多说话人环境下目标说话人语音提取方案①

叶于林, 莫建华, 刘 夏

(中国人民解放军 78438 部队, 成都 610066)

摘 要: 于目标声源的方位信息与非线性时频掩蔽语音欠定盲分离方法和 BP 说话人识别技术的研究基础上, 针对现实生活中多说话人交流场景, 设计并提出了一种行之有效的解决方案, 实现了对处于任意方位的任意目标说话人语音的提取. 该方案总体上分目标语音搜索与提取两个阶段, 搜索阶段采用了 BP 说话人识别技术, 提取阶段采用了一种改进的势函数聚类声源方位信息与非线性时频掩蔽的语音欠定盲分离方法. 实验结果表明: 该方案具有可行性, 可从混合语音流中有效提取处于任意方位的目标说话人语音, 且效果较好, 信噪比增益平均为 8.68dB, 相似系数为 85%, 识别率为 61%, 运行时间为 20.6S.

关键词: 欠定盲分离; 势函数聚类; 非线性时频掩蔽; BP 说话人识别

Extraction Scheme of Target Speaker's Speech Under Multi-Speaker Environment

YE Yu-Lin, MO Jian-Hua, LIU Xia

(78438 Troops of the Chinese People's Liberation Army, Chengdu 610066, China)

Abstract: Aiming at multi-speaker communication scene in real life, an effective solution is designed and proposed based on researches of underdetermined blind speech separation method of target sound source's azimuth information and nonlinear time-frequency masking and BP speaker recognition technology, which can extract any target speaker's speech in any orientation. The solution is generally divided into two stages, one is target speech search and the other is target speech extraction. The search stage uses BP speaker recognition technology. The speech extraction stage uses the method of underdetermined blind speech separation based on sound source azimuth information by an improved potential function clustering and nonlinear time-frequency masking. The results show that the solution is feasible. It can effectively extract the target speaker's speech in any position from the mixed speech stream. The average SNRG is 8.68dB, the similarity coefficient is 85%, the recognition rate is 61%, and the running time is 20.6S.

Key words: underdetermined blind source separation; potential function clustering; nonlinear time-frequency masking; BP speaker recognition

在现实生活中, 人们无时无刻不处在伴随着大量混响、噪声、干扰、背景音乐等嘈杂声学环境之中, 比如多人在一起交流讨论、会议活动等. 在这样的复杂环境中, 对于人类而言总是能够主动地关注、跟踪、锁定感兴趣的声音信号, 并有选择性的分辨提取所需的声音信息. 人耳这种听觉选择关注现象即所谓的"鸡尾酒会效应"[1](cocktail party effect), 也叫选择性关注(selective attention), 该现象表明了人耳听觉系统对声音信号的处理机理, 同时也展示了人类语音理解机制所特有的一种感知机能, 即人耳听觉系统的强大语音辨识和抗干扰能力.

随着语音处理及计算机技术的飞速发展, 如何让计算机语音系统来模拟人耳的听觉功能, 实现智能的跟踪、识别目标说话人并提取其语音, 无疑是一项具有挑战性的研究工作. 为使得该研究工作的顺利进行, 首先我们需设计出具有科学性、可行性、可靠性的解决方案, 而方案的设计以语音分离与说话人识别技术为基础. 为此, 我们必须对语音分离与说话人识别两

① 收稿时间:2015-07-10;收到修改稿时间:2015-08-12

项技术进行深入研究. 语音分离技术目前比较成熟且国内外普遍使用的主要是盲源分离, 其研究热点为欠定盲分离[2], 它主要采用稀疏分量分析(SCA), 利用语音信号在时频域的稀疏特性并采用两步法来实现盲源分离[3-5], 文献[3-5]提出了一些解决欠定盲分离的方法, 但还存在着一定的局限和缺陷, 如计算量大、混迭矩阵盲辨识复杂、空间方向扩散等问题; 说话人识别技术目前则主要围绕语音特征参数提取与识别方法两个方面进行研究, 特征参数以基于听觉模型的语音特征为主, 包括有 LPCC、MFCC 等特征参数[6], 识别方法主要有基于 DTW、VQ、HMM、GMM、BP、深度信念神经网络等说话人识别技术[7].

本文就多说话人环境中目标说话人语音提取展开研究, 主要基于目标声源的方位信息与非线性时频掩蔽语音欠定盲分离方法和 BP 说话人识别技术, 提出了一种具有一定主动性和选择性的目标说话人语音的提取解决方案并应用于现实生活中多说话人交流场景, 实现了智能跟踪识别目标说话人并提取其语音的研究目的, 且通过仿真实验验证了所提方案的有效性. 该研究对丰富和发展计算机听觉理论及其在声源定位、语音分离、语音/说话人识别、人工智能等研究领域都具有重要的意义, 同时对人耳听觉系统的研究也有着深远的影响.

1 多说话人环境下目标说话人语音欠定盲分离及识别技术

多说话人环境下目标说话人语音提取研究, 是一项非常复杂的语音处理系统工程, 主要涉及了语音分离和说话人识别两个方面的研究内容, 本文针对多说话人语音分离采用了一种基于目标声源方位信息与非线性时频掩蔽的语音欠定盲分离方法, 目标说话人识别采用了 BP 说话人识别技术.

1.1 基于目标声源方位信息与非线性时频掩蔽的语音欠定盲分离

1.1.1 基本原理

一般情况, 麦克风接收到的干扰信号和目标语音信号来自不同方位, 具有方位信息. 其方位信息可通过麦克风间的相对时延(ITD)与声强差(IID)来表征. 由于声源方位信息为声音信号固有的一种特性, 在频域中具有聚类特性, 在进行语音分离处理时可作为一种特定参数, 对它进行聚类分析, 其后估计出混合矩阵, 为语音分离提供条件. 同时语音信号又具有时频近似稀疏性, 根据时频域单源主导的相关理论, 采用非线性时频掩蔽可实现混合语音信号的盲分离及提取. 由此本文以语音信号的方位信息、时频稀疏性及人耳听觉感知的时频掩蔽效应为理论基础, 依据 P.Bofill[8]等提出的在 SCA 条件下欠定盲分离的两步分离法, 第一步采用势函数聚类分析估计混合矩阵或每个声源的方位信息; 第二步采用非线性时频掩蔽法提取目标(某方向)的语音. 以最终实现混合语音的目标语音分离提取.

1.1.2 实现步骤

本文研究的混合语音分离模型为衰减—时延模型, 在时域上的表达式为:

$$x_i(t)=\sum_{l=1}^{n}\alpha_{il}s_l(t-\tau_{il}) \qquad i=1,2..m \quad (1)$$

式(1)中, $x_i(t)$ 为第 i 个麦克风接收到的混合语音信号, $s_l(t)$ 为第 l 个源语音信号, α_{il} 表示衰减系数, τ_{il} 表示时间延迟, $t=1,...,N$ 表示离散时间.

针对此分离模型, 本文基于双麦克风阵进行语音欠定盲分离, 如图 1 所示.

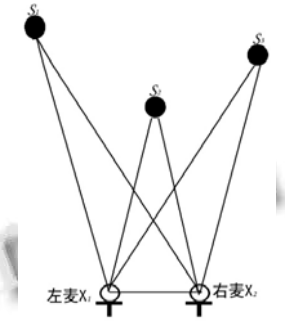

图 1 多声源到双麦克风的传输示意图

根据图 1, 对麦克风接收的混合语音信号进行短时傅里叶变换(STFT)可得:

$$X_1^k=\alpha_{1l}s_l^k e^{-j2\pi k\tau_{1l}/K}$$
$$X_2^k=\alpha_{2l}s_l^k e^{-j2\pi k\tau_{2l}/K}$$
$$\Delta H=\frac{X_2^k}{X_1^k}=\frac{\alpha_{2l}}{\alpha_{1l}}e^{-j2\pi k(\tau_{2l}-\tau_{1l})/K} \quad (2)$$

定义声源的方位信息为 ΔH, 它可通过相对幅度衰减和相对时延来表示:

$$\theta=\tan^{-1}(\frac{\alpha_{2l}}{\alpha_{1l}}),\tau=\tau_{2l}-\tau_{1l} \quad (3)$$

根据式(3), 在实际语音信号分离处理中, 声源的

方位信息可转换为计算不同声源处于不同方位而导致的不同的θ和τ. 依据以上分析, 本文混合语音分离的实现步骤如下:

① 通过双麦克风(1、2)接收并得到语音信号x_1,x_2, 对接收信号进行预处理、端点检测并对其进行短时傅里叶变换(STFT)得到频域信号$X_1(\omega_k,t)$, $X_2(\omega_k,t)$.

② 计算每个时频单元的声源方位信息, 计算公式为: $\Delta H_t(\omega_k)=\dfrac{X_2(\omega_k,t)}{X_1(\omega_k,t)}$.

③ 将一段语音的每个时频单元的方位信息$\Delta H_t(\omega_k)$转换为对应的相对幅度衰减和相对时延. 其中幅度衰减可转换成角度并限制取值范围$\theta(\omega_k,t)=\tan^{-1}(|\Delta H_t(\omega_k)|),0\le\theta\le\dfrac{\pi}{2}$; 时延$\tau(\omega_k,t)=\tau_{2l}-\tau_{1l}=\dfrac{\Delta\varphi}{\omega}$, $\Delta\varphi$表示声源每个时频单位方位信息$\Delta H_t(\omega_k)$的相位.

④ 将这一段语音的所有时频点(θ_k,τ_k)采用势函数聚类分析, 估计得到声源的个数及其各声源的相对衰减和时延(θ_i,τ_i).

⑤ 以估计得到的每个声源方位信息为基准, 通过比较各时频单元的$\Delta H_t(\omega_k)$与它的差别, 采用非线性时频掩蔽分离提取语音.

通过以上 5 个步骤可实现在欠定条件下混合语音的盲分离. 实现步骤中, 步骤 4 和 5 是关键, 体现了两步法的精髓. 本文在步骤 4 中采用了一种改进的势函数聚类算法[9], 步骤 5 中采用了非线性时频掩蔽语音分离法[10], 下面分别对其进行研究分析.

1.1.3 关键技术

1) 改进的势函数聚类算法

该算法是在原势函数的基础上改进得来, 原势函数表示为:

$$\phi(\theta,\lambda)=\sum_t l_t\varphi(\lambda(\theta-\theta(t))) \tag{4}$$

其基函数$\varphi(\alpha)$为:

$$\varphi(\alpha)=\begin{cases}1-\dfrac{\alpha}{\pi/4} & |\alpha|<\pi/4\\ 0 & \text{其它}\end{cases} \tag{5}$$

由于原势函数的聚类区间范围受限, 只能在接近于 0 到π的范围内, 且也不能同时对二维或多维数据进行聚类, 故本文对原势函数进行了改进, 使之能满足同时对本文声源方位信息的衰减和时延两个参数进行聚类.

改进的势函数聚类算法, 新势函数表示为:

$$\phi(\Delta)=\sum l^k\psi(\Delta) \tag{6}$$

式(6)中, $l^k=\sqrt{(x_1{}^k)^2+(x_2{}^k)^2}$为能量贡献因子, 一般聚类时为1; Δ为搜索变量与需要聚类数据的距离, 取值依据变量数据的维数而异, 针对二维变量数据聚类时, 以二维范数作为距离测度$\Delta=\sqrt{(\alpha-\alpha^k)^2+(\tau-\tau^k)^2}$, α^k和τ^k为需要聚类估计的两个变量(本文为衰减和时延变量), α和τ分别为α^k和τ^k的搜索变量, k为变量标记. 改进的势函数可同时对二维数据进行聚类, 扩宽了原势函数的聚类范围, 且聚类后无需配对就可得到混合矩阵.

新势函数的基函数表示为:

$$\psi(\Delta)=\frac{1}{1+(\Delta/\xi)^p} \tag{7}$$

式(7)中, 参数ξ、p为需要设置的可变参数, 本文分别设置为: 0.15, 10.

改进的势函数对二维变量聚类时, 需同时对α和τ进行搜索, 算法复杂度将呈平方增加, 计算量增大, 且耗时. 故本文采用两次聚类方式使得聚类效率进一步提升, 主要思想是第一次采用大间隔进行搜索聚类, 取值间隔越大, 精度越低, 搜索速度越快. 第一次聚类后会得到一个势函数曲面, 通过图形可直观的看出有n个波峰, 即有n个声源. 第二次采用小间隔搜索聚类, 取值是根据第一次聚类后n个波峰的坐标, 以最小间隔为搜索间隔.

$$\begin{aligned}&\alpha_1=\hat\alpha_1-k_\alpha:\lambda_{\alpha2}:\hat\alpha_1+k_\alpha,\tau_1=\hat\tau_1-k_\tau:\lambda_{\tau2}:\hat\tau_1+k_\tau\\ &\cdots\\ &\alpha_n=\hat\alpha_n-k_\alpha:\lambda_{\alpha2}:\hat\alpha_n+k_\alpha,\tau_n=\hat\tau_n-k_\tau:\lambda_{\tau2}:\hat\tau_n+k_\tau\end{aligned} \tag{8}$$

式(8)中, $(\hat\alpha_1,\hat\tau_1)...(\hat\alpha_n,\hat\tau_n)$为第一次聚类估计得到的$n$个波峰对应的值, K_α和K_τ是第二次搜索的α和τ的半径, $\lambda_{\alpha1}$和$\lambda_{\tau1}$分别为第一次聚类取值间隔, $\lambda_{\alpha2}$和$\lambda_{\tau2}$是第二次搜索取值间隔, 取n个波峰间隔对应的最小值. 同时在第二次聚类搜索时也可采用两种方式, 一是采用一维变量聚类方式, 即采用小间隔分别对α和τ进行波峰搜索聚类; 二是采用二维变量聚类方式, 即采用小间隔同时对α和τ进行波峰搜索聚类.

2) 非线性时频掩蔽分离法

在利用改进的势函数聚类估计出声源个数及混合矩阵的基础上, 可结合混合矩阵实现对混合语音的欠定盲分离, 本文主要采用了非线性时频掩蔽的分离方法. 具体实现如下:

定义每个时频点的掩蔽系数为 $M_k(\omega_k,t)$, 计算公式为:

$$M_k(\omega_k,t)=\frac{1}{1+(\alpha/\lambda)^p} \tag{9}$$

式(9)中, 参数 λ、p 为可变参数, 需要自定义, 本文分别设置为 8, 6. 参数 α 为声源某时频单元估计的方位信息与目标语音方位信息的差异. 在双麦克风阵条件, α 表示为:

$$\alpha=\frac{\left\|\Delta H_t(\omega_k)-\Delta H_0(\omega_k)\right\|}{\left\|\Delta H_t(\omega_k)\right\|+\left\|\Delta H_0(\omega_k)\right\|} \tag{10}$$

式(10)中, $\Delta H_t(\omega_k)=\frac{X_2(\omega_k,t)}{X_1(\omega_k,t)}$ 为声源某时频点估计的声源方位信息, $\Delta H_0(\omega_k)$ 为目标声源某时频点的方位信息. 参数 α 如果较小, 该时频点目标语音占主导成分, 掩蔽系数 $M_k(\omega_k,t)$ 接近 1, 反之掩蔽系数 $M_k(\omega_k,t)$ 接近 0. 利用参数 α、λ、p 就可计算出每个时频点的时频掩蔽系数 $M_k(\omega_k,t)$ 并进行时频掩蔽处理, 计算完后由 1.11 式提取出目标语音方向, 最后通过短时傅里叶逆变换(ISTFT)得到目标语音时域波形:

$$Y_j(\omega_t,t)=M_k(\omega_t,t)X_j(\omega_t,t)\quad j=1,2 \tag{11}$$

1.2 BP 说话人识别

随着说话人识别技术的研究发展, 人工神经网络方法也应用于说话人识别中, 其中 BP 神经网络[11]是目前应用最为广泛的神经网络模型之一.

BP 神经网络是一种将误差按逆方向进行传播修正的多层前馈网络, 应用在说话人识别中具有识别率高、识别时间短等优点. 网络组成的最小基本单元为神经元, 网络拓扑结构由输入层、隐层、输出层构成, 隐层可为多层; 网络学习算法主要采用 BP 算法[12,13], 基本思想是利用负梯度下降算法采取迭代运算来求解权值, 主要有前向计算(正向传播)和反向传播误差两个过程, 两个过程交替进行, 使得网络误差达到最小值, 同时保存网络的权值和偏差; 实现步骤包括网络构建、网络训练、网络识别三个步骤.

本文主要是针对多说话人环境中目标说话人语音提取方案进行研究设计, 说话人识别只是对分离语音进行目标说话人身份确认, 其技术本身不是本文研究重点, 故本文借鉴文献 13 的基本思想来实现说话人识别, 具体参数设置为: 网络结构为三层, 输入层、隐层(为一层)、输出层各层的神经元分别设置为 24、49、3; 网络训练基于遗传算法优化的 BP 算法, 学习率为 0.0001, 训练精度为 0.00001, 反向传播算法迭代次数为 1000, 网络输出准则设置为输出节点选取最大值置 1, 其它输出置 0; 语音特征参数为 24 维的差分 MFCC 特征参数, 前 12 维为静态参数, 后 12 维为动态参数.

2 多说话人环境下目标说话人语音提取方案设计

在基于目标说话人语音方位信息与非线性时频掩蔽语音欠定盲分离方法及 BP 说话人识别技术的研究基础上, 本文针对现实生活中多人谈话交流的情景, 设计并提出了一种目标说话人语音提取解决方案, 该方案总体上分两个阶段: 一是目标语音搜索阶段; 二是目标语音提取阶段.

2.1 方案设计

具体解决方案如图 2 所示.

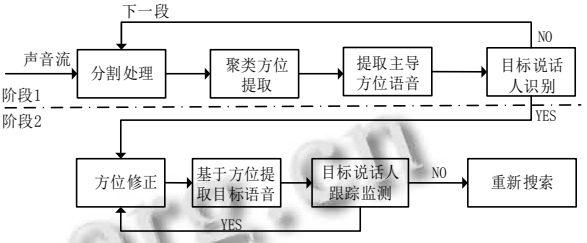

图 2 解决方案示意图

2.2 方案分析

2.2.1 实现步骤

对解决方案示意图进行研究分析, 可通过如下的步骤来实现目标说话人语音的提取:

① 对输入的混合语音流进行分割分段处理;

② 通过改进势函数聚类提取混合语音方位信息;

③ 依据主导方位信息通过非线性时频掩蔽法分离提取主导语音(注: 为简化计算, 搜索阶段每段语音只提取了主导语音, 即能量最大方位的语音);

④ 将提取的主导语音采用 BP 说话人识别进行目标说话人识别, 如果识别结果是目标说话人转向步骤 5, 否则转向步骤 1 对下一段语音流进行搜索处理.

⑤ 由于混合语音经时频掩蔽后, 分离出的语音

信号依然保留了原有的方位信息, 因此可对目标说话人语音方位信息进行进一步修正, 获得更准确的方位信息;

⑥ 基于修正的方位信息提取目标说话人语音;

⑦ 目标说话人的跟踪监测(同样采用 BP 说话人识别), 经识别如果还是目标说话人, 则转向步骤 5 继续对目标说话人的方位进行修正; 如果不是则进行重新搜索. (注: 此步目标说话人的跟踪监测, 不是时时进行, 本文采用了一定时间间隔(5 秒)进行抽查.)

2.2.2 相关步骤说明

以上解决方案的实现步骤中有三处处理需要特别说明: 一是语音流分割分段处理; 二是声源方位信息修正处理; 三是目标说话人的跟踪监测.

语音流分割分段处理: 为了实现对目标语音的实时搜索, 就需要将语音流分割成较小的段, 但较小的段会带来目标说话人的识别率下降, 本文综合考虑了各种因素以及实现的简单, 采用了固定段长分割方法, 具体分 20 帧一段.

声源方位信息修正处理: 一般情况通过势函数聚类分析可得到语音初估方位信息, 再利用初估的语音方位信息进行时频掩蔽提取语音, 在一定程度上提取出的语音有比较高的信噪比, 分离效果较好. 但为了进一步提升语音分离效果, 还需进一步对初估的语音方位信息进行修正, 使之更加精确. 由于掩蔽分离后的语音信号依然保留了声源的方位信息, 在进行方位修正时可直接对其进行修正. 对此, 本文采用了相关辨识法来获取更加准确的 $\Delta H_0(\omega)$ 估计, 计算公式如下:

$$\Delta\hat{H}_0(\omega)'' = \frac{G_{21}(\omega)}{G_{11}(\omega)} \tag{12}$$

式(12)中, $G_{21}(\omega)$ 和 $G_{11}(\omega)$ 分别表示麦克风 1、2 掩蔽分离后的语音信号的互功率谱和麦克风 1 掩蔽后分离后的语音信号的功率谱.

目标说话人的跟踪监测处理: 在实际情况下目标说话人的方位有可能发生变化, 因此方案设计了对目标说话人的跟踪监测. 具体进行跟踪监测不是随时进行的, 本文采用了一定的时间间隔(5s)进行抽查监测.

在以上的实现步骤中只包括了对目标说话人的识别及语音提取, 除此之外, 还需提前对待识别和分离的混合语音流进行相关的预处理、端点检测, 同时对说话人语音的 MFCC 特征参数提取、BP 网络学习等相关步骤, 以此为方案的实现奠定基础.

3 实验仿真

为了论证本文提出的解决方案的可行性, 下面进行实验仿真, 并对实验结果进行分析.

3.1 实验环境及数据

本文研究内容是模拟现实生活中多人(3 人, 2 男 1 女)交流情况, 对任意方位的目标说话人进行辨识且分离提取其语音. 由于现实环境中随时随地存在噪声, 故本文利用专业软件“Room Impulse Response 2.2”构建有混响情况的实验环境, 具体如图 3 所示.

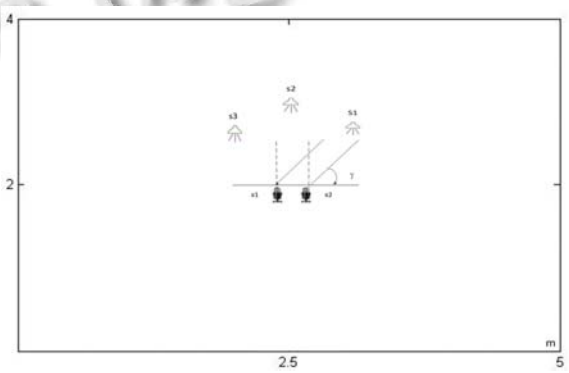

图 3 设置三个声源和双麦克风环境示意图

房间长、宽、高分别为(5、4、3)m, 房间混响时间 T=160ms, 两个全向麦克风置于房间的中心, 它们之间的距离 0.1m, 具体位置为: x_1(2.45, 2, 1.5)、x_2(2.55, 2, 1.5), 利用这两个麦克风接收来自三个不同方位的说话人语音, 即三个声源. 三个声源(3 个说话人, 2 男 1 女)成圆环状放置且在正半平面, 到两麦克风中心的距离为 1m, γ 为方位角(定义为声源相对于双麦克风阵(x_1,x_2)水平线的夹角), 三个声源的 γ 取值分别 45º, 90º, 135º, 利用声源的方位角可计算出声源对应的位置, 具体分别为: S_1 (3.21, 2.71, 1.75), S_2(2.5, 3, 1.70)和 S_3(1.79, 2.71, 1.65), 同时三个说话人与声源对应关系为: S_1 为男生 1、S_2 为女生 1、S_3 为男生 2. 声源到麦克风的冲击响应由专业软件生成, 观测信号由 3 个源信号和冲击响应卷积混合并叠加 -30dB 的高斯白噪声模拟环境噪声获得.

实验数据: 分别对三人采集时长为 20s, 文本内容不一样的 2 句语音作为识别训练样本, 待分离识别语音流时长为 150s. 语音格式为 wav 格式, 采样频率为 10kHz, 预加重系数为 0.9375, 分帧采用汉明窗(窗长

512, 帧移 256), 端点检测采用短时能量与短时过零率两者相结合的双门限检测方法.

3.2 性能评价参数

语音分离性能效果的评价, 本文采用基于信号的评价准则, 具体为信噪比增益与互相关系数两个评价指标. 信噪比增益定义如下:

$$SNRG=10\log_{10}\left\{\frac{E\left[|S(t)|^2\right]}{E\left[|y(t)-S(t)|^2\right]}\right\}-10\log_{10}\left\{\frac{E\left[|S(t)|^2\right]}{E\left[|x(t)-S(t)|^2\right]}\right\} \quad (13)$$

互相关系数定义为:

$$\xi_{sy}=\frac{|E[y(t)s(t)]|}{\sqrt{|E[y^2(t)]E[s^2(t)]|}} \quad (14)$$

式(13)、(14)中, $E[.]$为求均值运算, $S(t)$ 为纯净的目标语音信号, $y(t)$ 为提取的目标语音信号, $x(t)$ 为麦克风接收的混合语音信号. 式(13)中, $SNRG$ 的值越大说明分离效果越好, 反之则较差; 另式(14)中, ξ_{sy} 表示纯净的目标语音信号与提取的目标语音信号的互相关系数, 如果 $\xi_{sy}=1$, 说明提取的目标语音信号与纯净的目标语音信号完全相同, 分离效果非常好. 由于估计误差的存在, ξ_{sy} 只能接近 1; 如果 ξ_{sy} 的值趋于 0, 说明 $y(t)$ 与 $s(t)$ 不相关; 如果所有的 ξ_{sy} 值偏离 1 较远, 则表示分离未完成.

说话人识别性能的评价指标主要采用了识别率, 识别率是反映系统的识别正确率的高低.

3.3 实验及结果分析

实验一: 设目标说话人为女生 1 且位置相对固定, 方位角为 90º, 利用设计方案从混合语音流中识别她并提取她的语音流.

实验采用改进的势函数进行聚类分析, 其中参数 I_t, ξ, p 分别设置为 1、0.15、10, 两次聚类取值间隔大小分别为: $\lambda_{\alpha 1}=10,\lambda_{\tau 1}=2$; $\lambda_{\alpha 2}=0.8,\lambda_{\tau 2}=1;k_\alpha=2,k_\tau=1$; 非线性时频掩蔽语音分离, 参数 λ, p 分别设置为 8 和 6. 在构建的实验环境中根据实验条件进行仿真实验, 实验实得势函数聚类曲面效果图如 4 所示.

图 4 中 X 轴表示相对时间延迟, Y 轴表示相对幅度衰减(用角度表示), Z 轴表示信号的势能. 从图中可以看出有 3 个最大波峰, 即对应三个源信号, 由于环境噪声的存在, 其大的波峰后面还有小的干扰波峰. 波峰位置对应源信号势函数聚类估计的相对衰减、时延及势能, 根据对应的参数值, 得出估计的参数为:

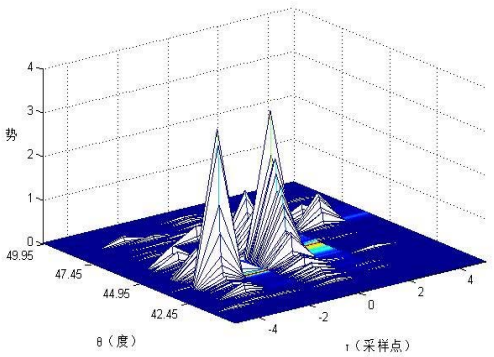

图 4 势函数聚类效果图

$$\Delta H_t(\omega)=\begin{bmatrix}1 & 1 & 1\\ 1.0576e^{-j2\pi k(-3)/K} & 1.0438 & 0.9587e^{-j2\pi k(3)/K}\end{bmatrix}$$

根据上面估计得到参数利用时频掩蔽可分离出目标说话人的语音信号. 实验一分离提取的目标说话人(女生 1)时域波形如图 5 所示.

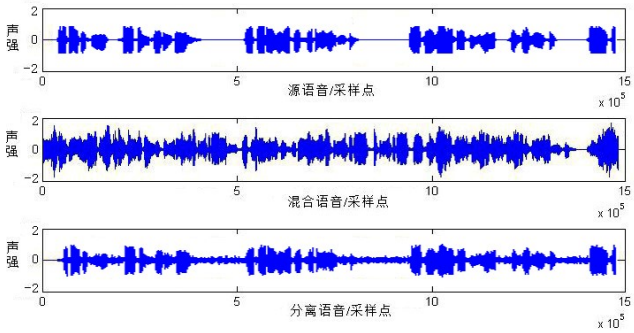

图 5 目标说话人语音提取波形

从图 5 可以看出, 采用本文设计的解决方案能成功的识别出了目标说话人并提取出对应的语音流, 由此说明了本文提出的解决方案具有可行性. 但由于目标说话人源语音信号在开始阶段没有出现, 方案首先需要对给定的语音流进行目标说话人语音搜索, 因此分离提取的目标说话人语音时域波形图的开始阶段和源语音有一定的差异.

实验数据结果如表 1.

表 1 目标说话人语音识别和分离实验结果

目标说话人	SNR_in	SNR_out	SNRG	ξ_{sy}	识别率	运行时间
女生 1	5.78dB	14.04dB	8.26dB	85%	65%	20.78S

(注: 实验数据结果为利用设计方案进行 100 次运行后统计平均.)

由表 1 可以直观的看出, 该方案提取的目标说话

人语音信号的信噪比增益、相似系数、识别率及运行时间都在我们所能接受的范围内, 且效果不错.

实验二: 设目标说话人同样为女生1, 位置发生变化, 方位角由 90 ° 变为 135º 并相对固定不动, 利用设计方案从混合语音流中识别她并提取她的语音流.

实验二分离提取的目标说话人(女生 1)时域波形如图 6 所示.

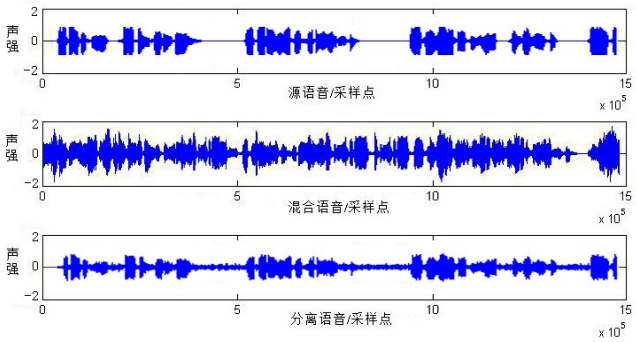

图 6　目标说话人语音提取波形

由图 6 可以看出, 同一目标说话人(女生 1)方位发生变化, 该方案同样能实现目标说话人的语音提取. 同时将图 5 与 6 进行对比分析可得出, 分离提取的目标说话人语音波形图形大致一样, 不同之处是分离提取语音的能量不一样, 实验二比实验一的能量要小一些, 主要是由于目标说话人的方位发生了变化(由 90º 变到 135º)导致的, 由此说明了说话人方位变化对分离效果有影响.

表 2　目标说话人语音识别和分离实验结果

目标说话人	SNR_in	SNR_out	SNRG	ξ_{sy}	识别率	运行时间
女生 1	5.78dB	13.45dB	7.67dB	81%	61%	20.73S

由表 1、2 对比可以看出, 在同样的 SNR_in 的情况下, 实验二的 SNR_out、SNRG、、识别率比实验一有所降低, 主要是因为目标说话人在新的位置(135º)的幅度衰减和时延要大一些, 由此进一步说明了目标说话人方位对分离效果有影响; 但运行时间没有发生多大变化, 几乎没有变化, 说明系统运行时间与目标说话人方位信息没有关系, 只与 SNR_in 有关.

实验三: 设目标说话人为男生 1 位置相对固定不动, 方位角为 45º, 利用设计方案从混合语音流中识别他并提他的语音流.

实验三分离提取的目标说话人(男生 1)时域波形如图 7 所示.

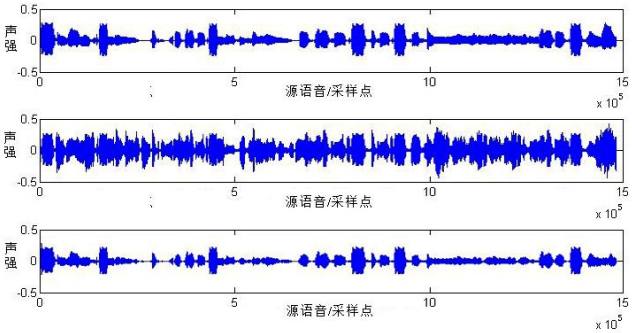

图 7　目标说话人语音提取波形

对图 7 分析可得, 分离提取目标说话人的语音信号波形图与源语音信号相差不大, 只是能量有所降低. 主要是由于目标说话人(男生 1)源语音信号一开始就有, 在利用设计方案进行目标说话人语音搜索时, 很快就搜索到; 同时由于方位的差异, 衰减和时延大小有所不同.

表 3　目标说话人语音识别和分离实验结果

目标说话人	SNR_in	SNR_out	SNRG	ξ_{sy}	识别率	运行时间
女生 1	-3.32dB	6.79dB	10.11dB	88%	58%	20.05S

根据表 3 的实验数据, 证明本文的设计方案对任意方位的目标说话人都可以识别并提取其语音, 且效果较好.

通过实验一、实验二、实验三的实验数据可得: 信噪比增益平均为 8.68dB, 相似系数为 85%, 识别率为 61%, 运行时间为 20.6S, 由此证明了本文设计的解决方案在现实多说话人的复杂环境中对任意方位的任意目标说话人进行识别并提取其语音具有可行性; 实验一、实验二进一步证明了同一目标说话人方位任意变化, 设计方案也能有效的识别目标说话人并提取其语音, 只是分离提取出的语音能量有所不同.

4　结语

本文主要基于目标声源方位信息与非线性时频掩蔽的语音欠定盲分离方法与 BP 说话人识别技术的研究基础上, 针对现实生活中多说话人交流的情景, 设计并提出了一种解决方案, 实现了对处于任意方位的目标说话人语音的提取, 以模拟实现“鸡尾酒效应”中人耳听觉系统智能辨识语音的能力. 通过仿真实验证明, 该方案具有科学性、可行性和有效性, 但在可靠性、稳定性和实用性方面还需进一步改进完善. 根据

研究体会, 具体应从下面几个方面进行着重考虑: 一是大规模声源的分离及识别(本文考虑了 3 个声源); 二是实际生活环境中, 声源信号实时移动或者是观测点相对声源信号移动情况下的语音分离(本文考虑的声源信号相对固定不动或者移动幅度不大的情况); 三是复杂环境下说话人语音分离和识别中噪声消除及语音增强; 四是利用音频和视频相结合的信息对说话人进行定位(本文只利用了语音的方位信息).

参考文献

1 Maddox J,郑佳.解开“鸡尾酒会效应”之谜.世界科学,1995,(1):23–23,40.

2 李从清,孙立新,龙东,任晓光.语音分离技术的研究现状与展望.声学技术,2008,27(5):779–787.

3 邱天爽,毕晓辉.稀疏分量分析在欠定盲源分离问题中的研究进展及应用.信号处理,2008,24(6):966–970.

4 李白燕,郭水旺,李应生.基于两步法稀疏分量分析的欠定盲源分离.电声技术,2010,34(9):64–67.

5 代勇,夏秀渝,陈林,叶于林.基于时频域的具有延迟的欠定盲分离.四川大学学报(工程科学版),2014,46(Z1):166–170.

6 余建潮,张瑞林.基于 MFCC 和 LPCC 的说话人识别.计算机工程与设计,2009,30(5):1189–1191.

7 单进.说话人识别技术研究.科技资讯,2010,(21):3–3.

8 Bofill P, Zibulevsky M. Underdetermined blind source separation using sparse representation. Signal Processing, 2001, 81(11): 2353–2362.

9 代勇,夏秀渝,陈林.一种改进的势函数聚类算法.电子技术应用,2013,39(11):107–110.

10 夏秀渝,何培宇.基于声源方位信息和非线性时频掩蔽的语音盲提取算法.声学学报,2013,38(2):224–230.

11 陈仁林,郭中华,朱兆伟.基于 BP 神经网络的说话人识别技术的实现.智能计算机与应用,2012,2(2):47–49.

12 陈仁林.基于神经网络的说话人识别算法研究[学位论文].银川:宁夏大学,2012.

13 兰胜坤.遗传算法优化 BP 神经网络的说话人识别系统.重庆理工大学学报(自然科学版),2013,27(10):91–95.