

基于听觉特性的迭代维纳滤波在 语音分离中的应用 *

卢 晶 沈春华 余 凯 徐柏龄

(南京大学声学研究所, 近代声学国家重点实验室 南京 210093)

2000 年 6 月 27 日收到

摘要 本文讨论了引入人耳听觉特性的迭代维纳滤波在语音分离中的应用, 即用矢量量化形成的码本反映目标话者的语音特征, 通过计算滤波结果与这一特征的匹配度来模拟人耳在“鸡尾酒会效应”中的注意力机制。实验结果表明这一方法有很好的效果。

关键词 语音分离, 迭代维纳滤波, 矢量量化

Iterative Wiener filtering based on the auditory characteristics for speech separation

Lu Jing, Shen Chunhua Yu Kai Xu Boling

(National Key Lab of Modern Acoustics, Institute of Acoustics, Nanjing University 210093)

Abstract This paper discusses the application of Iterative Wiener Filtering(IWF) with Auditory Characteristics in Speech Separation. A codebook formed by vector quantization is used to represent the speech characteristics of the desired speaker, and human's attention in "Cocktail Party Effect" is simulated by calculation of the matching degree between the filtering result and the codebook. Satisfactory speech separation results are obtained.

Key words Speech separation, Iterative Wiener Filtering, Vector quantization

1 引言

语音分离是语音信号处理中一个重要的研究领域, 其目的是在有多个话者的语音同时存在时提取特定话者(目标话者)的语音, 这对于鲁棒性语音识别以及信息抽取均具有十分重要的意义。人类的听觉系统在处理这类问题时,

一个很重要的特点便是注意力机制的引入, 即迅速提取出其所感兴趣的目标话者语音特征并将注意力集中于目标话者上, 这也就是我们通常所说的“鸡尾酒会效应”。由于语音信号的非平稳性, 传统的语音增强方法如自适应滤波^[1], 频谱减^[2]等方法在处理语音分离问题时存

* 国家自然科学基金资助项目 (69872014)

在诸多困难, 都要求干扰信号为平稳噪声。而根据空间特性所构成的话筒阵列^[3]虽然有一定的效果, 但当竞争话者与目标话者靠得很近时, 其性能会大大下降。如果我们能在语音分离的处理方法中模拟听觉系统的注意力机制, 可以大大提高语音分离系统的处理能力。基于这种考虑, 我们在迭代维纳滤波中引入目标话者特征以及匹配度计算, 有效地改进了语音分离算法。

2 迭代维纳滤波

Lim, Oppenheim, Hansen 等人所使用的迭代维纳滤波^[4]是一种基于全极点语音参数估计增强方法。这种滤波方法主要对在加性高斯噪声背景中的语音作最大后验概率 (MAP) 估计。滤波器的频率响应函数为

$$H(\omega) = \frac{P_{s_1}(\omega)}{P_{s_1}(\omega) + P_{s_2}(\omega)}$$

其中 $P_{s_1}(\omega)$ 和 $P_{s_2}(\omega)$ 分别是目标话者语音和噪声的功率谱密度函数。在实际应用中, $P_{s_2}(\omega)$ 可以在语音间隙的噪声段中获得, 而 $P_{s_1}(\omega)$ 却无法直接获得, 因此 Lim 等人提出了一个准最优迭代过程, 即把含噪语音作为第一次迭代的目标话者语音, 进行维纳滤波, 每一次滤波的结果作为下一次迭代的目标话者语音。在迭代维纳滤波中引入时变参量, 使之运用于语音分离, 其原理如图 1。

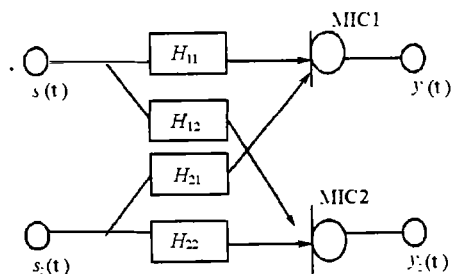


图 1 双话筒接收系统

在图 1 所示的双话筒接收系统中, 在传统维纳滤波方程中引入时变参量 α_i , 得到新的

滤波器形式如下:

$$H^{(i)}(\omega) = \frac{(1 + \alpha_i)P_{s_1}^{(i)}(\omega)}{P_{s_1}^{(i)}(\omega) + \alpha_i[|H_{11}(\omega)|/|H_{21}(\omega)|]^2 P_{y_2}(\omega)}$$

其中 $P_{s_1}^{(i)}(\omega)$ 和 $P_{y_2}(\omega)$ 分别是 $s_1(t)$ 和 $y_2(t)$ 的功率谱, 在每一帧迭代的开始, 用主通道信号 $y_1(t)$ 作为 $s_1(t)$ 的初值。 $H_{11}(\omega)$ 和 $H_{21}(\omega)$ 分别是目标话者所处位置到两个话筒的声场传输函数, 它们的比值可以在无竞争话者语音存在时计算出来; 而我们进行滤波的主要原因就是为了消除 $H_{21}(\omega)$ 和 $H_{11}(\omega)$ 这两个场的存在所导致的干扰信号。对于每一次迭代, 有

$$S_1^{i+1}(\omega) = S_1^{(i)}(\omega)H^{(i)}(\omega), \quad i = 0, 1, 2, \dots$$

时变参量 α_i 的初值 α_0 与两个通道所接收信号的信扰比有关, 通常取:

$$\alpha_0 = (0.01, \dots, 0.1) \left(\frac{A_{SIR_1}}{A_{SIR_2}} \right)^{-2}$$

其中 A_{SIR_1} 和 A_{SIR_2} 分别表示主从话筒接收到信号的信扰比, 它们的比值可以由只有目标话者语音存在和只有竞争话者语音存在这两种状态下的话筒接收信号估算出来。对于每一次迭代, α_i 可由如下公式进行更新:

$$\alpha_i = \frac{\alpha_0}{(1 + \alpha_0)^{i-1}}, \quad i = 1, 2, 3, \dots$$

在语音分离的过程中, 第一次迭代就用主话筒接收的信号 $y_1(t)$ 作为目标话者语音 $S_1^{(0)}(t)$ 的估值, 每一次滤波的结果作为下一次迭代的目标话者语音估值。时变参量 α_0 的引入能使迭代过程更有效地逼近最优滤波结果。迭代维纳滤波应用于语音分离中的一个重要问题是: 在开始阶段, 随着迭代次数的增加, 混合语音中竞争话者语音会受到明显抑制, 逐步向目标话者语音靠拢; 一旦超过了一定的迭代次数便会出现过滤波的现象, 这时候虽然竞争话者语音被抑制了, 但目标话者的语音会有较

大的失真, 可懂度会迅速下降。为了解决这个问题, 我们必须引入有效的收敛判据, 即对滤波的结果进行比较判断, 当某一次迭代滤波的结果最接近目标话者的原始语音时, 便停止这一语音分离过程。

3 基于听觉特性的迭代维纳滤波

如何在迭代维纳滤波中引入听觉系统的注意力机制从而形成有效的收敛判据? 矢量量化所形成的能够反映目标话者语音特征的码本是一种很好的解决方案。它的主要原理就是将每一次迭代的结果同目标话者语音特征相匹配, 如果在某一迭代时达到了最佳匹配, 我们就可以认为滤波结果达到最优。

3.2 码本的生成

采集一段目标话者的纯净语音, 分帧加窗计算出一系列 LPC 或 LPC 倒谱参数作为原始矢量准备用于聚类。聚类算法思想与 Linde 等人提出的经典 LBG 算法^[5]一致, 其目标是使所有的原始矢量相对于聚类中心(通常是 2^N 个)的平均失真度最小。主要步骤如下:

- ① 设定码本大小及一些与聚类有关的初值。
- ② 计算出所有矢量的中心作为初始最优码本。
- ③ 码本分裂, 即对于已经生成的码本, 对每个矢量进行大小为 ϵ 的微扰, 形成码本大小为 2ϵ 的新码本。
- ④ 以新码本为起点, 按照最近邻原则寻找最优码本
- ⑤ 重复步骤③和④直到码本大小达到预设值。

聚类的结果就作为反映目标话者特征的码本。

3.2 语音匹配度的衡量

通常两段语音的相似度用相应的语音参数距离来衡量, 我们为了衡量滤波结果与目标话者语音特征的匹配度, 同样要用到语音参数距离。对于 LPC 倒谱参数, 我们可以用欧氏距

离加以衡量, 即

$$d(X, Y) = \sqrt{\sum_{i=0}^{k-1} |x_i - y_i|^2} \quad (k \text{ 为倒谱阶数});$$

对于 LPC 参数, 则可以用经典的坂仓 (IS) 距离加以衡量, 即 $d(X, Y) = (X - Y)R(X)(X - Y)'$, 其中 $R(X)$ 是 LPC 参数 X 所对应的自相关矩阵。

在语音分离的过程中, 对于每一步迭代维纳滤波的结果, 生成相应的语音特征参数 (LPC 或 LPC 倒谱参数), 将每一帧参数与前面生成的码本进行比较, 找到其对应的码字并计算距离, 最后求得平均距离便可以作为本次滤波结果与目标话者语音特征的匹配度。根据迭代维纳滤波的特性, 当某一次滤波结果对应的匹配度达到最小值时, 就可以认为该次滤波结果是最优的。这一过程与人的听觉系统将注意力集中于某一特定话者的语音的过程相似。

4 仿真试验及其结果的客观评价

根据上述分析, 我们对引入听觉注意力机制的迭代维纳滤波用于语音分离进行了仿真试验。试验中, 语音采用 16 位 8kHz 采样, 每一帧语音长度为 256 点 32ms, 帧移 64 点, 加 Hamming 窗计算 LPC 倒谱参数, 生成大小为 512 的码本。

为了检测语音匹配度判据模拟听觉注意力机制的有效性, 一方面我们根据人耳的主观感觉进行判断, 另一方面必须引入有效的客观判据。目前国际上常用 Itakura-Saito (IS) 相似度测量^[6]作为客观评价标准, 即将分离后的语音与纯净的目标话者语音进行比较, 计算相应 LPC 参数的 IS 距离; 我们在本实验中, 还引入 LPC 倒谱参数线性距离客观判据, 即计算分离后的语音与纯净语音的 LPC 倒谱参数距离。

图 2, 图 3 和图 4 是一组典型的语音匹配度收敛判据和两个客观判据的比较。其中实线均表示语音匹配度对迭代维纳滤波的约束效果,

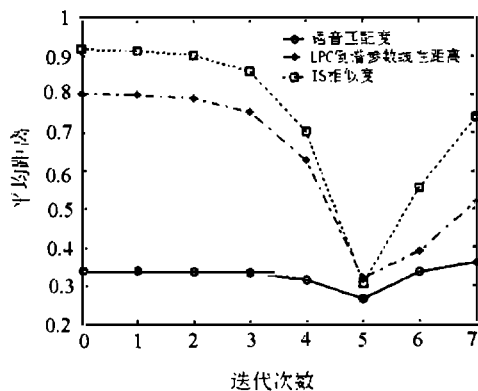


图2 两个男性话者语音分离收敛判据

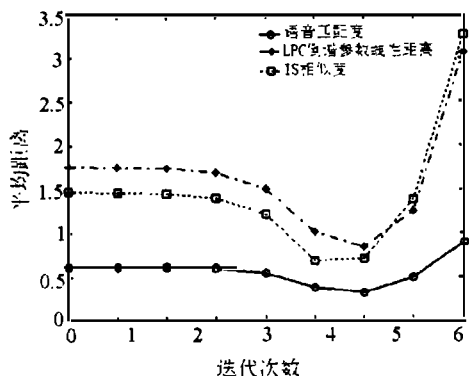


图3 两个女性话者语音分离收敛数据

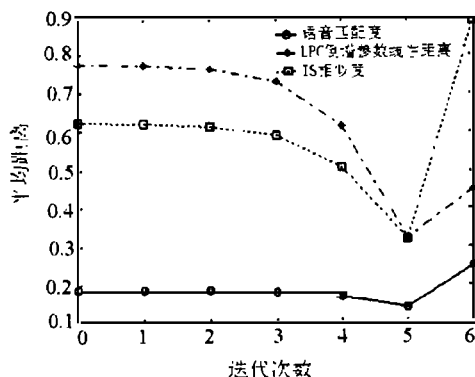


图4 两个异性话者语音分离收敛判据

点划线为 LPC 倒谱参数线性距离客观判据, 虚线为 IS 相似度判据。

图5中第2, 3, 4个波形分别对应于图4两个异性话者语音分离过程中第三次, 第五次和第六次迭代滤波的结果, 同第1个波形

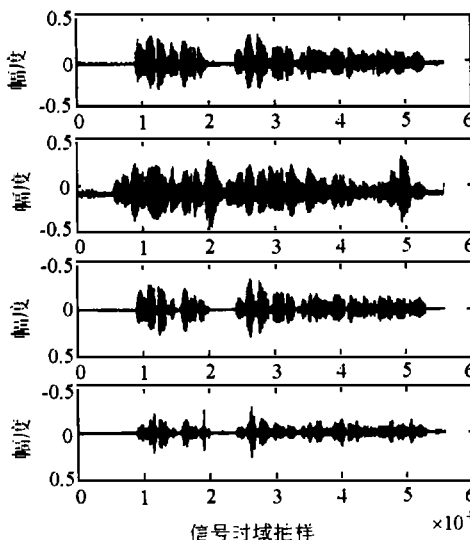


图5 语音分离波形图

(纯净目标话者语音)相比较不难看出第五步迭代确实对应于时域上的最优结果。

可以看出, 用反映目标话者特征的码本及相应的匹配度计算作为迭代维纳滤波的收敛判断是十分有效的, 在主观听觉上, 当滤波结果与目标话者语音特征匹配度最高, 即平均码本距离最小时, 如图2中的第五步迭代结果, 分离出的语音竞争话者语音的抑制达到最大, 同时又能保证目标话者语音具有良好的可懂度; 同两个客观判据的比较也证明了滤波结果与目标话者语音特征的匹配度变化曲线完全能够表征迭代维纳滤波用于语音分离的效果。

表1给出主通道不同信扰比情况下滤波结果对信扰比的改善(从通道信扰比: -12.68)。

表1 主通道不同输入信扰比下的目标语音分离信扰比变化情况

输入信扰比	输出信扰比	信扰比改善
-8.80	1.02	9.82
-7.64	1.90	9.54
-6.31	4.70	11.01
-4.72	6.23	10.95
-1.62	8.27	9.89
1.30	12.26	10.96

5 结束语

语音分离一直是语音信号处理中的一个难点,本文提出的基于听觉特性的迭代维纳滤波比较有效地解决了这一问题。我们引入反映目标话者特征的码本对滤波结果进行匹配度计算,在一定程度上模拟了人类听觉系统在解决“鸡尾酒会效应”问题中的注意力机制。

参 考 文 献

- 1 Zelinski R. ICASSP, 1998: 2578-2581.
- 2 Boll S F. *IEEE Trans on ASSP*, 1979, 2: 113-120.
- 3 Berkley D A, Mitchell O M. *Bell labs. Rec.*, Vol.52, 1974: 318-324.
- 4 Lim F S, Oppenheim A V, *IEEE Trans on ASSP*, 1978, 3: 197-210.
- 5 Linde Y, Buzo A, Gray R M. *IEEE Trans on Communication* 1980, 1: 84-95.
- 6 Itakura F. *IEEE Trans on ASSP*, 1975, 2: 67-72.

“首届高强度聚焦超声 (HIFU) 技术在医学中应用 国际学术研讨会”在重庆召开

HIFU 技术被誉为 21 世纪治疗肿瘤的新技术,是当今国际医学超声领域发展的一个突出的研究热点。我国在这一研究领域取得了国际公认的领先地位。

由中国有关单位主办,英国、美国有关单位协办,重庆医科大学承办的“首届高强度聚焦超声(HIFU) 技术在医学中应用国际研讨会”(First International Workshop on the Application of High Intensity Focused Ultrasound in Medicine) 于 2001 年 5 月 10-13 日在重庆市成功召开。会议计收到论文 85 篇,会上报告 47 篇,“会议论文集”收编了 74 篇。这些论文比较全面地反映了当今 HIFU 技术及其在医学中应用研究的最高水平。有 150 名代表参加了会议,其中国内代表 120 人,海外代表 30 人,后者分别来自英国、美国、法国、德国、俄罗斯、加拿大、以色列、澳大利亚、多哥等十多个国家。

本次研讨会显示出重庆医科大学以王智彪教授为首的科学工作者们在 HIFU 基础研究和临床应用和产业化开发中拥有的明显优势,其卓越成就获得与会代表高度赞誉和肯定。例如,国际声学委员会主席、美国华盛顿大学 L.A.Crum 教授在发言中讲到:“中国作为一个发展中国家,已在 HIFU 领域作出了令人叹服的成绩,中国重庆科学家们在 HIFU 的医学应用水平方面居世界领先地位”。

本次研讨会还充分展示了 HIFU 技术在肿瘤无创治疗领域的广阔前景和亟待进一步研究的现状。会后,在中、英、美等多国专家充分磋商的基础上,组织成立了“国际治疗超声学会(International Society for Therapeutic Ultrasound) ”(待批)。该学会由常务理事会负责管理工作,常务理事会中包括我国重庆医科大学王智彪教授、伍烽教授和南京大学冯若教授。英国皇家马斯顿医院 Gail ter Haar 博士被推选为学会主席,伍烽教授被推选为秘书长,秘书处设在重庆医科大学。

会议期间,重庆医科大学和重庆海扶(HIFU) 技术有限公司与英国牛津大学和英国超声治疗公司还就英方购买重庆 HIFU 肿瘤治疗设备正式签订了销售合同。这是我国具有完全自主知识产权的大型医疗设备首次向西方出口。

在大会闭幕式上,国际治疗超声学会主席 Gail ter Haar 博士宣布下届 HIFU 国际研讨会定于 2002 年 7 月 29 日-8 月 1 日在美国西雅图市召开。

与会代表对这次大会的成功组织、召开和作用给予了高度评价。中央电视台一台和四台分别报道了此次会议的召开,新华社、人民日报、光明日报、科技日报、健康报、China Daily 等均及时报道了这次会议的情况。

(声讯)