训练[101]. 上述模型基本需要分别对语音分离模型和语音识别模型进行训练或者预训练, 难以直接从头开始训练 (From scratch). 第二种思路则是直接对混合语音进行识别而没有显式的分离阶段. Weng 等使用多方式训练 (Multi-style training) 结合不同的目标函数, 针对多说话人复杂听觉环境中的不同情况生成相应的训练数据用来训练深度神经网络[102]. Qian 和 Yu 等对 PIT 进行拓展直接对混合语音进行识别, 使用交叉熵作为误差函数, 对所有可能的排列进行计算并选择最小的排列来更新模型参数[98, 103], 模型结构如图 6, 这里模型输出的标签为多元音素 (Senone). 基于 PIT 的多说话人语音识别系统由于其简洁性, 很容易和其他成熟的技术结合起来, 从而提高语音识别的正确率, 比如说话人自适应技术 (Speaker adaptation)[104], 序列判别训练 (Sequence discriminative training)[105], 知识蒸馏 (Knowledge distillation)[106] 和注意机制 (Attention mechanism)[107]. 但是上述没有显式分离阶段的模型, 在训练语音识别模型的时候需要使用一个预训练的单说话人语音识别模型做多元音素对齐 (Senone alignment)[101], 无法做到真正的端到端训练. 因此 Seki 等提出一个端到端的多说话人语音识别系统并采用 PIT 方法, 直接对输入的混合语音进行语音识别而无需使用音素级别的标签[100], 具体框图如图 7. 整个模型是一个 CTC/Attention 混合结构: 在编码器端共有三个层次的编码器, 分别为混合语音编码器, 说话人区分的编码器和识别编码器, 而在解码器端则使用 CTC 和基于注意的解码器. 编码器端的混合语音编码器相当于一个创建了能够区

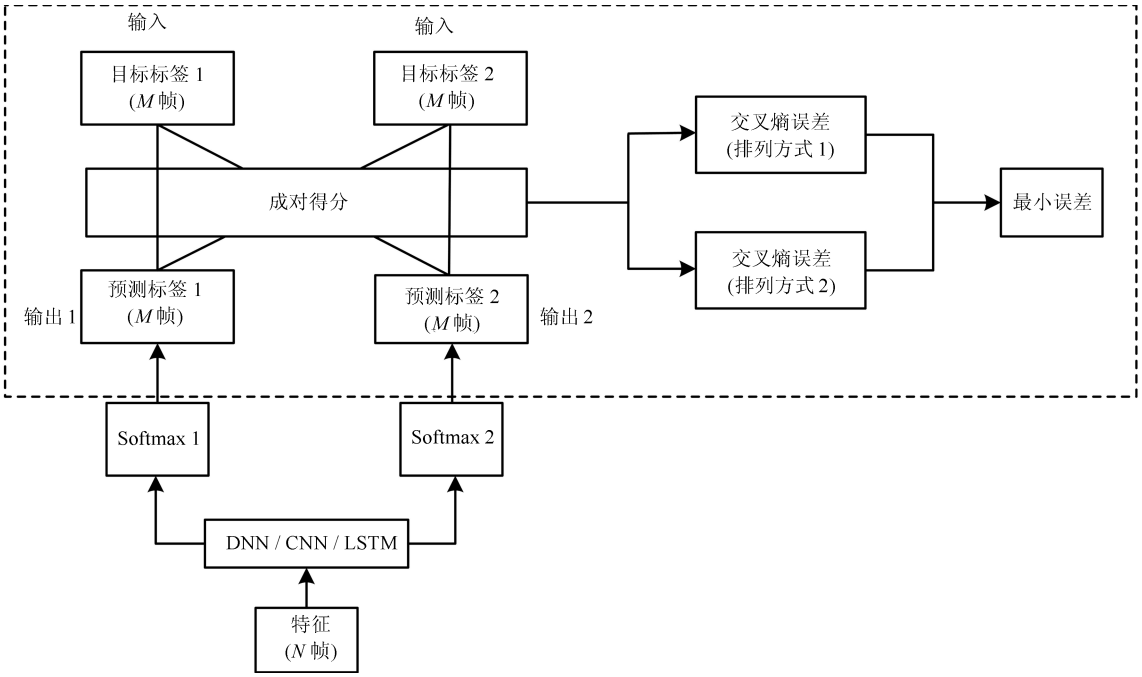



图 6　Qian 和 Yu 等提出的基于排列不变性训练方法的双说话人语音识别系统的结构[98, 103]

Fig. 6　The structure of the proposed direct two-speaker speech recognition system with PIT by Qian and Yu et al.[98, 103]

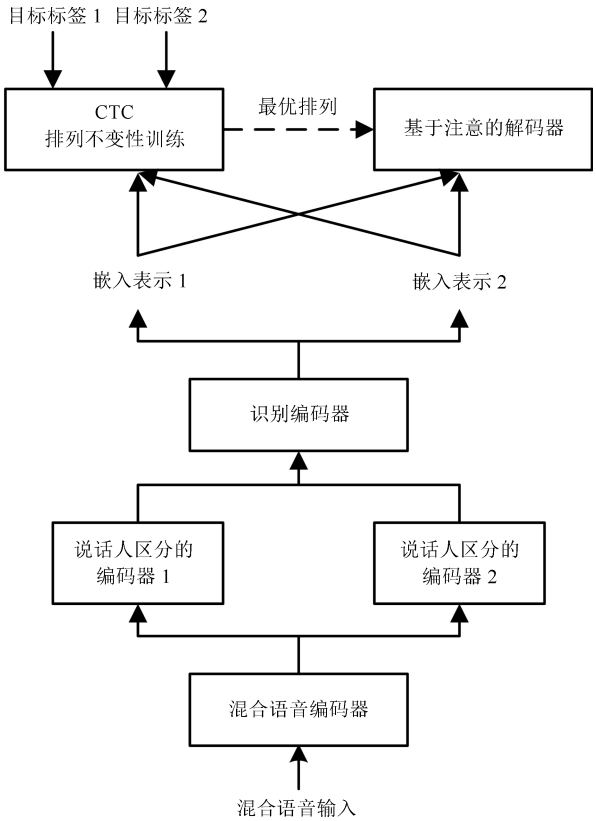



图 7　Seki 等提出的双说话人语音识别系统的结构[100]

Fig. 7　The structure of the proposed end-to-end two-speaker speech recognition system by Seki et al.[100]

分多个声源的嵌入向量的语音分离模块, 说话人区分的编码器则从上一阶段的输出提取出各个说话人的说话内容以备识别, 识别编码器则相当于一个编码了单个说话人的言语的声学模型用以最后的解码. 解码器端为减小计算成本, 通过采用 CTC 来确定所有可能的排列中误差最小的排列, 而基于注意的解码器则采用该排列进行解码. 实验表明该工作和之前端到端有显式分离和识别过程的模型[101] 效果相当, 但无需依赖预训练的语音分离系统. Chang 等则在文献 [100] 的基础上对其中的基于注意的解码器进行改进, 使得解码每个说话人的基于注意的解码器权值不共享, 以减轻编码器区分语音的负担[108].

由上面的工作可见, 近期一批对鸡尾酒会问题建模的计算模型进一步升级, 开始同时考虑语音分离之后的智能化处理. 语音分离只是朝向解决鸡尾酒会问题的第一步, 如何协同后续的智能化处理以进一步提升模型的性能, 将成为今后该领域研究的一个研究热点.

## 5　鸡尾酒会问题的思考与展望

近年来, 随着智能设备广泛进入日常生活的各个角落, 处理复杂听觉环境下的鸡尾酒会问题变成了非常受关注的一个领域, 在某种意义上成为了智能设备的关键入口和通道. 受益于大数据和深度学习技术的迅猛发展, 对鸡尾酒会问题建模的语音分

离计算模型已从原来的基于规则, 基于信号处理的方法逐渐变为了数据驱动型的, 基于深度学习的方法. 近期还出现了一批工作, 关注鸡尾酒会问题建模中的听觉机制建模和随后的智能化处理. 到目前为止, 尽管各类研究取得了一定的成果, 但是离真正解决鸡尾酒会问题还相去甚远. 可以预见, 未来若干年, 关于如何处理鸡尾酒会问题势必仍然是非常受瞩目的一个方向. 本文回顾了听觉研究的相关机制和对鸡尾酒会问题建模的相关模型. 我们认为, 针对鸡尾酒会问题的神经学机制以及计算模型方面, 目前还有一些非常值得探索的问题和方向, 主要包括:

1) 听觉系统是一个高度非线性的系统, 神经回路中神经元之间的连接十分复杂, 神经元对刺激采用多种编码方式, 主要有频率编码, 时间编码和群体编码这三种方式. 声音中富有丰富的时空结构, 而听觉系统对这些时空结构是高度敏感的. 而在最近的基于深度学习的语音分离算法中, 对语音的编码方式较为单一, 即神经元只使用频率编码, 可能不能充分挖掘利用语音中的时空结构. CASA 中基于神经网络的模型采用的振荡脉冲神经网络[52], 而脉冲神经网络在时间编码较有优势. 但是目前脉冲神经网络的性能与人工神经网络相比, 存在较大差距. 对语音时间编码的研究是一个值得探讨的问题.

2) 传统计算模型对复杂听觉场景的建模能力较为有限, 难以迁移到真实场景中. 尽管近年来通过扩大训练数据集覆盖大多数听觉环境, 运用深度学习, 模型在真实场景下的语音分离表现大幅度提升, 并且能够在开放数据集上取得不错的表现, 但相比人类处理鸡尾酒会问题的表现, 依旧有一定差距. 大多数模型都假定说话人的数目固定, 难以处理有不确定数目的说话人的情况, 比如 DC[71−72] 需要给定聚类的个数才能工作. 虽然 TDAA 模型[96] 通过结合迭代的自下而上推断过程和自上而下的注意过程, 令模型能够处理可变数目的说话人, 但该模型的表现大大受到自下而上推断过程得到的候选结果, 而其在开放数据集上的表现依然逊色于封闭数据集.

3) 仅仅用听觉模态的信息, 难以区分相类似的声音, 比如同性别说话人的声音. 近几年, 基于多感知整合的理论, 计算模型开始将视觉信息整合到语音分离当中, 一定程度上解决处理类似声音的问题. 利用听觉信息和视觉信息时间上的高度相关性, 可以进行自监督学习, 从而无需标记数据[88]. 尽管关于多感知整合发生在哪个阶段尚未有定论, 但是跨模态注意和多感知整合在大脑的某些处理层次中确实存在[22]. 目前已有工作从听觉注意出发, 对鸡尾酒会问题进行计算建模, 比如 ASAM[95], TDAA[96], 但是尚未有工作从视听觉多通道注意的角度对鸡尾酒会问题进行建模.

4) 值得注意的是, ASAM 还在建模中引入了长期记忆的机制, 但是目前其长期记忆的每个单元可能过于简化. 另外, 视听觉注意的触发时机问题, 也值得关注. 如何对视听觉注意与跨模态进行计算建模, 并将得到的时序模式根据一定的规则转存为长期记忆作为先验知识加以利用, 使得语音分离更有效率, 是未来值得探索的方向.

5) 语音分离之后的智能化处理. 近期多说话人语音识别计算模型的工作开始同时考虑语音分离之后的智能化处理[98, 100−108]. 语音分离虽然十分关键, 但只是朝向解决鸡尾酒会问题的第一步, 如何协同后续的智能化处理, 是今后解决鸡尾酒会问题一个值得研究的关键点.

6) 如何将预测加工融入到听觉计算模型中. 前面在第 1 节也提到, 大脑是一个具有预测能力的层次化结构, 在处理复杂听觉场景时, 总是试图预测接下来要到来的声音, 而且预测加工机制的存在使得人大脑的听觉中枢能够根据上下文实时补充和恢复单词中缺失的音素或音节. 但是现在几乎没有计算模型从这方面入手进行建模. 如果计算模型能够实时补全言语中被噪声掩盖的音素或音节而形成在语义上符合上下文内容的单词, 这势必是迈向鸡尾酒会问题计算建模解决方案的一大步.

综上所述, 我们认为要解决复杂听觉场景下的鸡尾酒会问题, 需要将计算模型和听觉研究中的一些相关机制深度结合起来. 听觉系统对刺激的编码策略, 听觉感知中的预测特性, 视听觉注意的整合和触发时机等等听觉和认知心理学研究中得到的一些基本成果, 应该如何借鉴到计算模型的建模中, 可能会成为解决鸡尾酒会问题的新的突破口.

## References


1 Cherry E C. Some experiments on the recognition of speech, with one and with two ears. *Journal of the Acoustical Society of America*, 1953, **25**(5): 975−979

2 Mesgarani N, Chang E F. Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, 2012, **485**(7397): 233−236

3 Pannese A, Grandjean D, Frühholz S. Subcortical processing in auditory communication. *Hearing Research*, 2015, **328**: 67−77

4 Moore B C J. *An Introduction to the Psychology of Hearing* (Sixth edition). Leiden: Brill, 2013.

5 Zhao Ling-Yun. Characteristics and Mechanisms of the Processing of Complex Time-frequency Information in Sub-cortical Auditory Pathway [Master thesis], Tsinghua University, China, 2010
(赵凌云. 皮层下听觉通路处理复杂时频信息的特性与机理研究[硕士学位论文], 清华大学, 中国, 2010)

6 Bizley J K, Cohen Y E. The what, where and how of auditory-object perception. *Nature Reviews Neuroscience*, 2013, **14**(10): 693−707

7 Bregman A S. *Auditory Scene Analysis: the Perceptual Organization of Sound*. Cambridge: MIT Press, 1990.

8 O′ Sullivan J A, Power A J, Mesgarani N, Rajaram S, Foxe J J, Shinn-Cunningham B G, et al. Attentional selection in a cocktail party environment can be decoded from single-trial EEG. *Cerebral Cortex*, 2014, **25**(7): 1697−1706

9 Shamma S A, Elhilali M, Micheyl C. Temporal coherence and attention in auditory scene analysis. *Trends in Neurosciences*, 2011, **34**(3): 114−123

10 Ayala Y A, Malmierca M S. Stimulus-specific adaptation and deviance detection in the inferior colliculus. *Frontiers in Neural Circuits*, 2012, **6**: Article No. 89

11 Ulanovsky N, Las L, Farkas D, Nelken I. Multiple time scales of adaptation in auditory cortex neurons. *The Journal of Neuroscience*, 2004, **24**(46): 10440−10453

12 Winkler I, Denham S L, Nelken I. Modeling the auditory scene: predictive regularity representations and perceptual objects. *Trends in Cognitive Sciences*, 2009, **13**(12): 532−540

13 Leonard M K, Baud M O, Sjerps M J, Chang E F. Perceptual restoration of masked speech in human cortex. *Nature Communications*, 2016, **7**: Article No. 13619

14 Gregory R L. Perceptions as hypotheses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 1980, **290**(1038): 181−197

15 Friston K. A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 2005, **360**(1456): 815−836

16 Bar M. The proactive brain: using analogies and associations to generate predictions. *Trends in Cognitive Sciences*, 2007, **11**(7): 280−289

17 Bubic A, Von Cramon D Y, Schubotz R I. Prediction, cognition and the brain. *Frontiers in Human Neuroscience*, 2010, **4**: Article No. 25

18 Pearce J M, Hall G. A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review*, 1980, **87**(6): 532−552

19 Kruschke J K. Toward a unified model of attention in associative learning. *Journal of Mathematical Psychology*, 2001, **45**(6): 812−863

20 Mackintosh N J. A theory of attention: variations in the associability of stimuli with reinforcement. *Psychological Review*, 1975, **82**(4): 276−298

21 Wills A J, Graham S, Koh Z, Mclaren I P L, Rolland M D. Effects of concurrent load on feature- and rule-based generalization in human contingency learning. *Journal of Experimental Psychology: Animal Behavior Processes*, 2011, **37**(3): 308−316

22 Koelewijn T, Bronkhorst A, Theeuwes J. Attention and the multiple stages of multisensory integration: a review of audiovisual studies. *Acta Psychologica*, 2010, **134**(3): 372−384

23 Shimojo S, Shams L. Sensory modalities are not separate modalities: plasticity and interactions. *Current Opinion in Neurobiology*, 2001, **11**(4): 505−509

24 Mcgurk H, Macdonald J. Hearing lips and seeing voices. *Nature*, 1976, **264**(5588): 746−748

25 Golumbic E Z, Cogan G B, Schroeder C E, Poeppel D. Visual input enhances selective speech envelope tracking in auditory cortex at a "cocktail party". *Journal of Neuroscience*, 2013. **33**(4): 1417−1426

26 Vroomen J, Bertelson P, De Gelder B. The ventriloquist effect does not depend on the direction of automatic visual attention. *Perception & Psychophysics*, 2001, **63**(4): 651−659

27 Schwartz J L, Berthommier F, Savariaux C. Audio-visual scene analysis: evidence for a "very-early" integration process in audio-visual speech perception. In: Proceedings of the 7th International Conference on Spoken Language Processing. Denver, USA: DBLP, 2002

28 Omata K, Mogi K. Fusion and combination in audio-visual integration. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 2008, **464**(2090): 319−340

29 Busse L, Roberts K C, Crist R E, Weissman D H, Woldorff M G. The spread of attention across modalities and space in a multisensory object. *Proceedings of the National Academy of Sciences of the United States of America*, 2005, **102**(51): 18751−18756

30 Talsma D, Woldorff M G. Selective attention and multisensory integration: multiple phases of effects on the evoked brain activity. *Journal of Cognitive Neuroscience*, 2005, **17**(7): 1098−1114

31 Calvert G A, Thesen T. Multisensory integration: methodological approaches and emerging principles in the human brain. *Journal of Physiology-Paris*, 2004, **98**(1−3): 191−205

32 Gannot S, Vincent E, Markovich-Golan S, Ozerov A. A consolidated perspective on multimicrophone speech enhancement and source separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2017, **25**(4): 692−730

33 Adel H, Souad M, Alaqeeli A, Hamid A. Beamforming techniques for multichannel audio signal separation. *International Journal of Digital Content Technology and Its Applications*, 2012, **6**(20): 659−667

34 Sawada H, Araki S, Mukai R, Makino S. Blind extraction of dominant target sources using ICA and time-frequency masking. *IEEE Transactions on Audio, Speech, and Language Processing*, 2006, **14**(6): 2165−2173

35 Chen Z. Single Channel Auditory Source Separation with Neural Network [Ph. D. dissertation], Columbia University, USA, 2017

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

37 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

38 Loizou P C. *Speech Enhancement: Theory and Practice*. Boca Raton: CRC Press, Inc., 2007.

39 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
(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)

40 Lee D D, Seung H S. Algorithms for non-negative matrix factorization. In: Proceedings of the 13th International Conference on Neural Information Processing Systems (NIPS). Denver, USA: MIT Press, 2000. 535−541

41 Hoyer P O. Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, 2004, **5**: 1457−1469

42 Schmidt M N, Olsson R K. Single-channel speech separation using sparse non-negative matrix factorization. In: Proceedings of the 9th International Conference on Spoken Language Processing. Pittsburgh, USA: INTERSPEECH, 2006.

43 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

44 Chen Z, Ellis D P W. Speech enhancement by sparse, low-rank, and dictionary spectrogram decomposition. In: Proceedings of the 2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. New Paltz, USA: IEEE, 2013. 1−4

45 Zhang L J, Chen Z G, Zheng M, He X F. Robust nonnegative matrix factorization. *Frontiers of Electrical and Electronic Engineering in China*, 2011, **6**(2): 192−200

46 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 (ICASSP). Brisbane, Australia: IEEE, 2015. 66−70

47 Hsu C C, Chi T S, Chien J T. Discriminative layered nonnegative matrix factorization for speech separation. In: Proceedings of the 17th Annual Conference of the International Speech Communication Association (INTERSPEECH 2016). San Francisco, USA: INTERSPEECH, 2015. 560−564

48 Rouat J. Computational auditory scene analysis: principles, algorithms, and applications (Wang D. and Brown G.J., Eds.; 2006) [Book review]. *IEEE Transactions on Neural Networks*, 2008, **19**(1): Article No. 199

49 Szabò B T, Denham S L, Winkler I. Computational models of auditory scene analysis: a review. *Frontiers in Neuroscience*, 2016, **10**: Article No. 524

50 Barniv D, Nelken I. Auditory streaming as an online classification process with evidence accumulation. *PLoS One*, 2015, **10**(12): Article No. e0144788

51 Llinàs R R. Intrinsic electrical properties of mammalian neurons and CNS function: a historical perspective. *Frontiers in Cellular Neuroscience*, 2014, **8**: Article No. 320

52 Wang D L, Chang P. An oscillatory correlation model of auditory streaming. *Cognitive Neurodynamics*, 2008, **2**(1): 7−19

53 Mill R W, Böhm T M, Bendixen A, Winkler I, Denham S L. Modelling the emergence and dynamics of perceptual organisation in auditory streaming. *PLoS Computational Biology*, 2013, **9**(3): Article No. e1002925

54 Elhilali M, Shamma S A. A cocktail party with a cortical twist: how cortical mechanisms contribute to sound segregation. *Journal of the Acoustical Society of America*, 2008, **124**(6): 3751−3771

55 Ma L. Auditory Streaming: Behavior, Physiology, and Modeling [Ph. D. dissertation], University of Maryland, USA, 2011

56 Krishnan L, Elhilali M, Shamma S. Segregating complex sound sources through temporal coherence. *PLoS Computational Biology*, 2014, **10**(12): Article No. e1003985

57 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

58 Wang Yan-Nan. Speaker-Independent Single-Channel Speech Separation Based on Deep Learning [Ph. D. dissertation], University of Science and Technology of China, China, 2017
(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], 中国科学技术大学, 中国, 2017)

59 Narayanan A, Wang D L. Ideal ratio mask estimation using deep neural networks for robust speech recognition. In: Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Vancouver, Canada: IEEE, 2013. 7092−7096

60 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 (ICASSP). Florence, Italy: IEEE, 2014. 1562−1566

61 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

62 Du J, Tu Y H, Xu Y, Dai L R, Lee C H. Speech separation of a target speaker based on deep neural networks. In: Proceedings of the 12th International Conference on Signal Processing (ICSP). Hangzhou, China: IEEE, 2014. 473−477

63 Tu Y H, Du J, Xu Y, Dai L R, Lee C H. Speech separation based on improved deep neural networks with dual outputs of speech features for both target and interfering speakers. In: Proceedings of the 9th International Symposium on Chinese Spoken Language Processing. Singapore: IEEE, 2014. 250−254

64 Du J, Tu Y H, Dai L R, Lee C H. A regression approach to single-channel speech separation via high-resolution deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2016, **24**(8): 1424−1437

65 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 (GlobalSIP). Atlanta, USA: IEEE, 2014. 577−581

66 Wang D L, Chen J T. Supervised speech separation based on deep learning: an overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2018, **26**(10): 1702−1726

67 Zhang X L, Wang D L. A deep ensemble learning method for monaural speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2016, **24**(5): 967−977

68 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, 2015. 259−267

69 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

70 Chen Z, Luo Y, Mesgarani N. Deep attractor network for single-microphone speaker separation. In: Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New Orleans, USA: IEEE, 2017. 246−250

71 Hershey J R, Chen Z, Le Roux J, Watanabe S. Deep clustering: discriminative embeddings for segmentation and separation. In: Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Shanghai, China: IEEE, 2016. 31−35

72 Isik Y, Le Roux J, Chen Z, Watanabe S, Hershey J R. Single-channel multi-speaker separation using deep clustering. In: Proceedings of the 2016 INTERSPEECH. Broadway, USA: Mitsubishi Electric Research Laboratories, Inc., 2016. 545−549

73 Yu D, Kolbaek M, Tan Z H, Jensen J. Permutation invariant training of deep models for speaker-independent multi-talker speech separation. In: Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New Orleans, USA: IEEE, 2017. 241−245

74 Kolbaek M, Yu D, Tan Z H, Jensen J. 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

75 Kuhl P K. Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics*, 1991, **50**(2): 93−107

76 Gabbay A, Ephrat A, Halperin T, Peleg S. Seeing through noise: speaker separation and enhancement using visually-derived speech. arXiv preprint arXiv: 1708.06767, 2017.

77 Rajaram S, Nefian A V, Huang T S. Bayesian separation of audio-visual speech sources. In: Proceedings of the 2004 International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Montreal, Canada: IEEE, 2004. 657−660

78 Sodoyer D, Girin L, Jutten C, Schwartz J L. Developing an audio-visual speech source separation algorithm. *Speech Communication*, 2004, **44**(1−4): 113−125

79 Dansereau R M. Co-channel audiovisual speech separation using spectral matching constraints. In: Proceedings of the 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Montreal, Canada: IEEE, 2004. 645−648

80 Wang W W, Cosker D, Hicks Y, Sanei S, Chambers J. Video assisted speech source separation. In: Proceedings of the 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Philadelphia, USA: IEEE, 2005. 425−428

81 Rivet B, Girin L, Jutten C. Mixing audiovisual speech processing and blind source separation for the extraction of speech signals from convolutive mixtures. *IEEE Transactions on Audio, Speech, and Language Processing*, 2007, **15**(1): 96−108

82 Barzelay Z, Schechner Y Y. Harmony in motion. In: Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition. Minneapolis, USA: IEEE, 2007. 4358−4366

83 Casanovas A L, Monaci G, Vandergheynst P, Gribonval R. Blind audiovisual source separation based on sparse redundant representations. *IEEE Transactions on Multimedia*, 2010, **12**(5): 358−371

84 Hou J C, Wang S S, Lai Y H, Tsao Y, Chang H W, Wang H M. Audio-visual speech enhancement based on multimodal deep convolutional neural network. arXiv preprint arXiv: 1703.10893, 2017

85 Ephrat A, Halperin T, Peleg S. Improved speech reconstruction from silent video. In: Proceedings of the 2017 International Conference on Computer Vision Workshops (ICCV). Venice, Italy: IEEE, 2017. 455−462

86 Gabbay A, Shamir A, Peleg S. Visual speech enhancement. In: Proceedings of the 2018 INTERSPEECH. Hyderabad, India: INTERSPEECH, 2018. 1170−1174

87 DeSa V R. Learning classification with unlabeled data. In: Proceedings of the 6th International Conference on Neural Information Processing Systems (NIPS). Denver, USA: Morgan Kaufmann Publishers Inc., 1993. 112−119

88 Owens A, Efros A A. Audio-visual scene analysis with self-supervised multisensory features. arXiv preprint arXiv: 1804.03641, 2018

89 Ephrat A, Mosseri I, Lang O, Dekel T, Wilson K W, Hassidim A, Freeman W T, Rubinstein M. Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation. *ACM Transactions on Graphics*, 2018, **37**(4): Article No. 112

90 Senocak A, Oh T H, Kim J, Yang M H, Kweon I S. Learning to localize sound source in visual scenes. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 4358−4366

91 Arandjelovič R, Zisserman A. Objects that sound. arXiv preprint arXiv: 1712.06651, 2017.

92 Zhao H, Gan C, Rouditchenko A, Vondrick C, McDermott J, Torralba A. The sound of pixels. In: Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer, 2018.

93 Ciocca V. The auditory organization of complex sounds. *Frontiers in Bioscience*, 2008, **13**(13): 148−169

94 Elhilali M. Modeling the cocktail party problem. *The Auditory System at the Cocktail Party*. Cham: Springer, 2017.

95 Xu J M, Shi J, Liu G C, Chen X Y, Xu B. Modeling attention and memory for auditory selection in a Cocktail Party environment. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI). New Orleans, USA: AIAA, 2018.

96 Shi J, Xu J M, Liu G C, Xu B. Listen, think and listen again: capturing top-down auditory attention for speaker-independent speech separation. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI). Stockholm, Sweden: IJCAI, 2018. 4353−4360

97 Kristjansson T T, Hershey J R, Olsen P A, Rennie S J, Gopinath R. Super-human multi-talker speech recognition: The IBM 2006 speech separation challenge system. In: Proceedings of the 9th International Conference on Spoken Language Processing. Pittsburgh, USA: INTERSPEECH, 2006.

98 Qian Y M, Chang X K, Yu D. Single-channel multi-talker speech recognition with permutation invariant training. *Speech Communication*, 2018, **104**: 1−11

99 Qian Y M, Weng C, Chang X K, Wang S, Yu D. Past review, current progress, and challenges ahead on the cocktail party problem. *Frontiers of Information Technology & Electronic Engineering*, 2018, **19**(1): 40−63

100 Seki H, Hori T, Watanabe S, Le Roux J, Hershey J R. A purely end-to-end system for multi-speaker speech recognition. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL). Melbourne, Australia: ACL, 2018. 2620−2630

101 Settle S, Le Roux J, Hori T, Watanabe S, Hershey J R. End-to-end multi-speaker speech recognition. In: Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, Canada: IEEE, 2018. 4819−4823

102 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

103 Yu D, Chang X K, Qian Y M. Recognizing multi-talker speech with permutation invariant training. In: Proceedings of INTERSPEECH. Stockholm, Sweden: INTERSPEECH, 2017. 2456−2460

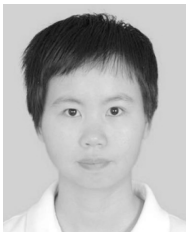
104 Chang X K, Qian Y M, Yu D. Adaptive permutation invariant training with auxiliary information for monaural multi-talker speech recognition. In: Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, Canada: IEEE, 2018. 5974−5978

105 Chen Z H, Droppo J, Li J Y, Xiong W. Progressive Joint Modeling in Unsupervised Single-Channel Overlapped Speech Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2018, **26**(1): 184−196

106 Tan T, Qian Y M, Yu D. Knowledge transfer in permutation invariant training for single-channel multi-talker speech recognition. In: Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, Canada: IEEE, 2018. 5714−5718

107 Chang X K, Qian Y M, Yu D. Monaural multi-talker speech recognition with attention mechanism and gated convolutional networks. In: Proceedings of INTERSPEECH. Hyderabad, India: INTERSPEECH, 2018. 1586−1590

108 Chang X K, Qian Y M, Yu K, Watanabe S. End-to-end monaural multi-speaker ASR system without pretraining. arXiv preprint arXiv: 1811.02062, 2018.


**黄雅婷** 中国科学院自动化研究所博士研究生. 主要研究方向是语音分离, 听觉模型, 类脑智能. 本文共同第一作者.
E-mail: huangyating2016@ia.ac.cn
(**HUANG Ya-Ting** Ph. D. candidate at the Institute of Automation, Chinese Academy of Sciences. Her research interest covers speech separation, auditory model, and brain-inspired intelligence. Co-first author of this paper.)



**石 晶** 中国科学院自动化研究所博士研究生. 主要研究方向是语音分离, 听觉模型, 自然语言处理, 深度学习. 本文共同第一作者.
E-mail: shijing2014@ia.ac.cn
(**SHI Jing** Ph. D. candidate at the Institute of Automation, Chinese Academy of Sciences. His research interest covers speech separation, auditory model, natural language processing and deep learning. Co-first author of this paper.)



**许家铭** 中国科学院自动化研究所副研究员. 主要研究方向为语音处理与听觉注意, 智能问答和对话, 深度学习和强化学习. 本文通信作者.
E-mail: jiaming.xu@ia.ac.cn
(**XU Jia-Ming** Associate professor at the Institute of Automation, Chinese Academy of Sciences. His research interest covers speech processing and auditory attention, question & answering and dialog system, deep learning and reinforcement learning. Corresponding author of this paper.)



**徐 波** 中科院自动化所所长, 研究员. 中科院脑科学与智能技术卓越创新中心副主任. 长期从事人工智能研究, 主要研究方向为类脑智能, 类脑认知计算模型, 自然语言处理与理解, 类脑机器人.
E-mail: xubo@ia.ac.cn
(**XU Bo** Professor, president of the Institute of the Automation, Chinese Academy of Sciences, and deputy director of the Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences. His research interest covers brain-inspired intelligence, brain-inspired cognitive models, natural language processing and understanding, brain-inspired robotics.)