基于声音定位和听觉掩蔽效应的语音分离研究

赵鹤鸣, 葛 良, 陈雪勤, 俞一彪
(苏州大学电子信息学院, 江苏苏州 215021)

摘 要: 人耳具有在嘈杂环境中将感兴趣的语言信息提取出来的能力, 而双耳听觉特性有助于这种能力的加强. 据此本文提出了一种基于声音定位和听觉掩蔽效应的混叠语音分离方法. 根据声音到达双耳的时间差和强度差在时频域内确定相应的掩蔽系数, 该系数是二值的, 以直接去除干扰信号, 保留有用信号并达到语音分离的目的. 实验表明, 本文提出的方法是有效的. 该方法不仅适用于混叠语音为浊音情形, 对清音的情况同样适用, 因而比基于基音提取的语音分离方法的适用范围更广.

关键词: 双耳时间差; 双耳强度差; 声音定位; 语音分离; 掩蔽效应
中图分类号: TN912.13 文献标识码: A 文章编号: 0372-2112 (2005) 01-0158-03

Speech Separation Based on Sound Localization and Auditory Masking Effect

ZHAO He-ming, GE Liang, CHEN Xue-qin, YU Yi-biao
(School of Electronics and Information Engineering, Soochow University, Suzhou, Jiangsu 215021, China)

Abstract: Human has the ability to attend to a single interested speech in a noised condition and this ability can be improved in the presence of binaural cues. In this paper a speech separation method is presented based on sound localization and auditory masking effect. By two important parameters-the interaural time differences (ITD) and interaural intensity differences (IID)-we estimate the binary masking coefficients in corresponding time-frequency regions. The coefficients are helpful of speech separation by holding interested signal and reducing noise signal. Experiments indicate that the approach described here is efficient not only for voiced speech but also for unvoiced speech and it has more extensive applications than pitch-based speech separation algorithms.

Key words: interaural time differences; interaural intensity differences; sound localization; speech separation; masking effect

1 引言

噪声环境下的语音信号处理一直是国内外学者十分重视且具有应用意义的研究课题, 这方面的研究尤以语言与噪声(含干扰语音)的自动分离最为困难, 但这正是实现语音处理技术走向实用的有效途径之一[1]. 自计算听觉场景分析(CASA)提出以来[2,3], 语音分离的研究更为人们所重视. 混叠语音信号可以不同的基音频率、谐波、相位、声音强度及空间方位等构成. CASA 就是模拟人的听觉感知特性, 根据上述不同信息, 从混叠信号中抽取相关特征形成感兴趣的单一语音流. 目前这方面的研究主要集中在利用基音和谐波信息实现混叠语音自动分离[4], 显然这种方法仅局限于含浊音的情形.

听觉研究表明[5], 双耳听觉的辨别功能比单耳好, 特别是在噪杂环境中选择性地注意感兴趣的声音并准确地定位等复杂声信息处理时更是如此. 由于从声源到两耳距离的不同及传声途径中屏障条件的不同, 从某一方位发出的声音到达两耳时便有时间差 ITD(interaural time differences)和强度差 IID(interaural intensity differences), 在神经中枢对输入声信息进行整合时, 此时间差和强度差便是声源定位的主要依据[6]. 基于这种特性, 本文提出了一种基于声源定位的语音分离系统: 首先将混叠后的语音流在时频域上分解, 并求出各分解片段上的 ITD 和 IID 值. 由于各 ITD 和 IID 值与相应频率片段上的信号能量比呈单调递增关系, 因而通过阈值的比较可得出各自的掩蔽系数, 以此来判断各片段具体属于哪个声源或直接将干扰语音去除. 最后将属于同一声源的片段组合起来, 就可以得到分离后的语音.

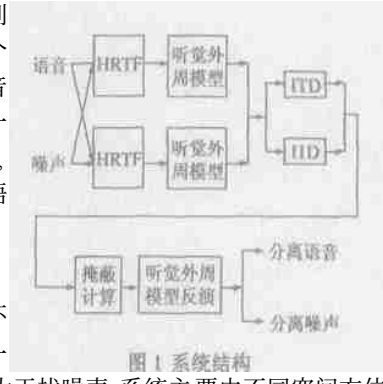

图 1 系统结构

2 系统结构

系统输入由两路不同方位的信号组成: 一路为语音信号, 另一路为干扰噪声. 系统主要由不同空间方位信号模拟、听觉外周模型、ITD 和 IID 估计. 掩蔽计算和听觉外周模型逆变换等部分组成. 结构如图 1 所示.

为了获取不同空间位置(不同得水平角和方位角)进入双耳的声信号, 可采用两种方法: 一是调整声源位置采用双通道同时采样; 二是输入信号经 HRTF(head-related transfer function)卷积获得不同空间位置的模拟输入. 由于 ITD 可达 10Ls 数量级, 为避免测量系统引起的误差, 本文采用第二种方法并用 MIT Media 实验室提供的 HRTF 系统冲激响应数据[7]对输入语音和噪声分别卷积后进行叠加, 以模拟不同空间方位输入至左右耳的混叠语音信号.

听觉外周模型首先由 128 个 Gammatone 滤波器[8](对应频

收稿日期: 2003-10-13; 修回日期: 2004-04-30
基金项目: 国家自然科学基金(No. 60172016)

率选择范围为80Hz至5kHz)对左、右耳混叠信号按时间帧进行频率分解,每个 Gammatone 滤波器的输出再经半波整流和饱和非线性处理,最后提取听神经发放率[2]. 听神经发放率与听神经纤维相应域频率范围内信号的强度大致成比例关系.

3　ITD、IID 估计和掩蔽计算

ITD 和 IID 是人的听觉系统用于确定声音位置方向的最基本的信息. 研究表明,在低频段(< 1. 5kHz) ITD 起决定作用,而在高频范围内主要由 IID 起作用[9, 10].

双耳时间差 ITD 的估计可通过听觉外周模型得到的两耳听神经发放率信号的互相关得到. 设左、右耳的发放率信号分别用 $p_l(i,t)$ 和 $p_r(i,t)$ 表示,其中 i 为频率通道, t 为时间点,则对于延迟 S 的互相关可用下式计算:

$$c(i,j,S)=\sum_{k=0}^{K-1} p_l(i,t-k)p_r(i,t-k-S)w(k) \qquad (1)$$

式中 w(#)取长度为 512 的矩形窗(采样率 44. 1kHz,对应约11. 6ms). 由文献[10]可知, ITD 的最大值约为 800Ls,所以取时间延时 S 的范围为- 1ms 到+ 1ms. 通过求取互相关最大值对应的时间延迟 S^i_{max}即可得到该时频片段上的 ITD 值.

设 L_i 为第 i 频率通道对应的双耳强度差 IID,它可用下式直接计算求得:

$$L_i = 20\log_{10}\frac{\sum_t p_l(i,t)^2}{\sum_t p_r(i,t)^2} \qquad (2)$$

由 ITD 和 IID 值可进一步确定掩蔽系数,以达到语音分离的目的. 为此,首先引入信号能量比的概念,并作如下定义:

$$E_{i,j}=\frac{\sum_t S^2_{i,j}(t)}{\sum_t S^2_{i,j}(t)+\sum_t n^2_{i,j}(t)} \qquad (3)$$

$\sum_t S^2_{i,j}(t)$、$\sum_t n^2_{i,j}(t)$ 分别表示第 i 频率通道和第 j 时间帧的语音信号、噪声信号对应的能量. 如果 $E_{i,j}> 0.5$ 表明语音大于噪声,则应当保留这个语音占主导的信号片段,反之,噪声占主导应当舍去.

为探讨 $E_{i,j}$ 与 ITD、IID 之间的关系,可在不同声源条件下分别计算出每个时频片段上的对应值并加以比较,为此,两个声源分别取二十个语音和二十个噪声在 5 种不同方位角(H_1, H_2)进行实验并计算. 结果表明,在低频端(< 1. 5kHz) $E_{i,j}$ 随 ITD 的增加单调递增,在高频端 $E_{i,j}$ 随 IID 的增加单调递增. 这一计算结果与心理声学预期的结论一致.

根据以上讨论,我们可直接由 ITD 和 IID 值来确定掩蔽系数. 听觉掩蔽效应表明,当两个强度不等的声音作用于人耳时,强度较高频率成份的存在会掩蔽强度较低的频率成份,使其变得不易察觉. 为简单起见,这里采用理想二值掩蔽,即直接保留强信号分量,舍去弱信号分量. 近期研究表明,采用类似理想二值掩蔽是可行的,作为噪声环境下语音识别系统的前端处理十分有利于提高系统的鲁棒性[11]. 为此对于第 i 频率通道、j 时间帧的掩蔽系数 F(i,j)由下式确定:

$$F(i,j)=\begin{cases}1 & \text{if}(f_i F\ f_c \text{ and } S_{max}(i,j) > T^{(S)}(i,j))\\ 1 & \text{if}(f_i > f_c \text{ and } L(i,j) > T^{(L)}(i,j))\\ 0 & \text{其它}\end{cases} \qquad (4)$$

式中 f_c 取 1. 5kHz, $T^{(S)}(i,j)$、$T^{(L)}(i,j)$分别为对应的 ITD 和 IID 阈值. 实验结果表明, ITD 和 IID 值取决于声源方位,与具体声音内容无关,这与已有的理论分析是吻合的. 图 2 给出了三组不同信号在同方位下对应的 128 个频率通道的 ITD 和 IID 曲线,其中(a)为男声浊音/ a0和噪声信号 noise1、(b)为男声清音/ s0和噪声信号 noise2、(c)为女声浊音/ e0和噪声信号 noise3.

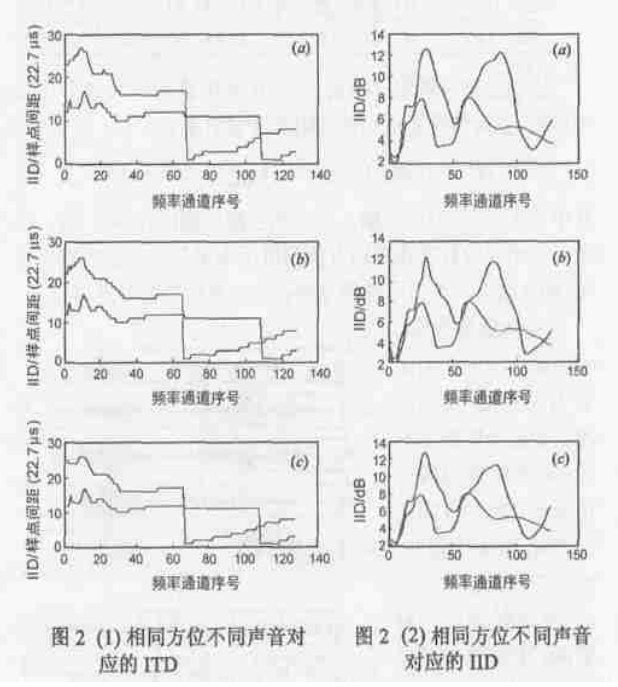

图 2 (1) 相同方位不同声音对应的 ITD　　图 2 (2) 相同方位不同声音对应的 IID

为确定计算掩蔽系数所需的 ITD 和 IID 阈值,在系统训练阶段对不同方位各采用 20 组混合语音数据计算出相应的 ITD 和 IID,然后取其平均值作为所需的阈值.

对输入的混合语音逐帧在各频率通道上求取掩蔽系数,即可形成掩蔽矩阵. 该掩蔽矩阵中相同元素(1 或 0)的位置决定了混合语音分量在各时频片段上的归属. 将属于同一信号源的各个片段按听觉外周模型的逆过程(包括听神经发放率、半波整流以及 Gammatone 滤波)进行反演运算即可获得各个信号分量[12],达到语音分离的目的. 以上过程完全体现了计算听觉场景分析的概念和新的思路.

4　实验结果

实验采用五组测试数据(分别用 A、B、C、D、E 表示),采样率为 44. 1kHz(选择较高采样率的目的是为提高计算 ITD 参数的精度),每组数据分别包含十个不同的语音信号和十个噪声(或另一干扰语音)信号. 然后各取一个语音和一个噪声按不同空间方位作混合输入. 其中 A 组:声源 1 为十个不同的汉语单词,分别由五位男生和五位女生发音,声源 2 为由收音机产生的十个噪声信号. B 组:声源 1 为五男五女发音的十个不同的汉语短句,声源 2 发音与 A 组相同. C 组:声源 1 与 A 组相同,声源 2 为十个强度不同的电话铃声. D 组:声源 1 与 B 组相同,声源 2 与 C 组相同. E 组:声源 1 与声源 2 均为由五男五女发音的十个不同的汉语句子. 实验中选取五组典型的不同空间方位角(H_1, H_2),如表 1 所示.

表 1 语音分离前后的信噪比

SNR(dB) / 方位角 H_1, H_2	A		B		C		D		E	
	前	后	前	后	前	后	前	后	前	后
0, 115	12.7	47.6	10.3	35.4	11.4	41.3	9.3	37.9	2.3	32.3
0, 90	11.8	30.4	10.7	25.3	12.3	30.6	8.9	23.8	3.5	22.1
0, 60	12.3	40.3	10.5	32.4	11.2	38.8	9.2	32.6	3.4	26.7
15, 60	11.2	39.2	12.3	32.5	9.9	40.3	12.0	34.9	0.2	25.3
135, 140	12.4	18.7	9.6	20.1	10.2	17.7	12.3	18.6	1.2	5.4

实验结果分别用信噪比变化、主观听觉和信号波形比较来评价. 分离信号的信噪比 SNR 按下式计算:

$$SNR(dB) = 10\lg\left(\sum_t s(t)^2\right) / \left[\sum_t [s(t) - \hat{s}(t)]^2\right] \quad (5)$$

其中 $s(t)$、$\hat{s}(t)$ 分别为输入信号和分离后的语音信号. 表 1 给出了五组实验数据在不同方位角情况下语音分离前后的信噪比. 图 3 给出了两个语音信号混叠经分离后的信号波形.

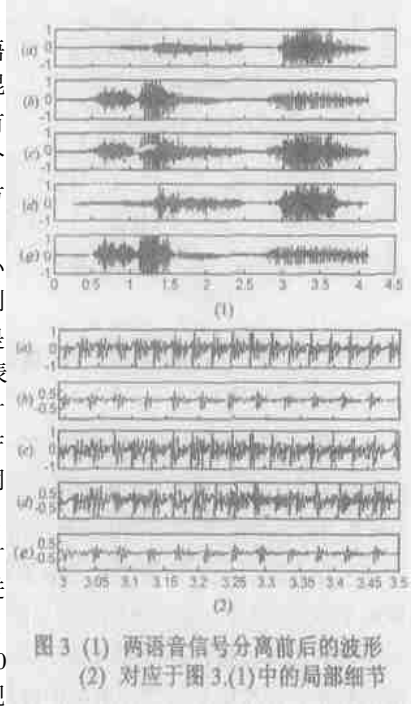

图 3 (1) 两语音信号分离前后的波形
(2) 对应于图 3.(1)中的局部细节

实验结果表明, 本文提出的方法对语音与噪声的分离、混叠语音的分离都是有效的, 特别是当两个声源有一定的空间方位差时, 效果比较好. 但当空间方位差较小时, 分离效果则受到影响, 其原因主要是当 H_1, H_2 很接近(如表中第五种情况)时, 计算 ITD 和 IID 容易产生偏差, 因而统计阈值有一定的离散性, 容易造成掩蔽系数计算错误. 为此我们进行了一些实验统计, 例如在方位角小于 10 度的情况下, 会出现 $E_{i,j} > 0.5$, 而 $S_{max}(i,j) < T^{(S)}(i,j)$、$L(i,j) < T^{(L)}(i,j)$ 或 $E_{i,j} < 0.5$, 而 $S_{max}(i,j) > T^{(S)}(i,j)$、$L(i,j) > T^{(L)}(i,j)$ 等情况, 且这种情况在中频段较严重, 达 12% 左右, 低、高频端的错误率分别在 3% 和 5% 以内. 事实上, 这种误差的出现也可用听觉现象来解释, 因为当两个方位相近的语音同时出现时, 人耳相对难以集中注意力辨识其中的一个声音. 另外, 本文提出的方法对混叠语音本身并没有限制, 它既适合于浊音, 也适合于清音, 在混叠信号信噪比不高的情况下也能有较好的分离效果. 主观听觉也表明, 各种不同情况下的混叠信号经分离后, 语音信号的清晰度和可懂度明显提高.

5 结论

本文提出了一种符合计算听觉场景分析概念的混叠语音分离的新思路和新方法. 该方法利用双耳听觉特性和掩蔽效应, 以 ITD 和 IID 作为感知要素实现混叠语音的分离, 取得很好的效果. 该方法既适用于浊音、又适用于清音, 对混叠语音的声源没有限制, 因而比现有的基于基音提取的语音分离方法更具普遍意义. 本文一个限制是混叠信号是经 HRTF 模拟得到的, 但这并不影响本文所提方法的有效性. 因为 HRTF 本身是经大量实验得到的, 这方面的工作已形成了声学研究的专门领域. 今后需进一步开展的工作有: (1) 声源方位较接近情况下 ITD 和 IID 估计准确性的提高. (2) 混叠语音推广至三个以上信号源. (3) 实际声源方位的具体确定以使系统实用化.

参考文献:

[1] D L Wang, G J Brown. Separation of speech from interfering sounds based on oscillatory correlation[J]. IEEE Trans, 1999, NN210(3): 684-697.

[2] G J Brown, M Cooke. Computational auditory scene analysis[J]. Computer Speech and Language, 1994, 8(24): 297-336.

[3] D F Rosenthal, H G Okuno. Computational Auditory Scene Analysis [M]. Mahwah: Lawrence Erlbaum, 1998.

[4] A J W Kouwe, D L Wang. A Comparison of Auditory and Blind Separation Techniques for Speech Segregation[J]. IEEE Trans, 2001, SAP29(3): 189-194.

[5] 梁之安. 听觉感受和辨别的神经机制[M]. 上海: 上海科技教育出版社, 1999. 119-130.

[6] W Roman, D L Wang. Speech segregation based on sound localization [A]. Proc IJCNN[C]. Washington DC: IEEE, 2001. 2861-2866.

[7] W Gardner, K Martin. HRTF measurements of a KEMAR[J]. J Acoust Soc Am, 1995, 97(6): 3901-3908.

[8] R Patterson, et al. An efficient auditory filterbank based on the gammatone functions[R]. APU Report No. 2341, Cambrige, Applied Psychology Unit. 1988.

[9] F Wightman, D Kistler. The dominant role of low-frequency interaural time differences in sound localization[J]. J Acoust Soc Am, 1992, 91(3): 1648-1660.

[10] J Blauert. Spatial Hearing-The Psychophysics of Human Sound Localization[M]. Cambridge: MIT Press, 1997.

[11] M P Cooke, P Green, L Josifovski and A Vizinho. Robust automatic speech recognition with missing and unreliable acoustic data [J]. Speech Comm, 2001, 34: 264-285.

[12] He-ming Zhao, Yong-qi Wang, Xu-qin Chen. Auditory model inversion and its application[A]. Proc. ICNNSP. 03[C]. Nanjin, China: ICNNSP, 2003, 868-871.

作者简介:

赵鹤鸣 男, 1957 年 8 月生于江苏无锡, 苏州大学电子信息学院教授, 博士生导师, 主要研究领域包括语音信号处理、神经计算和多媒体通信. E-mail: hmzhao@suda.edu.cn.

葛 良 男, 1977 年 7 月生于无锡, 硕士研究生, 主要研究方向为语音信号处理.

陈雪勤 女, 1974 年生于扬州, 博士研究生, 主要研究方向为语音信号处理、神经网络及其应用.