中国科学: 信息科学 2015 年 第 45 卷 第 10 期: 1310–1327
www.scichina.com info.scichina.com
《中国科学》杂志社
SCIENCE CHINA PRESS

面向感知中国的新一代信息技术专刊 · 论文

具有选择注意能力的语音拾取技术

王晓飞, 国雁萌, 葛凤培, 吴超, 付强*, 颜永红

中国科学院声学研究所, 中国科学院语言声学与内容理解重点实验室, 北京 100190
* 通信作者. E-mail: fuqiang@hccl.ioa.ac.cn

收稿日期: 2015–06–26; 接受日期: 2015–07–31; 网络出版日期: 2015–10–13
国家自然科学基金项目 (批准号: 11161140319)、中国科学院战略性先导科技专项 (批准号: XDA06030100, XDA06030500)、国家高技术研究发展计划 (863 计划) 项目 (批准号: 2012AA012503) 和中国科学院重点部署项目 (批准号: KGZD-EW-103-2) 资助

摘要 随着语音通信和人机语音交互系统的不断普及, 人们越来越期待抛开话筒和耳机等繁琐的设备, 实现像人类对话一般自然的人机语音交流. 然而, 语音毕竟只是一种声波, 在空气中传输时难免受到各种影响, 例如声波的衰减、墙壁和障碍物的多次反射以及同时存在的其他声源等. 如果不采用近讲的拾音方式, 那么这些因素都会对传播中的语音声波造成干扰. 特别是当多个语音系统和多个说话人处于同一环境时, 如何确保系统正确接收语音信息, 决定了语音系统能否走向实用. 本文参考人类的听觉注意机理, 提出充分利用对目标语音及干扰声源的先验知识, 检测和提升目标语音, 并通过将传声器阵列、语音唤醒、目标语音检测、语音增强、混响抑制等一系列技术相结合, 实现抗干扰的目标语音拾取.

关键词 传声器阵列 唤醒词 声学回波控制 语音增强 目标语音检测 混响抑制

1 引言

语音是人类最自然最常用的沟通方式. 人类通过发声器官, 把文本信息调制为声波并发送到空气中, 并由对方的听觉系统接收, 即可实现面对面的信息交流. 直接的语音交流会受到声学环境的制约, 不仅传输距离有限, 而且信息传输的效率和可靠性都易受周围环境的影响. 然而, 自电话机问世以来, 人们通过手持话筒或佩戴耳机等设备, 能够实现超视距的语音传播. 这不仅大大提高了人类对语音交流的预期, 而且促使这种被称为 “近讲模式” (close-talking mode) 的拾音方式获得了广泛接受. 近讲模式通过直接在用户口边采集语音, 显著弱化了声学环境的影响, 并推动了语音识别、话者识别等语音技术走向实际应用.

但是, 近讲模式与人类自然的语音交流模式相差很大, 不管手持、佩戴还是靠近拾音设备, 用户都会感觉到一定不便, 在某些情况下甚至无法实现. 随着智能家居、智能车载和可穿戴设备等新技术和新应用的不断涌现, 越来越多的智能设备具备了语音功能, 人们更需要摆脱近讲模式的束缚, 以更灵活自然的方式实现语音拾取, 这也带来一系列的技术挑战.

首先, 实际使用环境中除了语音源之外, 可能存在其他声源, 例如电视、空调、脚步声等. 它们辐射出的声波会与语音混合在一起, 影响到语音信息的提取, 这称为背景噪声问题. 一种特别的情况是,

引用格式: 王晓飞, 国雁萌, 葛凤培, 等. 具有选择注意能力的语音拾取技术. 中国科学: 信息科学, 2015, 45: 1310–1327, doi: 10.1360/N112014-00343

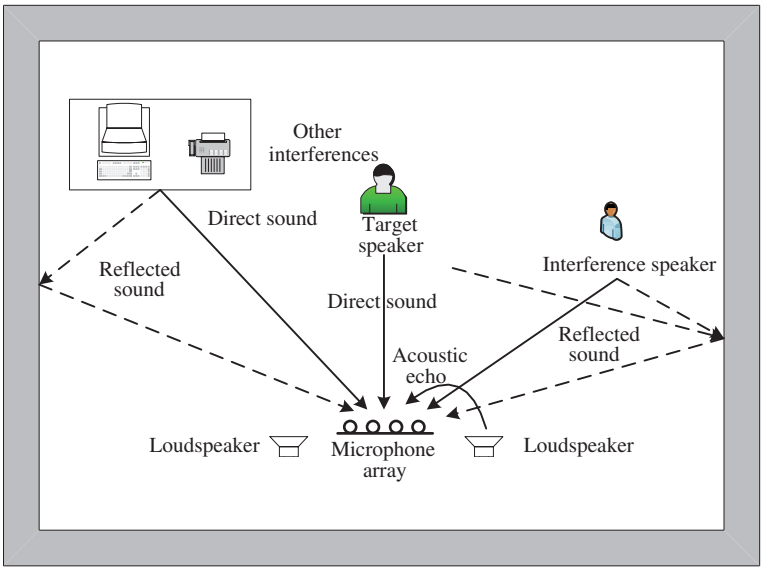


图 1 封闭空间内的拾音系统和声学环境示意图

Figure 1 Schematic diagram of the pickup system and acoustic environment in the enclosed space

如果语音系统以扬声器发出声音, 例如智能电视播放节目, 那么发射出的声波又会被语音系统采集回来, 成为一种噪声干扰, 这叫做 “声学回波”. 由于声学回波往往能量较强, 如果不进行特殊处理, 会严重影响语音系统的正常工作.

其次, 语音系统所处的声学环境复杂多样, 声波的传播也会受其影响. 用户发声后, 声波以球面波的方式向外传播并不断衰减, 遇到障碍物后会被部分吸收并反射, 最终到达语音系统的声波不仅包括衰减后的直达声, 还可能包括反射声和多次反射声, 这称为 “混响” 问题. 当用户距离拾音设备较远时, 混响会严重影响拾音的效果. 同时, 由于声音传播路径随着说话人和障碍物的移动而改变, 混响的影响往往是时变的. 因此, 要可靠提取语音信息, 混响也是必须解决的问题.

再次, 作为基本的信息传输方式, 用户的语音具有很多功能, 例如与周围人交流、和远方用户语音通信、控制多台智能设备等, 并且随时会在各功能间切换. 用户所发的语音声波可以到达附近的每个语音系统, 而每个语音系统也可以接收周围所有人的语音. 因此, 语音系统接收的所有信号中, 包含大量的无关语音, 只有从中检测出针对本系统的用户语音, 即 “目标语音”, 才能正确处理语音信息.

因此, 要实现自然的语音拾取, 首先需要从复杂的混合声音中检测和提取目标语音, 如图 1 所示. 这尽管在人类语音交流中非常轻松自然, 但对于语音处理系统, 则必须同时区分环境声音/自身扬声器声音、语音/非语音、目标语音/无关语音, 并抑制各种干扰和畸变. 这些问题在传统语音处理系统中从未同时出现. 为对其建立统一的应对方案, 本文引入了人类听觉机制的选择注意机理 (auditory attention) [1]. 相对来讲, 人类听觉之所以能处理多声源和有混响的问题, 甚至还能在多人说话时检测和跟踪自己感兴趣的语音 [2], 主要因为人类听觉具有特定的选择注意能力. 当人类对某种目标声音感兴趣时, 能够根据具体任务和环境, 选取目标语音与周围声音最有区分性的特征, 并根据先验知识进行比对和筛选, 排除干扰声音从而获得目标语音.

使目标语音区分于其他声音的特征有很多, 而要充分利用这类特征进行检测, 则需优先考虑先验知识最多和最可靠的特征. 例如, 当扬声器播放声音时, 与扬声器声音相关的声音都可以认为是回声干扰. 如果目标语音的语义已知, 那么语义就是明显的区分性特征; 如果目标语音的声波到达方向 DOA

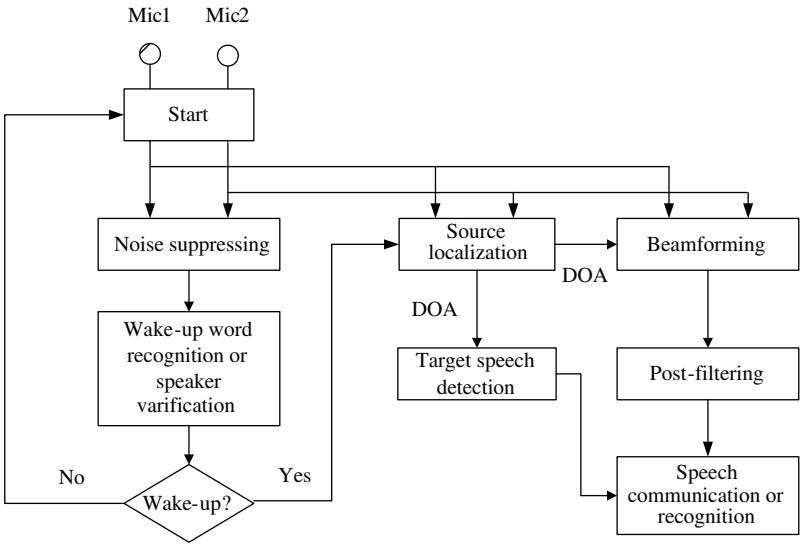


图 2　具有选择注意能力的目标语音拾取算法框图

Figure 2　Algorithm diagram of speech picking for speech systems with auditory attention ability

(direction of arrival) 已知, 那么通过检测 DOA 信息可以去除大量无关声音. 通过对各种区分性信息的检测和比较, 最终可以抑制无关声音的影响, 并从混合声音中筛选出目标语音段.

为了获得目标语音, 我们构建了如图 2 所示的远讲语音拾取系统. 它包含了语音唤醒技术、传声器阵列技术、语音增强技术等多种功能, 目的在于获得更加纯净的目标语音. 后文将对每一个关键技术点作详尽地综述, 并针对若干关键问题提出了改进方案.

2　传声器阵列

在近讲模式下, 由于用户与传声器非常接近, 仅用一个传声器即可保证一定的信噪比, 对信号的处理在时域和频域完成. 但若采用较自然的拾音方式, 为应对噪声、混响和多说话人等问题, 往往需要引入声场的空间信息. 为此, 一般将一组传声器按照特定拓扑结构排布在空间中, 并进行同步采音, 称为传声器阵列 [3], 这是拾取声场空间信息的技术基础. 利用传声器阵列, 可以对来自不同角度的声音进行相应的处理, 如增强或抑制指定方向的声音, 或进行声源定位和语音分离.

传声器阵列中的各个传声器接收到的声波在时间、相位和强度上都有微小差距, 而传声器阵列算法则通过对它们的比较, 获得声源的空间信息. 以阵元之间的距离为参考, 与传声器阵列较近的声源可称为近场 (near-field) 声源, 此时须将声波建模为球面波, 并可利用声波的声强信息; 与传声器距离较大的声源则称为远场 (far-field) 声源, 所发射的声波可以看作平面波, 声波到达传声器的的声强差也可以忽略.

传声器阵列拾取声场空间信息的能力取决于 3 个方面: 拓扑结构、阵列孔径和一致性. 拓扑结构包括传声器的几何排布及相互间距离. 典型的几何排布包括线阵、圆阵、平面阵、球阵等, 它们影响到阵列对空间区域的划分. 例如, 线阵只能在平面内 180° 范围内区分声源, 平面阵只能对来自平面一侧的声源进行区分. 阵列的规模是指传声器的个数, 最小的阵列仅有两个传声器, 大规模的阵列则可能包括上千个传声器. 阵列规模越大, 在空间中的采样点就越多, 所以能够采用更复杂和精密的算法并

取得更佳性能, 而相应的代价是制造成本和耗电量都比较大. 如果规模和阵形都已确定, 那么空间采样的分辨率取决于相邻传声器之间的距离. 根据空间采样定理, 当声波半波长小于两阵元间距时, 会出现 “空间混叠” 现象. 所以, 如果阵元间距较大, 尽管分辨的精度更高, 但高频更易出现混叠; 反之, 较小的阵元间距尽管能够处理更高频的信号, 但精度又比较低. 因此, 传声器的拓扑需要根据实际情况进行折中, 合理配置阵元数量、间距和阵形. 在实际使用中, 硬件的特性也非常重要, 它包括每个传声器频响特性、各传声器频响是否一致、采样的同步性等, 这往往对阵列算法有关键影响.

3 目标语音检测

在智能语音系统中, 如何从连续音频信号里检测到目标语音存在的时间段, 不仅关系到语音处理的速度, 更决定了系统的可靠性甚至可用性. 例如, 由于大多数语音增强方法是基于检测估计机制, 准确的目标语音检测有助于语音信号的增强处理; 语音识别同样需要避免无关声音的干扰, 否则会产生错误的识别结果, 并导致系统功能异常.

长期以来, 由于采用近讲模式, 目标语音检测可以通过时频统计特征实现, 称为话音活动检测 VAD (voice activity detection). 如果以自然方式进行语音拾取, 则需要结合具体情况, 选用不同的先验信息. 例如, 如果房间内没有方向性干扰声源, 可以直接根据声源的方向性进行检测 [4]; 如果声音的波达方向已知, 则可根据 DOA 进行检测 [5~7]; 如果说话人身份特征已知, 则通过区分说话人实现目标语音检测 [8,9]; 如果对话音的语义内容或领域有先验知识, 则可以通过语义信息实现检测 [10,11], 这种根据语义检测目标语音的方法, 目前已经在语音唤醒方面取得了实际应用.

如果语音唤醒获取目标语音的起点, 则必须进一步提取目标语音的其他区分性特征, 才能完成目标语音检测. 所选用的特征需要根据具体应用的环境条件来选取. 例如, 利用传声器阵列取得目标语音的 DOA, 或提取目标说话人身份特征等, 都可在语音通信和识别开启后用于目标语音检测; 而语义的领域和语音关键词等信息则有助于检测目标语音的终点. 结合功能和目标选用适当的区分性特征, 并不断丰富和更新先验知识, 这正体现了听觉的选择注意特点 [12].

为了得到复杂环境中更加准确的目标语音信号检测性能, 本文提出了一种在信号处理层利用时频信息和空间信息获得目标语音信号出现区间的方法. 作为基本的拓扑结构, 双通道方法通常具有组合灵活、低开销这样的特点. 我们提出了一种基于波束参考能量比 BRR (beam-to-reference ratio) 的混响鲁棒的双通道目标语音信号检测方法. 为了使目标语音信号检测方法适应不同的混响环境, 一种基于两通道相干分析的直达声混响声能量比 DRR (direct-to-reverberate energy ratio) 估计方法被引入进来. 从空域滤波的角度分析了时频域理论上的基于 BRR 的目标语音信号检测阈值, 并且将其从自由场假设推广到一般的混响场假设, DRR 作为这样一个参量被带入推导当中. 在此基础上, 针对小间距传声器阵列的空间混叠问题造成的检测虚警, 提出一个新的旁瓣抑制机制对该问题加以解决 [13].

利用 ROC (receiver operating characteristics) 曲线分析所提出的目标语音检测方法. ROC 曲线是在虚警率 FAR (false acceptance rate) 和漏警率 FRR (false rejection rate) 这两个矛盾对立面之间寻求一种平衡. 其中, 语音段和非语音段采用人工标记.

数据库是在实际的房间中录制, 双传声器间距 $d_{\mathrm{mic}} = 8.5$ cm, 房间配置如图 3 所示. 目标语音信号和干扰信号使用专业级扬声器播放. 其中, 目标语音信号包含 64 句汉语句子, 分别来自于两男两女, 每人 16 句话. 通过调整扬声器的音量, 目标语音信号和环境平稳噪声信噪比约为 5 dB (在扬声器处), 基本上是实际场景中信噪比. 与目标语音信号内容无关的干扰信号被放置在目标语音区域外的扬声器播放, 根据播放时的音量, 目标信号与干扰信号能量比被分别控制在 5 dB, 0 dB, −5 dB, −10 dB (目

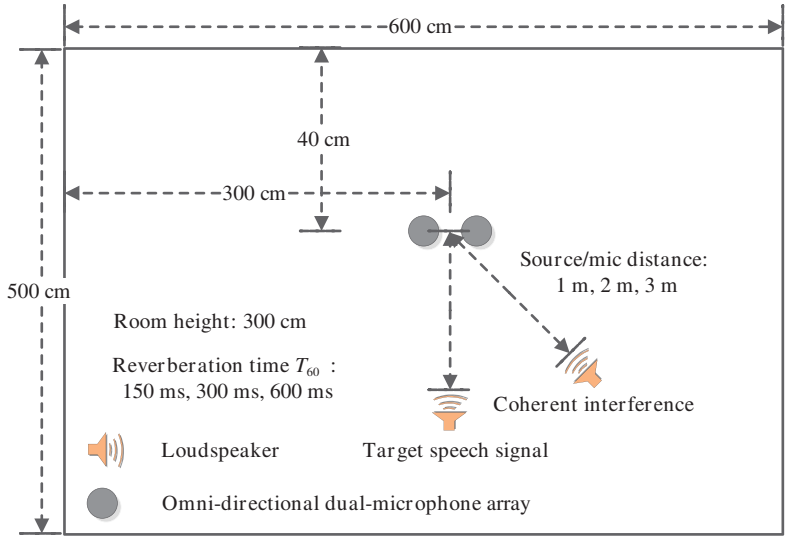


图 3　目标语音检测实验房间配置

Figure 3　Experimental room configuration of target speech detection

表 1　每个实验集中目标语音信号和相干干扰信号 DOA 方向描述

Table 1　Description of DOA distribution on both target speech signal and coherent interference signal in each experimental set

DOA of target speech signal (°)	DOA of interferant signal (°)
0	30, 45, 60, 90
30	0, 60, 90
45	0, 30, 90
60	90

标干扰同时出现时).

通过数据采集构建了 18 种数据库用来测试所提出的目标语音检测算法和对比算法. 这 18 个数据库包含 3 种房间类型 (T_{60} 分别为 150, 300 和 600 ms), 3 种目标到传声器阵列的距离 (1, 2, 3 m), 以及两种类型的相干噪声 (来自 TIMIT 数据库的语音和小琴音乐). 采音硬件采用 B&K 型号 3050-B-060 多通道采集器和型号 4958 传声器, 录音采用 24 bit 量化和 32 kHz 采样率. 最终 18 个数据库的录制音频被降采样到 16 bit 量化 16 kHz 采样率.

为了验证算法的鲁棒性, 我们录制了来自阵列正前方的目标语音信号, 也录制了来自其他目标方向的目标语音信号的情况. 表 1 给出了目标语音信号和干扰信号的 DOA 分布, 在 18 个数据库中, 每个数据库包含了 11 种情况.

如下几种算法用来同所提出的目标语音信号算法作性能比较, 包括经典的单通道方法 AMR2 [14]、基于 DOA 一致性 (DOA homogeneity) 的双通道方法 [15] 以及 Cohen 提出的双通道目标语音信号检测方法 [16]. 本文同时给出了未加旁瓣抑制 (标记为未加 SS) 和加旁瓣抑制 (标记为加 SS) 过程的目标语音信号检测性能.

该实验是完全仿真实际中远讲的情况, 例如家庭场景. 在两种噪声分别作为干扰的场景下, AMR2 方法都表现出了很高的虚警率, 这是因为 AMR2 方法是建立在单声源假设下的. 同样在这两种场景

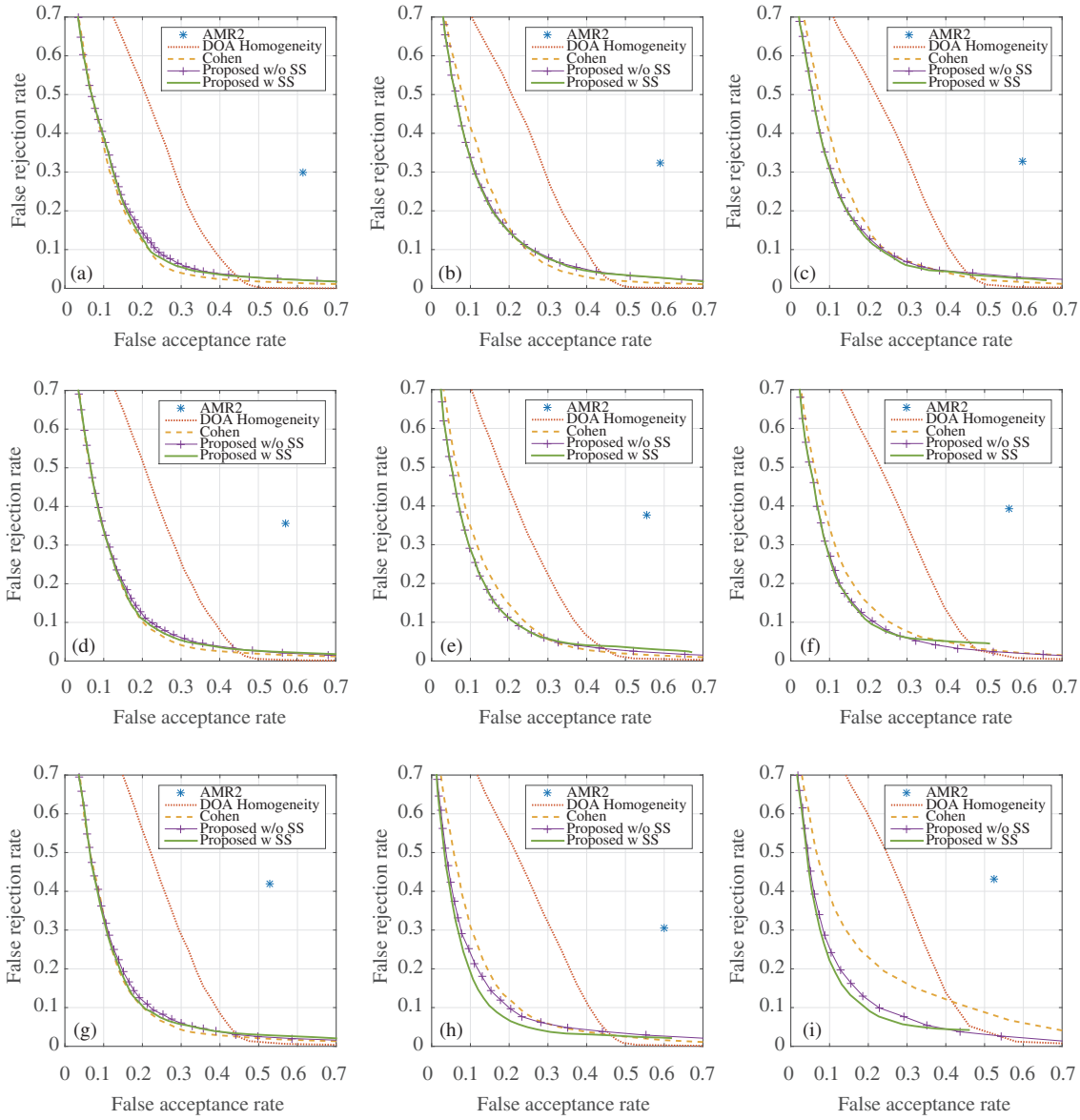


图 4 相干干扰声源为语音时的 ROC 曲线

Figure 4 The ROC curves when coherent interference is speech. (a)∼(c) Target source/mic distance=1 m, $T_{60} \approx 150$, 300, 600 ms; (d)∼(f) target source/mic distance=2 m, $T_{60} \approx 150$, 300, 600 ms; (g)∼(i) target source/mic distance=3 m, $T_{60} \approx 150$, 300, 600 ms

下, 尽管利用了空间信息, 由于单声源的假设, 基于 DOA 一致性的方法没有表现出很好的性能, 所有的图都表明该方法不像 Cohen 提出的方法以及本文提出的方法具有更好地虚漏警表现.

众所周知, 随着房间混响时间的增大或者目标声源到传声器距离的增加, 或者二者皆增加, 混响对目标语音的影响是在增大的. 图 4 给出了以语音作为干扰的目标语音检测性能, 9 幅图表明本文提出的方法对于不同环境不同距离性能保持了稳定. 在混响时间 $T_{60} \approx 150$ ms 的房间中, 声学环境相对理想, 接近理想自由场, 所提出的未加 SS 过程和加 SS 过程的目标语音检测方法同 Cohen 提出的方法性能基本相当. 但是, 随着混响时间增加 ($T_{60} \approx 300, 600$ ms) 并且声源到传声器距离增加 (从 1 m 到

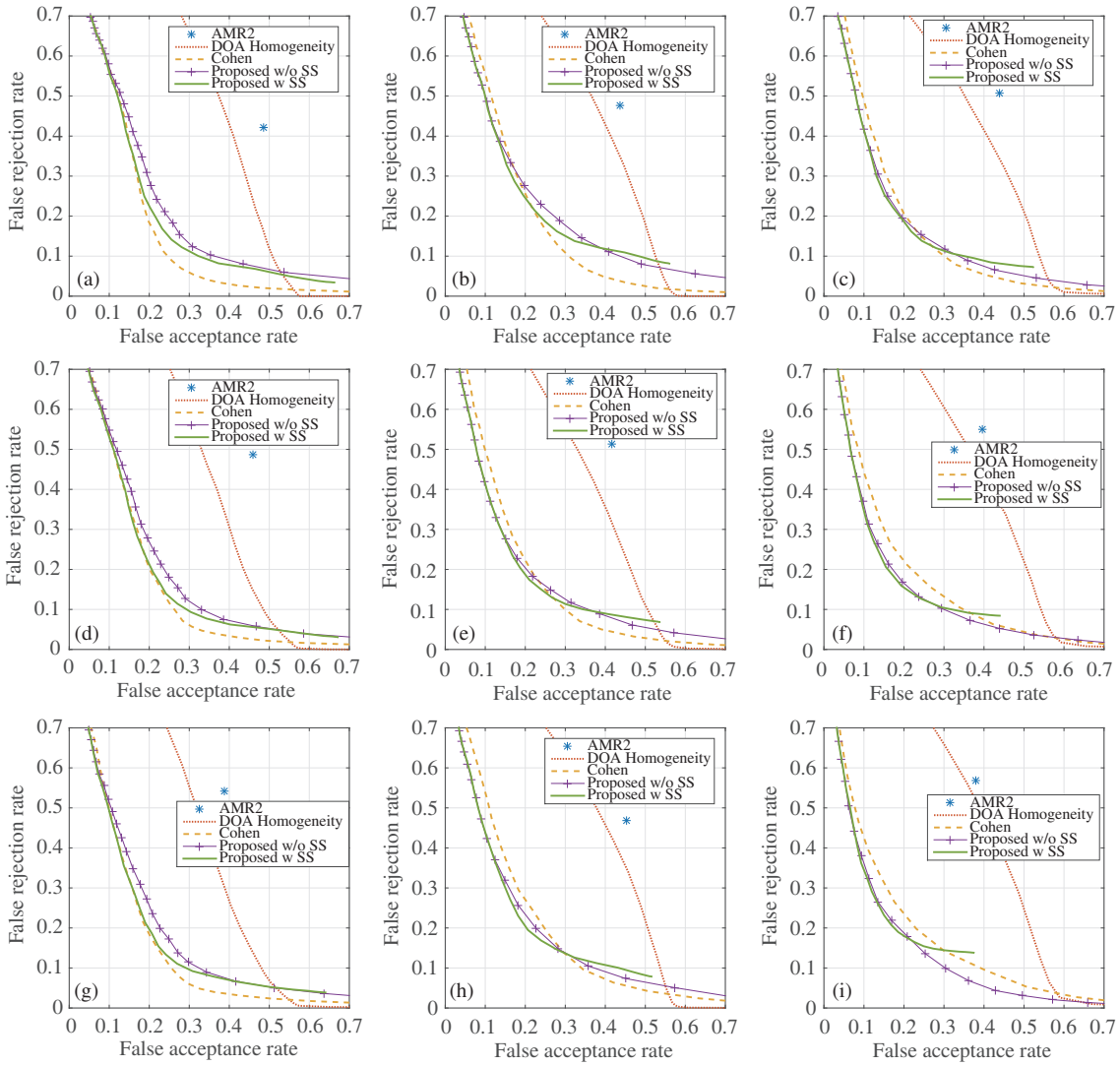


图 5　相干干扰声源为音乐时的 ROC 曲线

Figure 5　The ROC curves when coherent interference is music. (a)∼(c) Target source/mic distance=1 m, $T_{60} \approx 150$, 300, 600 ms; (d)∼(f) target source/mic distance=2 m, $T_{60} \approx 150$, 300, 600 ms; (g)∼(i) target source/mic distance=3 m, $T_{60} \approx 150$, 300, 600 ms

3 m), 所提出的未加 SS 过程的算法相对于 Cohen 提出的方法表现出了更加鲁棒的特性, 因为 Cohen 基于自由场假设下的条件不再满足, 而本文的方法可以自适应地根据环境调整每个时频点的检测阈值, 从而提高了估计似然的准确性. 若考虑到旁瓣对方法的影响, 当加入 SS 过程之后, 所提出的目标语音检测方法, 表现出了更好的混响鲁棒性.

图 5 表明, 当干扰信号为音乐时, 可得到同干扰信号为语音时同样的趋势. 但值得一提的是, 当音乐作为干扰时, 目标语音检测的整体性能相对语音是有所下降的, 这主要是因为从全带角度看, 不再具有观测信号只来自一个声源的语音这样的假设, 但是在每个时频点, 依然假设该时频点是被某个声源所主导的, 当音乐作为干扰信号出现时, 由于其谐波结构相对于语音更加复杂, 这种假设带来的副作用会被加剧.

通过上述实验, 可以探究混响对于目标语音检测造成的影响. 由于除了直达声之外, 来自于房间墙壁、地面以及房间陈设的反射会被传声器同时采集, 这些混响声会模糊掉原本较为清晰的空间区分特性, 从而也就降低了目标语音信号和干扰信号的区分特性. 在目标语音检测算法当中, 固定阈值一定是不能满足这种复杂环境下的需求的, 本文的一个重要出发点即为寻得某种环境学习机制, 使得方法能够克服混响带来的种种问题, 而 DRR 作为这样的一个参数, 在所提出的方法中发挥了重要的作用. 图 4 和 5 同样通过鲁棒的 ROC 曲线印证了这一点.

4 语音唤醒

语音唤醒也称为激活词语音识别技术, 是通过识别和监测特定的语音内容, 实现语音控制系统的智能开启, 从而消除无关语音的干扰. 通过为每个语音系统指定特定的唤醒词 (wake-up-word), 当用户需要启动该语音系统时, 只要说出特定的唤醒词, 就可以使该系统处于激活并接收语音的状态, 而其他的系统则不会受其影响. 引入语音唤醒技术, 能够利用目标语音中先验的语义信息, 有效排除无关语音的干扰, 准确发现目标语音的起始点.

目前激活词语音识别技术主要分为两类: 一种为 DTW 模式的模板匹配方法, 另一种是基于统计模型的语音识别方法 [10]. 基于 DTW 的方法计算复杂度低, 在限定说话人和说话内容的前提下具有一定的可用性, 但是它受到模板选择的限制, 鲁棒性和推广性不太好. 基于统计模型的语音唤醒系统采用了语音识别的框架. 首先利用大量语音训练声学模型, 然后采用识别解码策略对激活词内容进行检测并计算置信度, 从而完成分类判别. 也可以将语音识别和关键词检索算法相结合 [11,17~19], 实现对唤醒词的检测. 这种方法因为采用了描述能力较强的声学模型, 所以不需要限定说话人, 还可以更换激活词内容, 即使使用环境发生一定的变化, 系统也能保持不错的效果. 因此, 这类基于语音识别的方法更为稳定并容易扩展, 也受到更多关注. 例如, 通过在前端进行预筛选, 可以剔除非正常语音片段 [20], 能够减轻后端识别压力, 提高实现效率; 对语音识别的声学特征 [11,17~19] 或韵律特征 [21] 进行改进, 也可以用于改善性能.

目前, 语音唤醒技术已经在某些手机客户端上取得了实用, 并已出现了说话人相关和说话人无关的英文激活识别芯片, 但它们主要适用于距离较近且背景安静的条件. 在国内的某些智能家电中, 也已经应用了远讲模式的语音唤醒技术. 随着穿戴式设备的不断涌现, 先进行语音唤醒继而开启语音识别等交互功能可能会成为未来语音应用的常规模式.

5 混响抑制

当声音信号在封闭空间传播时, 不仅会不断衰减, 而且会因墙面和物体等障碍物的反射而出现多径传播现象. 如果传声器距离声源较远, 则拾取得到的声音中往往会含许多经过延迟和衰减后的声波. 其中直接到达的声波称为直达声, 到达较早且能量较强的反射声称为早期反射声, 而到达较晚, 能量较小且持续时间较长的反射声则称为晚期反射声. 早期混响主要带来频谱的染色效应. 而晚期混响主要引起自我掩蔽和重叠掩蔽.

对语音来说, 一定的混响不仅有利于改善听感, 而且可以提高语音识别的正确率; 但当语音受混响影响较大时 (主要反映在房间混响时间 T_{60} 的增大和说话人/传声器距离的增加), 混响会导致语音信号时间和频谱上的拖尾效应, 会同时改变语音信号的包络和精细结构, 从而严重影响语音信号的质量和可懂度. 一般用房间声学冲激响应 (room impulse response, RIR) 来表征声波传播到传声器时受

混响影响的情况. RIR 往往随着房间内物体的移动而不断改变, 因此无法先验确知. 同时, 由于语音源信号也非先验可知, 因此, 要在拾取到的信号中消除混响影响, 是一个比较困难的盲问题. 目前的研究主要包括 3 种思路: 1. 波束形成方法; 2. 盲系统辨识的方法; 3. 谱增强的方法.

5.1 波束形成方法

波束形成方法是利用传声器阵列对声波的定向选择作用, 对来自声源方向的声波进行增强, 而抑制其他方向的声波. 由于反射声的波达方向一般与声源方向不同, 这种方法能够有效抑制混响的影响. 同时, 这类方法还能抑制来自非目标方向上的噪声, 所以非常适用于混响环境中的降噪处理. 波束形成方法可以分为固定波束形成和自适应波束形成, 因为混响具有一定扩散性质, 所以固定波束形成更适合用于混响抑制. 最简单的固定波束形成方法即为延迟相加波束形成 (delay-and-sum beamformer), 它对传声器信号进行时延补偿, 继而对多路信号进行有权重相加, 从而重构相干信号 (直达声), 而不相干信号 (混响或扩散噪声) 则受到抑制 [22]. 除延迟相加波束形成外, 还有若干具有固定指向性的固定波束形成方法 [23, 24], 能够在一定情况下更好地抑制混响的影响.

5.2 盲系统辨识和逆滤波方法

混响语音信号可以看作语音源和房间冲激响应的卷积, 如果能够估计房间冲激响应, 并对观测信号进行逆滤波, 则在理论上能够完全消除混响. 为此, 有一类混响抑制方法通过盲系统辨识估计房间冲激响应, 并以此消除混响.

常见的多通道盲系统辨识方法利用了二阶统计量, 即基于通道间的互相关估计房间冲激响应. 这类方法包括基于最小二乘估计的盲系统辨识方法 [25]、基于空间分解的方法 [26] 以及多通道 LMS 和 Newton 自适应方法 [27~29]. 但是, 这些盲系统辨识方法都建立在以下两个基本假设之上, 即: 1. 声源到不同传声器之间的传递函数不存在共同的零点; 2. 源信号的相关矩阵是满秩的.

由于上述假设并非总是成立, 所以出现了基于多通道线性预测 (linear prediction) 的方法. 它们使用白化的输入语音信号, 继而通过多通道线性预测方法来辨识系统 [30~32]. 但是这类方法同样面临一定问题. 首先, 房间冲激响应通常是时变的, 因此需要系统辨识必须是自适应完成的; 其次, 房间冲激响应的阶数通常很长, 因此需要精度很高的鲁棒算法, 并要求设备能够适应很高的计算复杂度. 更关键的是, 这类系统辨识的鲁棒性会受到噪声的影响, 而在实际环境中, 噪声其实是无法避免的.

当已知声源到传声器的声学冲激响应时, 混响抑制可以通过通道求逆得到, 但由于声学冲激响应不是最小相位系统, 零极点都不在单位圆内, 直接求逆通常会面临稳定性问题. 在单通道求逆的方法中, 通常将声学冲激响应分解为最小相位系统和全通系统的级联. 而在多通道求逆方法中, 可以考虑 MINT 方法 [33] 实现准确的逆滤波, 其代价是较高的滤波器阶数及复杂的矩阵运算. 为降低滤波器阶数, 可以采用基于子带和自适应的 MINT 方法 [34]. 同时, 为满足噪声环境中和估计误差下的鲁棒性, 还可以考虑多通道均衡方法 [35].

尽管完全的去混响在理论上可以通过盲系统辨识和盲均衡实现, 但在实际应用当中, 由于房间冲激响应的时变性及噪声干扰的影响, 该类方法稳定性不佳且计算量巨大, 因此具有很大的局限性.

5.3 谱增强的方法

谱增强是噪声抑制的常用方法, 通过对噪声频谱的估计, 可以在频谱上抑制干扰并提升语音. 如果将混响看作一种噪声干扰, 则可将该方法用于混响抑制, 从而去除混响成分并改善语音质量.

最早的谱增强方法是在倒谱域展开的 [36], 它认为反射声信号会在倒谱域表现出一定的峰值, 因此, 可以利用峰值检测方法抑制反射声造成的峰值, 从而消除混响成分. 但是这种方法只适用于简单的混响模型. 为此, 很多方法基于时空联合的多通道线性预测方法, 在残差域抑制混响, 并通过自回归系数 (AR 系数) 恢复纯净语音信号 [37~39], 但是该类方法的基本假设是混响语音信号同纯净语音信号的 AR 系数相同, 该假设在环境中存在噪声时往往无法满足.

通过将单通道语音降噪中的谱减法进行扩展, 可以实现单通道混响抑制 [40], 并可进一步扩展为多通道混响抑制 [41]. 其中, 混响声的功率谱可通过广义统计混响模型估计获得. 混响声一般用指数衰减模型建模, 其衰减因子是由房间的混响时间 (T_{60}) 决定的.

波束形成方法由于受扩散性质混响的影响, 只能提供有限的信混比增益; 盲系统辨识方法受计算复杂度和噪声鲁棒性影响, 在实际中受到局限; 相对来讲, 由于同噪声抑制的框架具有一致性, 谱增强方法体现了很大的优势. 但是, 单纯的谱增强方法往往存在语音失真问题, 需要在语音失真和混响抑制程度之间进行折中 [42]. 为此, 可以根据对声学场景的自动辨识 [43], 适当选用不同的语音谱增强策略, 能够有效兼顾语音保真, 噪声抑制和混响抑制, 并保持较低的计算复杂度.

6 语音增强

语音增强是一种语音处理技术, 它根据输入的声音信号, 提升其中的目标语音成分并抑制噪声成分, 达到改善听感和识别率的效果. 结合人类听觉注意机理, 要改善语音增强性能, 同样应尽可能利用目标语音的先验知识.

根据采用的传声器数量, 语音增强可分为单通道方法和多通道方法. 其中单通道方法能利用的先验知识较少, 主要适合处理平稳背景噪声或噪声特征已知的情况, 而多通道方法则可以利用目标语音的 DOA 特征, 从而抑制来自非目标方向的非平稳噪声. 另外, 在用扬声器播放声音的场景中, 如果能够获得声学回声的参考信号, 则可直接用该信息消除回声的干扰.

6.1 单通道语音增强

单通道语音增强主要利用语音和噪声在时频域分布的差异而实现噪声消除. 根据是否采用数学模型对语音机噪声进行描述, 单通道语音增强可以分为参数类方法和非参数类方法两类. 其中, 参数类方法以数学模型 (如 AR 模型) 描述语音及噪声信号, 再根据带噪信号对模型中的参数予以估计, 最后构建算法 (如 Kalman 滤波器) 对带噪信号进行处理. 非参数类方法致力于估计语音以及噪声信号的统计特性, 借以恢复目标信号. 在现有的单通道语音增强方法中, 非参数类方法占据了主导地位.

在非参数类方法当中, 最早出现的是谱减法和 Wiener 滤波的方法, 此后出现了基于统计模型的方法. 典型的基于统计模型的方法包括基于 Gauss 模型假设下的短时幅度最小均方误差 (MMSE) [44] 和短时对数幅度估计子 (MMSE-LSA) [45] 的方法. 在此基础上, 通过对每个时频点的语音信号建立存在概率模型, 可以对 MMSE-LSA 的算法 (OMLSA) [46] 进行改进. 此后, 人们用 Gamma 模型和 Laplace 模型等分布模型取代 Gauss 模型, 以更加符合语音的先验分布, 改善语音增强的性能.

近些年来, 基于最优化模型的算法逐渐受到重视, 如最小方差无失真相应 (MVDR) 算法等. 在 MVDR 算法中, 假设连续若干帧的目标信号的帧间自相关矩阵是稳定的, 并且可以被有效实时估计 [47]. 这类方法有助于减小语音失真.

单通道语音增强的两个核心问题是噪声估计和先验信噪比估计, 前者是降低噪声的关键因素, 而

后者则关系到残留 “音乐噪声” 的程度. 单通道增强算法在很多情况下能够显著提高信噪比, 尤其对平稳噪声 (白噪声、车噪等) 有较好的消除效果. 但对于非平稳信号, 由于无法可靠地估计噪声和信噪比, 单通道语音增强的效果还不够理想. 与此同时, 研究表明单通道语音增强算法并不能有效地提高语音可懂度. 因此, 引入声源的空间分布特征, 采用具有空间选择性的多通道语音增强算法, 能够更有效地消除非平稳噪声干扰.

6.2 多通道语音增强

多通道语音增强利用了传声器阵列拾取空间信息的能力, 可以结合时域、频域以及空间信息, 获得带有空间区分性的接收能力. 根据不同的处理原理, 多通道语音增强方法可以分为 3 类: 波束形成及后滤波、基于听觉场景分析的方法、基于盲源分离的方法.

6.2.1 波束形成及后滤波

波束形成算法又被称为空域滤波, 传声器阵列的空间选择性集中体现于此. 这一族算法大致可分为两类, 即固定波束形成 (fixed beamforming) 和自适应波束形成 (adaptive beamforming).

固定波束形成算法使用一组经优化的滤波器以增强处于某特定方向 (位置) 的声源, 而同时尽可能地抵制来自其他方向 (位置) 的声源, 达到提高信噪比的效果. 典型的固定波束形成算法有延迟相加、超指向性等 [22]. 固定波束形成的滤波器系数需要在使用前进行设定, 且不随时间或输入信号的变化而变化, 因此, 当声学环境复杂而多变时, 它对噪声的抑制能力往往不够充分.

自适应波束形成的滤波器系数随输入数据的变化而发生改变, 从而能适应时变的声学环境, 得到更好的结果. 其中, 线性约束最小方差算法 (linearly constrained minimum variance, LCMV)[48] 是最早的自适应波束形成算法. 此后的广义旁瓣抵消算法 (generalized sidelobe canceller, GSC) [49] 将 LCMV 中的带约束条件的优化问题转化为无约束优化问题, 不仅实现简单, 而且效果显著, 因此获得了广泛应用. 然而, 在干扰噪声源的数目多于所用传声器数目的情况下, 特别是在扩散噪声场和混响环境中, 自适应波束形成器的性能会明显下降. 另外, 在非平稳干扰和瞬态噪声存在的环境下, 自适应波束形成器中自适应滤波器的收敛也会面临问题, 相关的改进研究 [50, 51] 虽然取得了一定进步, 但鲁棒性和滤波器收敛问题仍未获得彻底解决.

波束形成能够抑制方向性干扰, 但要进一步提高信噪比增益, 常常需要接后置滤波器, 进一步抑制残余噪声. 常见的后置滤波包括基于对通道间噪声相关性的后置滤波器 [52, 53] 以及基于目标语音信号检测的方法 [54]. 其中, 基于目标语音检测的方法通过估计目标语音信号存在概率, 构建每个时频点的增益函数, 能够显著提高非平稳噪声干扰下的信噪比, 但存在语音失真的问题. 将鲁棒 GSC 和后滤波方法结合起来, 基于子带后滤波反馈控制进行多通道增强 [42], 则不仅能保证自适应波束形成的鲁棒性, 而且提供了足够的信噪比增益.

6.2.2 基于听觉场景分析的语音增强

听觉场景分析源自英国心理学家 Cherry 在 1953 年的发现, 即人类听觉系统能够从复杂的混合声音中有效地选择并跟踪某一说话人的声音, 这称为 “鸡尾酒会问题 (cocktail-party problem)” [2]. 即使在极其嘈杂的环境中, 人们仍然能够专注并理解所感兴趣的目标语音信号, 而忽略或抑制其他干扰噪声信号. 这个过程被称作听觉场景分析 (auditory scene analysis, ASA). 人类听觉在噪声环境中对目标语音的加工过程, 利用了多种特征信息 (如语音的基频、谱包络的连续性和声源的位置信息). 对听觉

场景分析过程进行建模, 并用计算机来模拟实现, 即计算听觉场景分析 (computational auditory scene analysis, CASA). CASA 致力于将声场中的各声源相互分离, 最初的单通道 CASA 主要利用语音的基频特征进行语音信号的分离和增强; 此后出现的双耳 CASA 则可以同时利用声源的方位信息, 能够更好地跟踪、分离或增强 [55] 目标语音.

6.2.3 盲源分离

盲源分离算法 (blind source separation, BSS) 与上述两种算法有较大差别 [56]. 之前介绍的阵列算法中, 阵列拓扑结构和目标声源的位置信息往往是已知的. 盲源分离则不依赖这些先验信息, 只考虑信号源彼此独立的假设条件, 利用所接收到的多通道信号进行分析处理, 对各个信号源信号进行分离. 尽管这类算法近年来引起了较多的关注, 但距离实际应用仍然有一定距离. 首先, 该算法假设噪声和干扰是统计独立的, 而这一点在实际环境中很难得到满足; 其次, 在算法的具体实施中, 会受到诸如运算量, 分离后排列不确定等问题的困扰.

6.3 声学回声消除

当语音系统用扬声器播放声音时, 发射出的又会被传声器采集回来, 并成为噪声干扰, 这称为 “声学回声” (acoustic echo), 它会严重降低语音拾取的质量. 以双端语音通信系统为例, 如图 1, 远端说话人的语音在本地经扬声器播放, 通过房间传播后被传声器接收, 然后和本地传声器采集的近端说话人的语音一起送回到远端, 这样, 远端说话人就会听到自己的回声. 为此, 利用声学回声消除 (acoustic echo cancellation, AEC) 技术, 通过对扬声器和麦克风之间的回声传播路径 (loudspeaker-enclosure-microphone, LEM) 进行系统辨识, 自适应地估计回声信号, 从而消除近端传声器中的回声成分, 提高语音质量.

但是, AEC 技术中的系统辨识问题存在以下几个难点: 1. 对 LEM 路径进行充分地建模通常需要数千阶的滤波器, 计算量较大; 2. 自适应滤波器的输入信号自相关性较强, 会降低滤波器的收敛性能; 3. 当近端存在语音, 即出现 “双讲” (double-talk, DT) 问题时, 自适应滤波器很可能发散, 无法消除回声. 利用频域自适应滤波器可以较好地解决前两个问题. 一方面, 在频域利用重叠保留技术能够降低计算量 [57], 另一方面, 离散 Fourier 变换 (discrete fourier transform, DFT) 能够将输入信号去相关 [58], 从而提高滤波器的收敛速度. 为保证 “双讲” 情况下仍然能够有效地估计回声, 通常需要自适应滤波器在存在近端语音时减慢甚至停止迭代, 不存在近端语音时快速迭代以实现快速收敛. 许多文献都提出了算法来解决这个问题, 这些算法可以分为两类: 基于双讲检测 (double-talk detection, DTD) 的算法和基于变步长 (variable step size, VSS) 的算法. 基于 “双讲” 检测的算法检测出近端语音后停止滤波器迭代, 以保持对 “双讲” 的鲁棒性. 现有的 DTD 算法有能量比较法 [59,60] 和相关比较法 [61~63] 等. 但是, 除了算法各自的缺陷, 基于 “双讲” 检测的算法有一个共同的问题: DTD 固有的延迟问题. 这导致滤波器经常在 “双讲” 被检测到之前就已经发散, 因而, 基于 VSS 的算法成为研究的热点. 基于 VSS 的算法主要是通过对每次迭代的步长进行控制, 使自适应滤波器保持较好的收敛性能的同时对 “双讲” 也比较鲁棒. 一类比较经典的算法是以最小化系统误差为目标, 通过直接估计 [64,65] 或者递归迭代估计 [66,67] 的方法估计最优迭代步长. 但是, 这类方法得到的最优步长都与未知的残余回声相关, 因而这类方法的性能受到残余回声估计准确性的影响. 另一类经典算法是基于鲁棒估计量的方法. 这类方法将 AEC 中的系统辨识问题看做一个线性回归问题, 并将鲁棒统计量引入到自适应滤波理论中, 使得滤波器的自适应迭代对近端语音有较好的鲁棒性 [68~70], 这类算法在实际应用中对 “双讲” 问

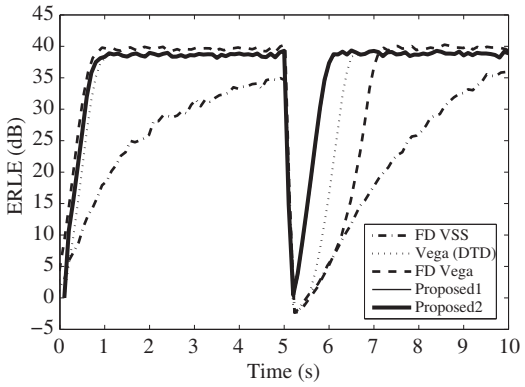


图 6 ERLE, 参考输入为白噪声, 回声路径在 5 s 发生改变

Figure 6 ERLE. The input signal is a white Gaussion noise. Echo path change takes place after 5 s

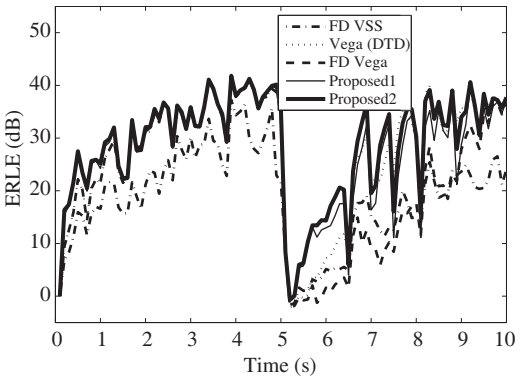


图 7 ERLE, 参考输入为语音, 回声路径在 5 s 发生改变

Figure 7 ERLE. The input signal is speech. Echo path change takes place after 5 s

题表现出不错的鲁棒性.

我们提出一种基于鲁棒统计量的频域步长控制方法, 在每个频率点上, 以最小化后验误差为准则, 对滤波器的迭代步长加以限制, 同时自适应地更新该限制参数. 跟传统的基于鲁棒统计量的步长控制算法相比, 可以在频域的各个频率点上对滤波器迭代步长施加不同的限制, 将该限制与参考信号的幅度联系起来, 滤波器迭代步长与参考信号幅度正相关, 可以有效改善滤波器的收敛性能, 并保持原有的鲁棒性. 同时, 借鉴 PNLMS (proportionate normalized least mean square) 里面更新滤波器系数的思想, 利用前一帧的滤波器系数, 对限制参数进行自适应地成比例更新, 进一步提高收敛速度.

算法的收敛性能和对近端干扰的鲁棒性通过仿真实验进行验证. 提出的算法分两种情况进行仿真, 不更新限制参数的算法记为 Proposed 1, 自适应更新限制参数的算法记为 Proposed 2. 对照算法包括: (1) 传统的基于鲁棒统计量的算法 [69], 该算法根据是否需要双讲检测器分两种情况实现; (2) 频域变步长控制算法 [64]. 首先, 比较 “单讲” 条件下在白噪声和语音作为参考信号时, 算法的收敛速度和稳态性能, 用回声返回损失增益 (echo return loss rnhancement, ERLE) 作为测度. 仿真结果分别如图 6 和 7 所示, 其中, 房间冲击响应在 5 s 处发生改变. 从两个图中都可以看到, 提出的算法和 Vega 的方法具有最好的初始收敛能力. 并且, 提出算法在房间冲激响应发生改变后的跟踪速度相比于其他算法具有明显优势. 值得一提的是, Proposed 2 比 Proposed 1 有更快的收敛速度, 这得益于限制参数的自适应更新. 其次, 比较算法在存在近端干扰时的鲁棒性, 用系统失调量 (misalignment) 作为测度. 仿真结果如图 8 所示, 在 8 s 处出现近端干扰语音. 从图中可以看到, 提出的算法无论在 “单讲” 还是 “双讲” 段, 都具有最小的系统失调, 即最好的抗近端干扰能力.

7 结论

要实现自然方式的语音拾取, 必须解决复杂环境中的各类干扰问题. 为此, 借鉴了人类的听觉注意机理, 在语音拾取的每个阶段引入了区别目标语音和干扰声音的区分性特征, 并通过对区分性特征的不断获取和更新, 模拟听觉中不断更新先验知识的过程. 其中, 最初的区分性特征可以来自预先设定的唤醒词, 此后则可以根据应用目的和场景的不同, 分别引入目标语音 DOA、说话人身份和语义等

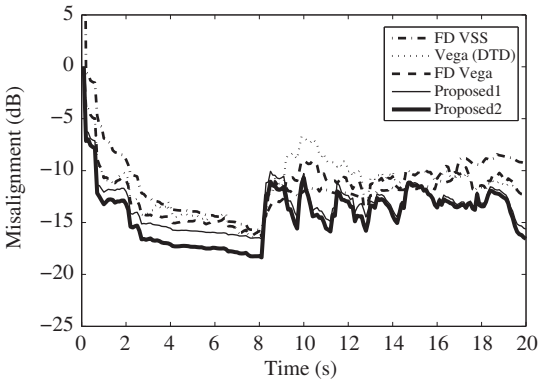


图 8 系统失调量, 参考输入为语音, 8 s 开始存在近端干扰语音, 近端干扰对参考信号能量比为 0 dB

Figure 8 Misalignment. Far-end input signal is speech. Double-talk occurs after 8 s. The near-end signal to far-end echo ratio is 0 dB

特征, 结合语音增强、混响抑制和声学回声消除等技术, 共同完成目标语音拾取, 最终实现具有选择注意能力的语音拾取.

参考文献

1 Driver J. A selective review of selective attention research from the past century. Brit J Psychol, 2001, 92: 53–78

2 Blauert J. Spatial Hearing: the Psychophysics of Human Sound Localization. Boston: MIT Press, 1997

3 Brandstein M, Ward D. Microphone Arrays. Berlin: Springer, 2001

4 Guo Y M, Li K, Fu Q, et al. A two-microphone based voice activity detection for distant-talking speech in wide range of direction of arrival. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Kyoto: IEEE, 2012. 4901–4904

5 Kim H D, Komatani K, Ogata T, et al. Two-channel-based voice activity detection for humanoid robots in noisy home environments. In: Proceedings of IEEE International Conference on Robotics and Automation. Pasadena: IEEE, 2008. 3495–3501

6 Ishizuka K, Araki S, Kawahara T. Speech activity detection for multi-party conversation analyses based on likelihood ratio test on spatial magnitude. IEEE Trans Audio Speech Lang Process, 2010, 18: 1354–1365

7 Araki S, Fujimoto M, Ishizuka K, et al. A DOA based speaker diarization system for real meetings. In: Proceedings of Hands-Free Speech Communication and Microphone Arrays. Trento: IEEE, 2008. 29–32

8 Ji M, Kim S, Kim H, et al. Reliable speaker identification using multiple microphones in ubiquitous robot companion Environment. In: Proceedings of 16th IEEE International Symposium on Robot and Human Interactive Communication. Jeju: IEEE, 2007. 673–677

9 Cheng S S, Wang H M, Fu H C. BIC-based speaker segmentation using divide-and-conquer strategies with application to speaker diarization. IEEE Trans Audio Speech Lang Process, 2010, 18: 141–157

10 Lee H, Chang S, Yook D, et al. A voice trigger system using keyword and speaker recognition for mobile devices. IEEE Trans Consum Electron, 2009, 55: 2377–2384

11 Kepuska V Z, Klein T B. A novel wake-up-word speech recognition system, wake-up-word recognition task, technology and evaluation. Nonlinear Anal Theory Methods Appl, 2009, 71: 2772–2789

12 Kalinli O, Narayanan S. Prominence detection using auditory attention cues and task-dependent high level information. IEEE Trans Audio Speech Lang Process, 2009, 17: 1009–1024

13 Wang X F, Guo Y M, Wu C, et al. A reverberation robust target speech detection method using dual-microphone in distant-talking scene. Speech Commun, 2015, 72: 47–58

14 Cornu E, Sheikhzadeh H, Brennan R L, et al. Etsi amr-2 vad: evaluation and ultra low-resource implementation. In:

Proceedings of International Conference on Multimedia and Expo (ICME03). Maryland: IEEE, 2003. 841–844
15 Rubio J E, Ishizuka K, Sawada H, et al. Two-microphone voice activity detection based on the homogeneity of the direction of arrival estimates. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Honolulu: IEEE, 2007. 385–388
16 Cohen I, Berdugo B. Two-channel signal detection and speech enhancement based on the transient beam-to-reference ratio. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Hongkong: IEEE, 2003. 233–236
17 Kepuska V Z, Eljhani M M, Hight B H. Wake-up-word feature extraction on FPGA. World J Eng Technol, 2014, 2: 1–12
18 Kepuska V Z. Wake-Up-Word Recognition. Bellingham: SPIE Newsroom, 2010
19 Kepuska V, Breitfeller J. Wake-up-word speech recognition application for first responder communication enhancement. In: Proceedings of SPIE 6201, Sensors, and Command, Control, Communications, and Intelligence (C3I) Technologies for Homeland Security and Homeland Defense V, Bellingham, 2006. 62011E
20 Cho N, Kim T, Shin S, et al. Voice activation system using acoustic event detection and keyword/speaker recognition. In: Proceedings of IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, 2011. 21–22
21 Shih C T. Investigation of prosodic features for wake-up-word speech recognition task. Dissertation for Master Degree. Melbourne: Florida Institute of Technology, 2009
22 Elko G W. Microphone array systems for hands-free telecommunication. Speech Commun, 1996, 20: 229–240
23 Allen J B, Berkley D A, Blauert J. Multimicrophone signal-processing technique to remove room reverberation from speech signals. J Acoust Soc Am, 1977, 62: 912–915
24 Flanagan J L, Johnston J D, Zahn R, et al. Computer-steered microphone arrays for sound transduction in large rooms. J Acoust Soc Am, 1985, 78: 1508–1518
25 Xu G, Liu H, Tong L, et al. A least-squares approach to blind channel identification. IEEE Trans Signal Process, 1995, 43: 2982–2993
26 Gannot S, Moonen M. Subspace methods for multimicrophone speech dereverberation. EURASIP J Appl Signal Process, 2003, 2003: 1074–1090
27 Huang Y A, Benesty J. Adaptive multi-channel least mean square and Newton algorithms for blind channel identification. Signal Process, 2002, 82: 1127–1138
28 Huang Y, Benesty J, Chen J. Optimal step size of the adaptive multichannel LMS algorithm for blind SIMO identification. IEEE Signal Process Lett, 2005, 12: 173–176
29 Huang Y A, Benesty J. A class of frequency-domain adaptive approaches to blind multichannel identification. IEEE Trans Signal Process, 2003, 51: 11–24
30 Triki M, Slock D T M. Delay and predict equalization for blind speech dereverberation. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. Toulouse: IEEE, 2006. V97–V100
31 Delcroix M, Hikichi T, Miyoshi M. Precise dereverberation using multichannel linear prediction. IEEE Trans Audio Speech Lang Process, 2007, 15: 430–440
32 Kinoshita K, Delcroix M, Nakatani T, et al. Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction. IEEE Trans Audio Speech Lang Process, 2009, 17: 534–545
33 Miyoshi M, Kaneda Y. Inverse filtering of room acoustics. IEEE Trans Acoust Speech Signal Process, 1988, 36: 145–152
34 Yamada H, Wang H, Itakura F. Recovering of broadband reverberant speech signal by sub-band MINT method. In: Proceedings of International Conference on Acoustics, Speech, and Signal Processing. Toronto: IEEE, 1991. 969–972
35 Nelson P A, Orduna-Bustamante F, Hamada H. Inverse filter design and equalization zones in multichannel sound reproduction. IEEE Trans Speech Audio Process, 1995, 3: 185–192
36 Oppenheim A V, Schafer R W, Stockham T G. Nonlinear filtering of multiplied and convolved signals. IEEE Trans Audio Electroacoust, 1968, AU-16: 437–466
37 Wu M, Wang D L. A two-stage algorithm for one-microphone reverberant speech enhancement. IEEE Trans Audio

Speech Lang Process, 2006, 14: 774–784
38 Gaubitch N D, Naylor P A. Spatiotemporal averagingmethod for enhancement of reverberant speech. In: Proceedings of 15th International Conference on Digital Signal Processing. Cardiff: IEEE, 2007. 607–610
39 Nakatani T, Miyoshi M, Kinoshita K. Single-microphone blind dereverberation. Speech Enhancement. Berlin: Springer, 2005. 247–270
40 Lebart K, Boucher J M, Denbigh P N. A new method based on spectral subtraction for speech dereverberation. Acta Acust United Ac, 2001, 87: 359–366
41 Habets E A P. Single- and multi-microphone speech dereverberation using spectral enhancement. Dissertation for Ph.D. Degree. Eindhoven: Technische Universiteit Eindhoven, 2007
42 Li K. Microphone array for speech enhancement in complex environments. Dissertation for Ph.D. Degree. Beijing: Institute of Acoustics, Chinese Academy of Sciences, 2012 [李凯. 复杂环境下基于传声器阵列的语音增强方法研究. 博士学位论文. 北京: 中国科学院声学研究所, 2012]
43 Wang X F, Guo Y M, Yang X, et al. Acoustic scene aware dereverberation using 2-channel spectral enhancement for REVERB challenge. In: Proceedings of IEEE workshop on REVERB challenge. Florence: IEEE, 2014. 1–8
44 Ephraim Y, Malah D. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. IEEE Trans Acoust Speech Signal Process, 1984, 32: 1109–1121
45 Ephraim Y, Malah D. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Trans Acoust Speech Signal Process, 1985, 33: 443–445
46 Cohen I, Berdugo B. Speech enhancement for non-stationary noise environments. Signal Process, 2001, 81: 2403–2418
47 Benesty J, Huang Y. A single-channel noise reduction MVDR filter. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP). Prague: IEEE, 2011. 273–276
48 Frost III O L. An algorithm for linearly constrained adaptive array processing. Proc IEEE, 1972, 60: 926–935
49 Griffiths L J, Jim C W. An alternative approach to linearly constrained adaptive beamforming. IEEE Trans Antenn Propag, 1982, 30: 27–34
50 Hoshuyama O, Sugiyama A, Hirano A. A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters. IEEE Trans Signal Process, 1999, 47: 2677–2684
51 Gannot S, Burshtein D, Weinstein E. Signal enhancement using beamforming and nonstationarity with applications to speech. IEEE Trans Signal Process, 2001, 49: 1614–1626
52 Zelinski R. A microphone array with adaptive post-filtering for noise reduction in reverberant rooms. In: Proceedings of International Conference on Acoustics, Speech, and Signal Processing. New York: IEEE, 1988. 2578–2581
53 McCowan I A, Bourlard H. Microphone array post-filter based on noise field coherence. IEEE Trans Speech Audio Process, 2003, 11: 709–716
54 Cohen I. Multichannel post-filtering in nonstationary noise environments. IEEE Trans Signal Process, 2004, 52: 1149–1160
55 Wang D, Brown G. Computational Auditory Scene Cnalysis: Principles, Algorithms, and Applications. New Jersey: Wiley-IEEE Press, 2006
56 Brandstein M S, Ward D B. Microphone Arrays: Signal Processing Techniques and Applications. Berlin: Springer, 2001
57 Shynk J J. Frequency-domain and multirate adaptive filtering. IEEE Signal Process Mag, 1992, 9: 14–37
58 Buchner H, Benesty J, Gansler T, et al. Robust extended multidelay filter and double-talk detector for acoustic echo cancellation. IEEE Trans Audio Speech Lang Process, 2006, 14: 1633–1644
59 Duttweiler D L. A twelve-channel digital echo canceler. IEEE Trans Commun, 1978, 26: 647–653
60 Wu C, Fu Q, Yan Y H. A robust double talk detection algorithm based on noise estimation and energy ratio. In: Proceedings of National Conference on Man-Machine Speech Communication, Guiyang, 2013 [吴超, 付强, 颜永红. 基于噪声估计和能量比的双讲检测方法. 第十二届全国人机语音通讯学术会议 (NCMMSC), 贵阳, 2013]
61 Ye H, Wu B X. A new double-talk detection algorithm based on the orthogonality theorem. IEEE Trans Commun, 1991, 39: 1542–1545

62 Cho J H, Morgan D R, Benesty J. An objective technique for evaluating doubletalk detectors in acoustic echo cancelers. IEEE Trans Speech Audio Process, 1999, 7: 718–724
63 Benesty J, Morgan D R, Cho J H. A new class of doubletalk detectors based on cross-correlation. IEEE Trans Speech Audio Process, 2000, 8: 168–172
64 Shi K, Ma X L. A frequency domain step-size control method for lms algorithms. IEEE Signal Process Lett, 2010, 17: 125–128
65 Jiang K Y, Wu C, Guo Y M, et al. Acoustic echo control with frequency-domain stage-wise regression. IEEE Singal Process Lett, 2014, 21: 1265–1269
66 Kuech E M F, Enzner G. State-space architecture of the partitioned-block-based acoustic echo controller. In: proceedings of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Florence: IEEE, 2014. 1309–1314
67 Kanadi M, Akhtar M T, Mitsuhashi W. A variable step-size-based ICA method for a fast and robust acoustic echo cancellation system without requiring double-talk detector. In: Proceedings of IEEE China Summit International Conference on Signal and Information Processing (ChinaSIP). Beijing: IEEE, 2013. 118–121
68 Papoulis E, Stathaki T. A normalized robust mixed-norm adaptive algorithm for system identification. IEEE Signal Process Lett, 2004, 11: 56–59
69 Vega L R, Rey H, Benesty J, et al. A new robust variable step-size NLMS algorithm. IEEE Trans Signal Process, 2008, 56: 1878–1893
70 Wu C, Jiang K Y, Guo Y M, et al. A robust step-size control algorithm for frequency domain acoustic echo cancellation. In: Proceedings of 15th Annual Conference of the International Speech Communication Association, Singapore, 2014. 2819–2823

Speech-picking for speech systems with auditory attention ability

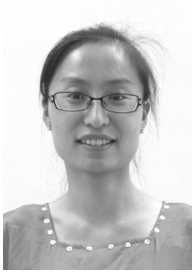
WANG XiaoFei, GUO YanMeng, GE FengPei, WU Chao, FU Qiang* & YAN YongHong

Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190*, China*
*E-mail: fuqiang@hccl.ioa.ac.cn

Abstract Currently, a natural speech-picking mode is badly needed in speech communication and in human-computer interaction systems. However, speech is usually corrupted by attenuation, multi-path propagation, and various interferences before it is received, especially when there exist several speech systems and users. It is important for practical speech systems to pick the correct speech signal within complex environments. In this paper, the mechanism of auditory attention ability is simulated through a target speech-picking system in which the a priori knowledge of the target speech and interference of sound sources are used carefully to detect and improve the target speech. The technologies of microphone arrays, wake-up-words, target speech detection, speech enhancement, and dereverberation are combined in this strategy to fulfill the task of robust target speech-picking.

Keywords microphone array, wake up word, acoustic echo control, speech enhancement, target speech detection, dereverberation

WANG XiaoFei was born in 1987. He received a Ph.D. degree in signal and information processing from the Institute of Acoustics, Chinese Academy of Sciences, Beijing, in 2015. Currently, he is an assistant research fellow at the Institute of Acoustics, Chinese Academy of Sciences. His research interests include speech enhancement, microphone array processing, and far-field speech recognition.



GE FengPei was born in 1982. She received a Ph.D. degree from the University of the Chinese Academy of Sciences, Beijing, in 2010. Currently, she is Associate Professor at the Institute of Acoustics, Chinese Academy of Sciences. Her research interests include automatic speech recognition, acoustic modeling, and wake-up-word detection.

GUO YanMeng was born in 1976. She received a Ph.D. degree in 2007 from the Institute of Acoustics, Chinese Academy of Sciences, Beijing. Currently, she is Associate Professor at the Institute of Acoustics, Chinese Academy of Sciences. Her research interests include microphone array signal processing, front-end processing of automatic speech recognition, and distant-talking speech recognition.

FU Qiang was born in 1972. He received a Ph.D. degree from Xidian University, Xi'an, in 2000. Currently, he is Professor at the Institute of Acoustics, Chinese Academy of Sciences, China. His research interests include speech analysis, microphone array processing, far-distant speech recognition, audio-visual signal processing, and machine learning for signal processing.