DOI:10.3969/j.issn.1673-4785.201109007

多个声源下基于人耳听觉特性的语音分离

罗元,童开国,张毅,邢武超,陈凯,陈红松,何春江,陈君
(重庆邮电大学 智能系统及机器人研究所,重庆 400065)

摘 要:受声学研究启发,结合人脑人耳听觉特性对语音的处理方式,建立了一个完整的模拟听觉中枢系统的语音分离模型.首先利用外周听觉模型对语音信号进行多频谱分析,然后建立重合神经元模型提取语音信号的特征,最后在脑下丘的神经细胞模型中完成对语音的分离.基于现有的语音识别方法,该模型能够很好地解决绝大多数的语音识别方法都只能在单声源和低噪声的环境下使用的问题.实验结果表明,该模型能够实现多声源环境下语音的分离并且具有较高的鲁棒性.随着研究的深入,基于人耳听觉特性的语音分离模型将有很广泛的应用前景.

关键词:多声源;人耳听觉特性;双耳时间差;双耳水平差;语音分离
中图分类号:TP311 **文献标志码**:A **文章编号**:1673-4785(2012)02-0121-08

Sound source separation of a multi-voice environment based on human ear listening properties

LUO Yuan, TONG Kaiguo, ZHANG Yi, XING Wuchao, CHEN Kai,
CHEN Hongsong, HE Chunjiang, CHEN Jun
(Research Center of Intelligent System and Robot, Chongqing University of Posts and Telecommunications, Chongqing 400065, China)

Abstract: Inspired by acoustics, an integrated voice separation model simulating the central auditory system was established to process a voice by imitating the listening properties of human ears. First, multi-spectral analysis of voice signals was carried out by a peripheral auditory model. Next, a coincidence neuron model was established to extract the features of voice signals. Last, the voices were separated in the cell model of the brain inferior colliculus. Compared to the majority of speech recognition models that can only be used in a single sound source and low-noise environment, this model is a good choice. Experimental results show that the model can separate voices in a multi-sound source environment, thus having a high robustness. With further research, speech separation models based on human ear listening properties will have a wide range of applications.

Keywords: multi-voice source environment; human ear listening properties; interaural time difference; interaural level difference; sound source separation

在多声源下,利用听觉中枢系统对语音分离已有20多年的研究历史,总体来说有3个阶段的模型.第1个模型是Bhadkamkar提出的,方法是构建COMS电路来处理双耳时间差(interaural time difference, ITD),这种方法简单、容易实现,适用于工程,但是精度不够高[1].第2个模型是Willert等提出的,方法是构建概率模型来估计声源的方位,结合了内侧上橄榄(medial superior olive, MSO)、外侧上橄榄(lateral superior olive, LSO)和脑下丘,并且利用贝叶斯理论来计算他们之间的联系,但是没利用生物电信号神经网络来模拟现实的神经元对语音的分离[2].第3个模型是Voutsas等提出的,构建尖峰神经网络多滞后线模型,利用ITD,对低频语音信号分离有良好的效果,但是由于只考虑ITD,对高于1.5 kHz的语音信号没有效果[3].

在过去的25年里,对于听觉中枢系统的结构和功能的研究已经有了长足的进步[4],脑下丘在听觉信息的获取过程中起到了非常关键的作用[5].

脑下丘是提取声音特征的一个枢纽和处理中心[6].在这里,声音中双耳时间差和水平差都被提取出来.听觉学研究表明,双耳的辨别功能比单耳好[7].根据从声源到两耳距离的不同及传声途径中屏蔽条件的不同,从某一方位发出的声音到达双耳

收稿日期:2011-09-28.
基金项目:科技部国际合作资助项目(2010DF12160);重庆市攻关计划资助项目(CSTC:2010AA2055).
通信作者:童开国. E-mail:359018647@qq.com.

时,便出现双耳时间差和双耳水平差,在听觉中枢系统对输入语音信息进行分离时,双耳时间差和水平差便是声源定位的重要依据[8].

脑下丘会控制内耳神经的听觉纤毛响应阈值,低频段(小于 1.5 kHz)的语音信号(在这个频段范围内 ITD 对语音离位更有效率)会经过 MSO 的中区传递给脑下丘;高频段(大于 1.5 kHz)的语音信号(在这个频段范围内 ILD 对语音分离更有效率)则可以同时经过 MSO 和 LSO 的中区传递给脑下丘,最后不同区域的信号分别输入给脑下丘[9]. 脑下丘的神经组织还有一个重要的特点:在物理上使用多层解剖结构对声音信号依照频率进行分解,每一层的神经细胞只对特定的频率分量进行响应,这种解剖特征被称为频率解剖特征,这种特征使得多频段语音输入在脑下丘中进行了空间隔离[10]. 这样,来自同一声源或者具有同样频率特征的声音就很容易被重合和提取出来,于是在嘈杂的多声源环境中,语音信号就分别被分离出来,重新生成信号流[11].

综上所述,听觉中枢系统对多声源噪声输入能够有效地进行分离,建立一个完整的模拟听觉中枢系统的语音分离模型,就可能解决目前动态复杂环境下的语音识别问题.

1 多声源环境下基于听觉中枢系统的语音分离模型

基于听觉中枢系统的语音分离模型如图 1.

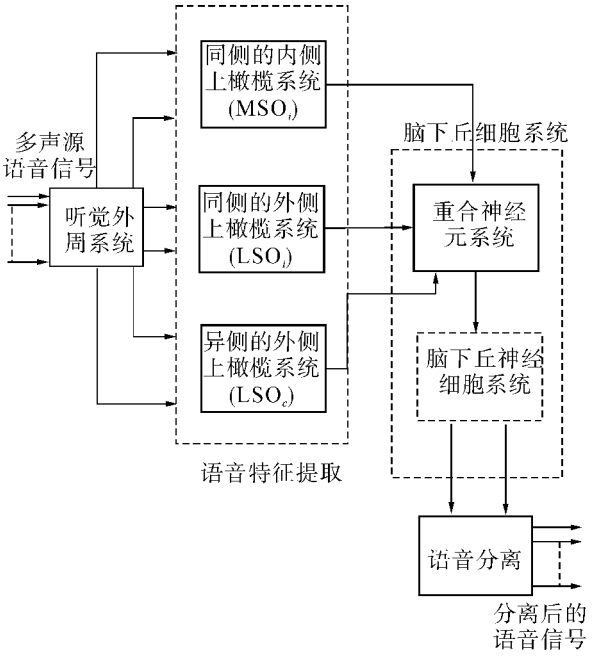


图 1 基于听觉中枢系统的语音分离模型

Fig. 1 The speech separation model based on central auditory system

图 1 是本文提出的多声源环境下基于听觉中枢系统的语音分离原理结构图,是一个完整的模拟听觉中枢系统的计算模型. 多路语音信号先经过听觉外周模型,根据频率的不同而被划分为不同的频率通道,然后经过上橄榄复合体(SOC,包括 MSO 和 LSO)进行语音信息提取,最后利用脑下丘细胞模型将多声源分离成单个的语音信号.

1.1 听觉外周模型

声学研究表明,位于耳蜗内部的基底膜具有频率分解的作用,不同频率的信号将激发基底膜的不同位置具有不同振动. 基于基底膜的特性,音频外围处理时,本文选择用 24 个二阶离散的 Gammatone (GT)滤波器组,取代常用的三角滤波器来进行多频率分析. Gammatone 函数的时域如式(1)所示:

$$g(t) = \frac{t^{n-1}\cos(2\pi f_0 + \theta)}{E^{2\pi bt}}u(t). \qquad (1)$$

式中:n 表示滤波器的阶数,选取 $n = 4$;参数 θ 为 Gammatone 滤波器的初始相位;$u(t)$ 为阶跃函数;参数 $b = b_1\mathrm{ERB}(f_0)$,$b_1 = 1.019$,$\mathrm{ERB}(f_0)$ 是 Gammatone 滤波器的等效矩阵带宽,并且它和 Gammatone 滤波器中心频率 f_0 有如下关系:

$$\mathrm{ERB}(f_0) = -2.47 + 0.108f_0.$$

图 2 是一组利用听觉外周模型的 Gammatone 滤波器组的频率响应图,是 24 个 Gammatone 滤波器组成的滤波器组,它的频率范围是 80 ~ 4 000 kHz. 对于输入的语音信号,经过听觉外周模型的多频率分析之后,根据频率的不同,分别在听觉中枢系统中的 24 个不同的频率通道内传递,便于语音信号在系统模型中的分离.

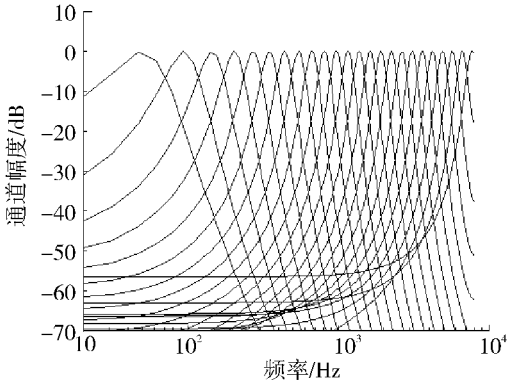


图 2 伽马通滤波器组的频率响应

Fig. 2 The frequency response of filter group consisted of Gammatone filters

1.2 重合神经元模型

重合神经元模型模拟突触和细胞体的响应,完成对语音信息的提取与融合. 本文分别选取了 Meddis 的通用突触函数模型和已经成熟应用的 Leaky

integrate-and-fire(LIF)模型来模拟突触和细胞体对语音信息的提取,然后又根据听觉神经中枢对 ITD 和 ILD 的信息整合的特点,提出了本文核心重合神经元模型,完成对语音信息的融合.

1.2.1　通用突触模型

语音信号在基底膜上引起的振动会造成递质通过可渗透膜向突触间隙释放,引起了听神经的发放. 渗透膜的渗透率 $h(t)$ 是变化的,决定于输入信号的振幅,每个 GT 滤波器输出要经过半波整流.

$$h(t) = \begin{cases} \dfrac{A + \mathrm{stim}(t)}{A + B + \mathrm{stim}(t)} g, & A + \mathrm{stim}(t) \geqslant 0; \\ 0, & A + \mathrm{stim}(t) < 0. \end{cases}$$

式中:$\mathrm{stim}(t)$ 是输入语音信号瞬时的幅度,A 为信号 $x(t)$ 的渗透阈值,g 是与渗透率相关的量,B 与最大渗透率有关.

图 3 是突触模型的原理图. 突触中内毛细细胞含有可以自由释放的神经递质量,用 $q(t)$ 表示,且有 $y[1-q(t)]$ 的补偿率. 突触裂隙内包含的神经递质量以 $c(t)$ 表示,它向内毛细胞返回的量为 $rc(t)$,并且有 $lc(t)$ 的神经递质量不断的丢掉,可用下列方程来描述突触子系统的操作过程:

$$\frac{\mathrm{d}q}{\mathrm{d}t} = y[1 - q(t)] + rc(t) - h(t)q(t), \quad (2)$$

$$\frac{\mathrm{d}c}{\mathrm{d}t} = h(t)q(t) - lc(t) - rc(t), \quad (3)$$

$$p(t) = hc(t)\mathrm{d}t. \quad (4)$$

式(2)~(4)组成了通用突触模型,其中,y、r、l、h 是相关的一些常数,$\mathrm{d}t$ 则是采样间隔,取值如表 1 所示.

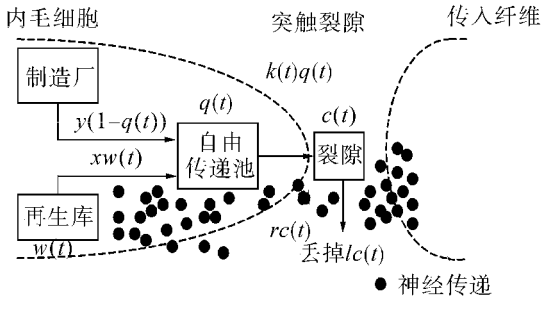


图 3　突触子模型的原理

Fig. 3　Synaptic model diagram

表 1　参数取值

Table 1　Parameters

参数	描述	数值/s
A	参透常量	2.00
B	参透常量	300.00
L	丢失速率	2 500.00
R	恢复速率	6 580.00
X	再加工速率	66.31
Y	再加工速率	8.00
G	恢复速率	2 000.00

1.2.2　通用细胞体模型

递质分子通过突触间隙递质扩散到突触后神经元而形成电流,电流向神经元的细胞体移动,形成一个逐渐增加的突触后电流 $I(t)$. 本文选择 LIF 模型来模拟通用细胞体的功能,如图 4 所示,包括 1 个电阻 R 以及 1 个与之并联的被外来电流 $I(t)$ 驱动的电容 C,其中,

$$u(t) = u_r \exp\left[-\frac{t - t^{(f)}}{\tau_m}\right] + \frac{1}{C}\int_0^{t-t^{(f)}} \exp\left[-\frac{s}{\tau_m}\right] \cdot I(t-s)\,\mathrm{d}s.$$

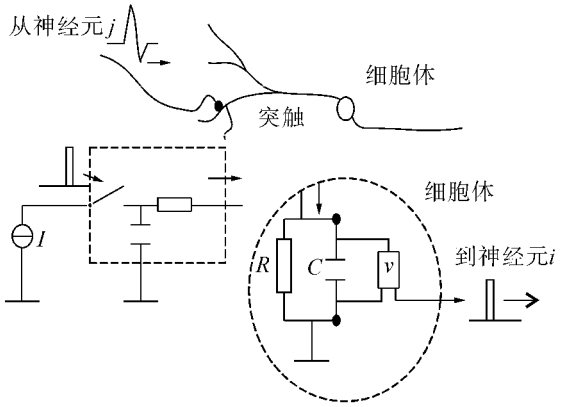


图 4　Leaky integrate-and-fire 模型的结构

Fig. 4　Schematic of the leaky integrate-and-fire model constant

初始膜电位是 u_r,τ_m 是一个常量,分别代表电阻 R 和电容 C. C 是被 $I(t)$ 充电的电容,φ 为行动电位. 如果在 t 时刻,当 $u(t) = \varphi$ 时,细胞体将会释放一个脉冲,然后 $u(t)$ 被重设为初始电压 0.

1.2.3　重合神经元模型

在已有的通用突触模型和通用细胞模型的基础上,本文根据生物学原理提出重合神经元模型,分别用于对 ITD 和 ILD 信息进行融合.

ITD 通路,异侧耳朵的脉冲序列的发射要经过变化的延迟线 Δt_i,表示延迟脉冲序列为 $S_{\mathrm{CP}}(\Delta t_i, f_j)$,这里 C 代表异侧,f_j 代表频率通道 j. 类似地,$S_{\mathrm{IP}}(\Delta T, f_j)$ 代表同侧耳朵的固定延迟脉冲序列带有一个固定的延迟时间 ΔT. 为了计算 ITD $S_{\mathrm{CP}}(\Delta t_i, f_j)$ 和 $S_{\mathrm{IP}}(\Delta T, f_j)$,被输入到 ITD 的重合模型. ITD 重合模型计算的输出是一个新的脉冲序列,即为 $S_{\mathrm{ITD}}((\Delta T - \Delta t_i), f_j)$. 脉冲 $S_{\mathrm{ITD}}((\Delta T - \Delta t_i), f_j)$ 代表声音到达同侧耳朵比到达异侧耳朵,ITD $= \Delta T - \Delta t_i$. 图 5 就是 ITD 的重合模型,其中,ES 代表兴奋性突触.

ILD 通路没有使用 LIF 模型,检测到两侧声音等级用来计算等级差,并且相应的 ILD 细胞将释放

一个脉冲. 等级差异的计算公式是: $\Delta p^j = \log(p_I^j / p_C^j)$, 这里 p_I^j 和 p_C^j 分别代表频道 j 的同侧和异侧声音等级. 对于脉冲 $S_{\mathrm{ILD}}(\Delta p_j, f_j)$, 负的 ILD 值意味着声音等级将会是右耳的比左耳的低, 正的 ILD 值正好相反. 图 6 为 ILD 的重合模型, 其中 ipsi 和 contra 是代表异侧的 Gammatone 频率通道.

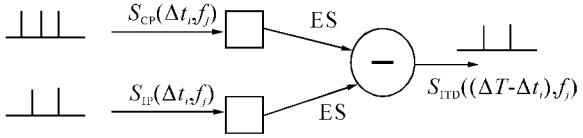


图 5 ITD 的重合模型

Fig. 5 ITD coincidence model

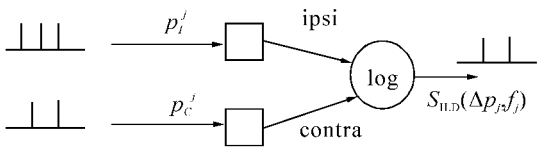


图 6 ILD 的重合模型

Fig. 6 ILD coincidence model

由以上分析, 在完成重合神经元之后, 为了提取并融合 ITD 和 ILD 传递的语音信息, 建立了 2 个加权阵列: ITD_w 和 ILD_w, 在所有的频率范围内, 利用乘以一个二维的 ITD/ILD 的矩阵加权阵列计算出一个加权的 ILD 和 ITD 映射.

$$\mathrm{ITD}_w^j = \frac{\sum_j (\max(f_j/1\,200, 1))}{\max(f_j/1\,200, 1)},$$

$$\mathrm{ILD}_w^j = \frac{\max(\log(f_j/1\,000, 0))}{\sum_j (\max(\log(f_j/1\,000, 0)))}.$$

式中: j 是频道指数. 加权的 ITD 和 ILD 映射信息最终被融合到一起, 也就是 MSO 和 LSO 的输出信息, 最后被输入到脑下丘的神经细胞内进行语音信息的提取和分离.

1.3 脑下丘细胞模型

脑下丘中一共有 Rebound Regular、Rebound Onse、/Sustained Regular、Onset 等几种细胞. 本文根据脑下丘的 Onset 神经细胞模型对多声源的语音信号进行分离的特征, 构造了 Oneset Cell 模型. 图 7 是脑下丘的 Onset 神经细胞模型的结构原理图.

对于 Onset Cell 模型, 每一个模型都有激活和非激活 2 个状态. 当细胞为激活状态时, 模型被实施为 LIF 模型的神经元, 直到释放了一个脉冲或者接受一个抑制性的输入, 然后细胞模型变为非激活状态. 当为非激活状态时, 也就是细胞模型为空置状态, 直到细胞模型在一段持续时间 t_s 内没有受到抑制并且输入为 0(无脉冲)后, 细胞模型会变为激活状态.

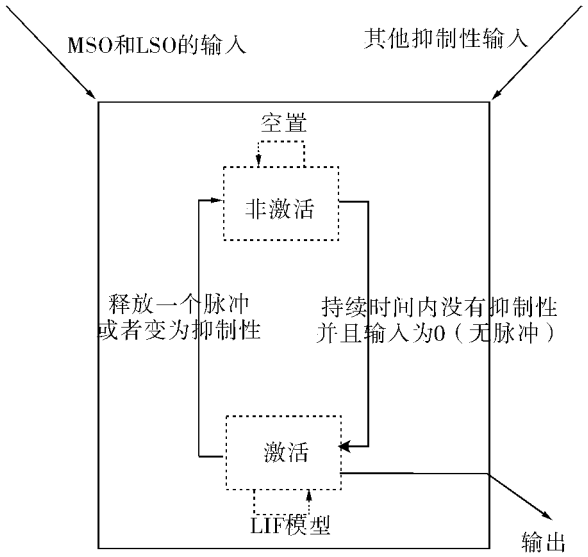


图 7 脑下丘的起始神经细胞模型

Fig. 7 The IC's onset cell

再利用 Onset Cell 模型对多声源语音信号进行分离时, 要用到信号能量比, 首先计算出神经细胞模型中语音信号的第 i 频率通道、第 j 时间帧能量 $\sum_i S_{i,j(t)}^2$ 和噪声信号能量 $\sum_i n_{i,j(t)}^2$, 然后计算出信号能量比:

$$E_{i,j} = \frac{\sum_i S_{i,j(t)}^2}{\sum_i S_{i,j(t)}^2 + \sum_i n_{i,j(t)}^2}.$$

如果 $E_{i,j} > 0.5$, 表明语音能量大于噪声能量, 应该保留这个语音占主导地位的信号片段; 反之, 如果 $E_{i,j} < 0.5$, 表明噪声能量占主导地位则应当舍去. 然后再利用 Onset 细胞模型获取 ITD 和 ILD 的值, 来构建掩蔽矩阵, 实现语音信号的分离. 本文采用二值掩蔽, 对于第 i 通道、第 j 时间帧的掩蔽系数可以定义为

$$\lambda(i,j) = \begin{cases} 1, f_i \leqslant f_c, \text{且}[\tau_{\max}(i,j)] > T^{(\tau)}(i,j)]; \\ 1, f_i > f_c, \text{且}[L(i,j)] > T^{(l)}(i,j)]; \\ 0, \text{其他}. \end{cases}$$

式中: $f_c = 1.5$ kHz, $T^{(\tau)}(i,j)$ 和 $T^{(l)}(i,j)$ 分别是 ITD 和 ILD 的阈值, $\tau_{\max}(i,j)$ 是第 i 频率通道、第 j 时间帧最大的时间延迟, $L(i,j)$ 是第 i 频率通道、第 j 时间帧的 ILD 值,

$$L(i,j) = 20\lg \frac{\sum_{i,j} p_1(i,j,t)^2}{\sum_{i,j} p_r(i,j,t)^2}.$$

式中: $p_1(i,j,t)$ 和 $p_r(i,j,t)$ 分别为第 i 频率通道、第 j

时间帧的左、右耳的信号发放率.

对多声源的语音信号在各频率通道和各时间帧上求掩蔽系数,然后再获得掩蔽矩阵. 矩阵中所有相同的元素 1 和所有相同的元素 0 为同一归属.

所有相同的元素 1 的矩阵中,信号的自相关函数的傅里叶变换等于该信号傅里叶变换幅度的平方. 如果用 $R_{xx}(\tau)$ 表示 $x(t)$ 的自相关,则 $x(t)$ 的功率谱 $|X(w)|^2$ 为

$$|X(w)|^2 = \int_{-\infty}^{\infty} R_{xx}(\tau)\exp(-jw\tau)\mathrm{d}\tau.$$

由此可得到听觉模型中神经发放率的短时幅度谱,接下来进行一种迭代算法,该算法在每次迭代中,重构信号的相位信息,以减少重建信号的短时傅里叶变换幅度与原已知信号的短时傅里叶变换幅度之间的平方误差,从而得到信号的估计值,然后将估计信号的傅里叶变换幅度值与原已知的傅里叶变换幅度值的平方误差最小化. 第 i 次迭代重构的信号 $x^{(i)}(n)$ 由式(5)表示:

$$x^{(i)}(n) = \frac{\sum_{m=-\infty}^{\infty} w(mS-n)\frac{1}{2\pi}\int_{-\pi}^{\pi}\hat{X}^{(i-1)}(m,n)\mathrm{e}(jw\tau)\mathrm{d}w}{\sum_{m=-\infty}^{\infty} w^2(mS-n)}. \quad (5)$$

式中:$w(mS-n)$ 为分析窗,S 为窗移. 可以根据 $x^{(i)}(n)$ 求出第 i 次迭代重构信号的短时傅里叶变化 $X^{(i)}(m,n)$,并由式(6)可以求出它与原来给定的短时幅度 $X_d(m,n)$ 之间的误差.

$$\mathrm{Error} = \sum_{m=-\infty}^{\infty}\sum_{n=0}^{N-1} \| X^{(i)}(m,n)| - |X_d(m,n)\|^2. \quad (6)$$

如果误差小于给定的值,迭代结束;否则计算出 $\hat{X}^{(i)}(m,n)$,按照式(5)进行下一次迭代.

$$\hat{X}^{(i)}(m,n) = |X_d(m,n)|\frac{X^{(i)}(m,n)}{|X^{(i)}(m,n)|}.$$

经过以上的运算,可以求出听觉模型中每个通道的神经发放率 $p(t)$. 下一步要从听神经发放率 $p(t)$ 恢复出半波整流后的信号 $h(t)$:

$$c(t) = \frac{p(t)}{h\mathrm{d}t}.$$

求得 $c(t)$ 后,经过推导可以依次求出 $q(t)$ 和 $h(t)$:

$$q(t) = y[1-q(t-1)]\mathrm{d}t - lc(t-1)\mathrm{d}t - c(t) - c(t-1) + q(t-1),$$

$$h(t) = \frac{\left[\frac{c(t)-c(t-1)}{\mathrm{d}t} + lc(t) + r(t)\right]}{q(t)}.$$

$h(t)$ 即为求得的半波整流后的信号表示. $h(t)$ 再次经过迭代就可以得出原始语音信号.

2　实验结果及分析

2.1　实验配置

本文选择具有代表性的国家“863”多语言基础资源库,通过与当前语音分离最为权威的尖峰神经网络的多滞后线模型比,来验证本文的模型.

国家“863”多语言基础资源库口语语音库中,包括1 500人的 El 语语音库,主要有电话语音、会议语音各 750 人和广播语音,每人发音长度至少为30 min,随意口语. 本文选择的实验测试集是“863”多语言基础资源库的广播语音库(包括访谈类和新闻类),总共有300 h的较大规模资料库,从中随机挑选了 20 个人(10 男 10 女)的 50 个汉语单词和句子.

选择上述构建好的 Oneset Cell 模型在 Intel Pentium 2.5 GHz、内存1 GB的微机上,利用 Matlab 对上述模型用以上的方案进行试验. 把这些测试数据总结为 3 类(分别用 A、B、C 表示),每个测试类分别包括 2 种语音信号和一个噪声(本文选择交通噪声)信号,采样率为44.1 kHz,选择 16 位的采样精度. A 类:声源 1 为男生汉语单词,声源 2 为女生汉语单词;B 类:声源 1 为男生汉语单词,声源 2 为女生汉语短句;C 类:声源 1 为女声汉语短句,声源 2 为男声短句单词.

2.2　实验结果

图 6 就是选取本文所用模型中 C 类的一个语音分离仿真结果.

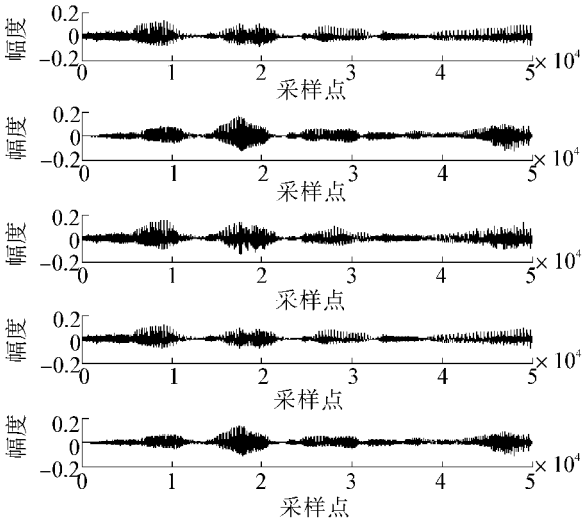
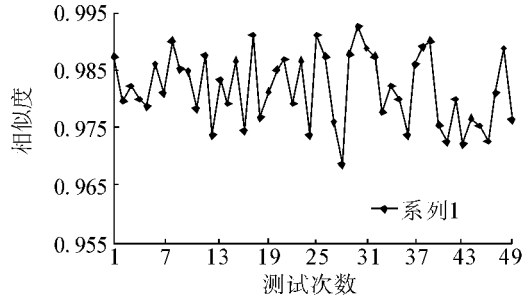


图 8　双语音信号源语音分离结果

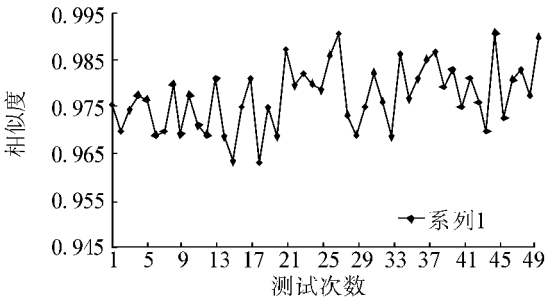
Fig. 8　Dual voice signal source separation results

第1幅图是原始的女生“中国向前进”,第2幅图是原始的男生“人民齐发展”,第3幅图是混叠后的信号,第4幅图是分离后的声源信号男生“人民齐发展”,第5幅图是分离后的声源信号女生“中国向前进”.

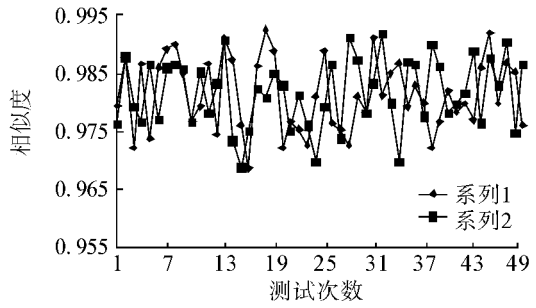
对于A、B、C 3类测试,做了大量实验之后,从每类测试中分别随机抽取了50组,结果对分离后的语音信号和原始的语音信号波形利用Matlab进行相似性对比.图9给出了相似度比较结果.



(a)A类



(b)B类



(c)C类

图9 3类语音信号分离前后相似度

Fig.9 Three similarity curve table of the third group

图9分别对应于A、B、C 3类测试的相似度对比结果,横坐标代表试验次数,纵坐标代表分离后语音信号和原始语音信号的相似度.由曲线可得,分离后的语音信号与原始的平均相似度可以达到0.97以上,由此可得,本文提出的完整的利用听觉中枢系统的模型对于多声源环境下的语音分离具有很高的鲁棒性.

接下来,本文对比Voutsas等构建的构建结合实际尖峰神经网络的多滞后线模型[3],该模型也利用生物学听觉中枢的相关原理,但是在提取多声源语音信号的特征时只利用了ITD信息,也就是说,该模型没有利用重合神经元融合ILD的信息.本文随机从国家“863”多语言基础资源库的口语语音库中挑选了25个小于1.5 kHz的词语和25个大于1.5 kHz的词语,利用该模型进行语音分离实验,并且将分离后的语音信号和原始的语音信号进行相似度比较,结果如图10所示.

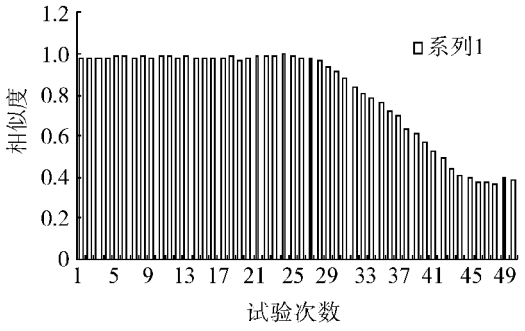


图10 Voutsas和Adamy的模型的语音信号分离前后相似度曲线

Fig.10 Similarity curve table of Voutsas and Adamy's model

由图10可得,对小于1.5 kHz频率的低频语音信号,采用Voutsas等构建的构建结合实际尖峰神经网络的多滞后线模型,其结果相似度可以达到0.975以上,但是对于大于1.5 kHz的语音信号却越来越弱.这一点正好符合生物学原理,ITD对低于1.5 kHz的语音信号的特征提取起作用,而对高于1.5 kHz的语音信号则会失去效果;ILD则正好相反.

由以上分析可得,相对于Voutsas和Adamy构建的构建结合实际尖峰神经网络的多滞后线模型,本文所提出的模型更好地模拟了人类听觉中枢对语音信号的特征提取和分离,能够在更广、更全的频率范围内有效地对多声源环境下的语音信号进行分离,并且具有较高的鲁棒性.对于第1类和第2类的测试实验,采用本文的方法还可以提高语音信号的信噪比.利用重合神经元融合的ITD和ILD的信息,选取了5组的方位角数据,按照信噪比计算公式:

$$\mathrm{SNR} = 10\lg\left(\sum_i \hat{s}(t)^2\right)\left\{\sum_i [s(t)-\hat{s}(t)]^2\right.,$$

计算的对比结果如表2所示.

表 2　2 组语音分离前后信噪比对比

Table 2　The contrast of signal to noise ratio

角度/(°)	第 1 组		第 2 组	
	分离前	分离后	分离前	分离后
0,25	17.2	50.2	12.4	49.2
0,45	16.7	49.5	12.1	48.2
45,75	16.8	50.1	12.9	48.6
100,130	15.3	49.4	12.8	46.8
140,145	11.8	21.1	11.5	20.7

由表 2 可得,当 2 个声源的入射具有一定空间方位差别时,分离后的信噪比有了大幅度的提高,当 2 个声源的入射的空间方位差别较小时,分离后的语音信号的信噪比与分离前的差别不大. 例如表 2 中,当方位角(θ_1,θ_2)选取为(135°,140°)时,重合神经元在计算 ITD 和 ILD 的信息时容易造成偏差,也就造成了掩蔽系数的计算错误. 这种现象也可以利用人的听觉现象来解释,当 2 个声源来自 2 个很相近方位角时,人的听觉系统难以分辨出其中的一个声音.

3　结论与展望

提出了一种在多声源环境中语音分离方法,建立了一个完整的人脑听觉中枢系统模型. 与现有的语音识别方法相比,本文模型很好地解决了绝大多数的语音识别方法都只能在单声源和低噪声的环境下使用的问题.

随着研究的深入,基于听觉中枢系统的语音分离模型将具有广泛的应用前景:1)智能机器人,可以提高语音系统识别率;2)助听设备,用于有听力障碍的残疾人;3)多媒体检索,辅助目前的文字检索;4)语音增强,去除音频文件中掺杂的一些干扰噪声.

参考文献:

[1] OZEROV A, VINCENT E, BIMBOT F. A general modular framework for audio source separation[C]//9th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA10). Saint-Malo, France, 2010: 33-40.

[2] VINCENT E, BERTIN N, BADEAU R. Harmonic and inharmonic on negative matrix factorization for polyphonic pitch transcription[C]//Proc of IEEE International Conference on Acoustics, Speech, and Signal Processing. Rennes Cedex, France, 2008: 109-112.

[3] FITZGERALD D, GAINZA M. Single channel vocal separation using median filtering and factorization techniques[J]. ISAST Transactions on Electronic and Signal Processing, 2010, 4(1): 62-73.

[4] 赵鹤鸣,葛良,陈雪勤,等. 基于声音定位和听觉掩蔽效应的语音分离研究[J]. 半导体学报, 2005, 33(1): 158-160.
ZHAO Heming, GE Liang, CHEN Xueqin, et al. Research based on sound localization and auditory masking effect of voice separation[J]. Journal of Semiconductors, 2005, 33(1): 158-160.

[5] LIU Jindong, ERWIN H, WERMTER S. Mobile robot broadband sound localisation using a biologically inspired spiking neural network[C]//Proceedings of IEEE/RSJ Int Conf on Intelligent Robots and Systems in Nice. [S. l.], 2008: 2191-2196.

[6] DURRIEU J L, RICHARD G, DAVID B. An iterative approach to monaural musical mixture desoloing[C]//Proc of IEEE International Conference on Acoustics, Speech, and Signal Processing. Paris, France, 2009: 105-108.

[7] KONIARIS C, CHATTERJEE S, KLEIJN W B. Towards effective singing voice extraction from stereophonic recordings[C]//2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP). Hatfield, UK, 2010: 233-236.

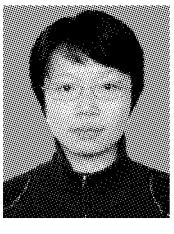
[8] BROWN G J, FERRY R T, MEDDIS R. A computer model of auditory efferent suppression: implications for the recognition of speech in noise[J]. Acoustical Society of America, 2010, 127(2): 943-954.

[9] DUONG N, VINCENT E, GRIBONVAL R. Spatial covariance models for under-determined reverberant audio source separation[C]//Applications of Signal Processing to Audio and Acoustics 2009 (WASPAA'09). Rennes, France, 2009: 129-132.

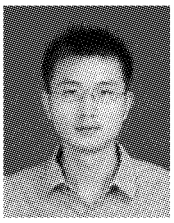
[10] DONG Yi, MIHALAS S, NIEBUR E. Improved integral equation solution for the first passage time of leaky integrate-and-fire neurons[J]. Neural Computation, 2011, 23(2): 421-434.

[11] VOUTSAS K, ADAMY J. A biologically inspired spiking neural network for sound source lateralization[J]. IEEE Trans Neural Networks, 2007, 18(6): 1785-1799.

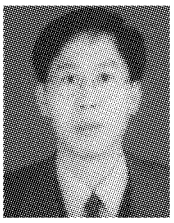
作者简介:



罗元,女,1972 年生,教授,博士. 近年来参与和负责了包括科技部国际合作项目、教育部留学回国人员项目、重庆市科研项目等多项国家级、省部级项目. 主要研究方向为机器视觉、人机交互、基于图像视频处理的测试. 近年来发表学术论文 60 余篇,其中 20 余篇被 SCI、EI 检索,获得国家发明专利 3 项.



童开国,男,1985年生,硕士研究生,主要研究方向为语音识别与智能机器人,发表学术论文4篇.



张毅,男,1966年生,教授,博士生导师,博士后,近年来承担了科技部国际合作项目、人事部留学人员科技活动项目择优资助重点项目以及重庆市科技攻关项目"轮椅式机器人导航与控制系统研发"课题;国际期刊 International Journal of Modelling, Identification and Control、International Journal of Automation and Computing 和 International Journal of Advanced Mechatronic Systems 关于智能系统及机器人专刊的编委.

2013年计算科学与工程会议
SIAM Conference on Computational Science & Engineering (CSE13)

Computational Science and Engineering (CS&E) is now widely accepted, along with theory and experiment, as the critical third pillar of scientific discovery. It is indispensable for leading edge investigation and engineering design in a vast number of industrial sectors, including for example, aerospace, automotive, biological chemical, and semiconductor technologies that all rely increasingly on advanced modeling and simulation. CS&E has also become essential at government agencies for informing policy and decisions relating to human health, resources, transportation, and defense. Finally, in many new areas such as medicine, the life sciences, management and marketing (e.g. data- and stream mining), and finance, techniques and algorithms from CS&E are of growing importance.

CS&E is by nature interdisciplinary. Its goals concern understanding and analyzing complex systems, predicting their behavior, and eventually optimizing processes and designs. CS&E thus grows out of physical applications, while depending on computer architecture, and having at its core powerful algorithms. At the frontiers of CS&E there remain many open problems and challenges, including for example, the validation and verification of computational models especially in the presence of uncertainties and the analysis and assimilation of very large data sets, including techniques for visualization and animation.

The SIAM CS&E conference seeks to enable in-depth technical discussions on a wide variety of major computational efforts on large problems in science and engineering, foster the interdisciplinary culture required to meet these large-scale challenges, and promote the training of the next generation of computational scientists.

Themes

Multiphysics and Multiscale Computations
Identification, Design, and Control
Surrogate and Reduced-order Modeling
Verification, Validation, Uncertainty Quantification
Discrete Simulations
Scientific Data Mining
Scalable Algorithms for Big Data
Simulations on Emerging Architectures
Exascale Challenges
Scientific Software and High-Performance Computing
Applications in Science, Engineering, and Industry
Computational Mathematics of Planet Earth
CSE Education

Website: http://www.siam.org/meetings/cse13/ .