语音信号的 Bark 子波变换及其在语音识别中的应用

付 强,易克初
(西安电子科技大学 ISN 国家重点实验室,西安 710071)

摘 要: 本文首先在语音感知实验基础上所得到的 Bark 尺度概念与子波的构造联系起来,提出了一种语音信号 Bark 子波及其变换. 在感知特性中,它与人耳的听觉系统是直接吻合的:在数学上,它是一种非正交的但却是超完备的,可逆且是自反演的子波变换. 应用在语音识别中时,利用子波的局部基性质,取较小的分析帧长,提高了前端处理的时间分辨率. 实验表明,所形成的新特征在对辅音区别能力方面明显优于目前应用于广泛的 MFCC 特征.

关键词: 子波变换; 语音识别; 特征提取
中图分类号: TN912 **文献标识码**: A **文章编号**: 0372-2112 (2000) 10-0102-04

Bark Wavelet Transform of Speech and Its Application in Speech Recognition

FU Qiang, YI Ke-chu
(Xidian University, National Key Lab. of ISN, Xi'an 710071, China)

Abstract: This paper firstly describes a novel perceptual wavelet and its transform of speech, which is based on Bark frequency scale concept coming from the speech perceptual experiment. In the nature of physics, it consists with human's cochlea filter properties; in the sense of mathematics, Bark wavelet transform is a kind of non-orthonormal but overcomplete, reversible and self-inverting wavelet transform. When it is applied in speech recognition, small analysis frame-length is used by using the property of local base of wavelet transform, which can improve the time-domain resolution of front-end process. Experiment shows that the formed novel feature is superior to MFCC in the capability of distinguishing the consonants.

Key words: wavelet transform; speech recognition; feature extraction

1 引言

传统的语音分析在短时平稳假定的基础上,采用固定窗傅里叶变换以获取时-频局部化信息. 不论对于任何音段,其时-频分辨率均是固定不变的,由不确定性原理可知[1],其时频分辨率不可能同时都很高. 其结果无疑将会模糊语音的细节特征,特别是对于那些平稳很差的"速变"语音段,如爆破音和塞擦音等,它将引起一些非常重要的信息丢失. 子波变换采用多辨率分析的思想,非均匀地划分时-频空间,频率较低的信号成分的持续周期较长,允许而且应当采用较低的时间分辨率和较高的频率分辨率进行分析,而频率较高的信号成分则正好相反. 这样对于频率成分复杂的语音信号,在服从不确定性原理的前提下,使不同的时-频区都能获得比较合适的时-频分辨率.

正因为这一明显的优点,近几年来在基于子波的语音分析这一课题上已有许多学者作了大量工作[2~5],但几乎都是直接采用了现有的子波基函数进行分析和应用. 在多分辨率分析框架下构造子波基的主要优点在于它满足正交性(包括半正交、双正交性)[6],而这一点对语音分析来说并不必要,因为听觉感知本身就存在较大的冗余. 并且,无论是二进子波、子波包还是 M 带子波变换,其频域划分都是一种倍频程关系,这与人耳所固有的对语音的频域感知特性并不完全吻合[7]. 有理采样子波(Rational Sampling)以及多子波变换(Multi-Wavelets)虽采用多个子波函数来获得更灵活的频域划分,但随着所需频带数的增加,其设计过程也急剧复杂起来[7].

为此,本文提出一种针对语音信号的 Bark 子波及其变换. 其基函数满足时间-感知频率(弯折频率,主观频率)上的最佳不确定性(时间-带宽积最小[1]),而分析尺度的伸缩则按照"临界带(critical band)"的概念来变化,并使得每一尺度下的带宽为一个"频率群"[8]. 通过这样的构造方法所得到的子波变换无疑将具有与听觉系统十分吻合的分析特性. 本文进一步证明了 Bark 子波是非正交但却是起完备的,可逆且是自反演的变换,即构成紧框架. 由于子波变换的局部化性质,它可以对语音信号在很小的分析帧长下仍具有很好的谱分辨率,这一点应用在语音识别系统的特征提取过程,可使系统对辅音的区别性增强,与基于短时傅里叶变换的 Mel 频率倒谱(MFCC)分析法对比实验验证了这一点.

2 Bark 子波的导出及变换定义

人耳基底膜具有与频谱分析器相似的作用,在 20 -

收稿日期:1999-04-21;修回日期:2000-01-11
基金项目:国家自然科学基金

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

16,000Hz 范围内的频率可分成 24 个频率群(临界带,Bark),频率群的划分相应于基底膜分成许多很小的部分,每一部分对应一个频率群,并且长度相等. 对应于同一基底膜部分的那些频率的声音,在大脑中似乎是叠加在一起进行评价的[8]. 这就是说,人类听觉系统对于声音频率的感知与实际频率的对应关系,是一种非线性映射关系,这就引出了所谓的 Bark 尺度的概念. Traunmullar 给出了线性频率与 Bark 频率之间的函数关系[9],即

$$b = 6.7\,a\sinh[(f - 20)/600] \tag{1}$$

其中,b 代表 Bark 频率,f 为线性频率.

以下即是构造 Bark 子波的基本思想:第一,由于在语音分析中时间与频率信息的同等重要性,所选择的子波母函数(以下简称母函数)应在 Bark 域满足时间-带宽最小,即为 Bark 域的高斯函数[1];第二,为与频率群的概念相一致,母子波在 Bark 域的带宽应相等且为单位宽度,即 1 个 Bark.

根据以上两点,选择母函数在 Bark 的形式为 $W(b) = \mathrm{e}^{-c_1 b^2}$,并且在单位带宽定义为 3dB 带宽的情况下 c_1 取为 4ln2. 假定被分析语音信号 $s(t) - S(f)$ 的线性频率带宽 $|f| \in [f_1, f_2]$,相应的 Bark 频率带宽为 $[b_1, b_2]$. 定义子波函数 Bark 域形式为

$$W_k(b) = W(b - b_1 - k\Delta b),\ k = 0,1,\cdots,K-1 \tag{2}$$

式中,Δb 为 $W_k(b)$ 的平移步长,根据 Bark 域等带宽原则,有 $\Delta b = (b_2 - b_1)/(K-1)$,$k$ 为尺度参数. 因为 $W(b) = \mathrm{e}^{-c_1 b^2}$,所以

$$W_k(b) = \mathrm{e}^{-4\ln 2(b - b_1 - k\Delta b)^2} = 2^{-4(b - b_1 - k\Delta b)^2},\ k = 0,1,\cdots,K-1 \tag{3}$$

再将式(1)带入式(3),就得到了线性频率下的 Bark 子波函数表达式

$$W_k(f) = c_2 \cdot 2^{-4[6.7\,a\sinh([f - 20]/600) - (b_1 + k\Delta b)]^2} \tag{4}$$

这里 $b = 6.7\,a\sinh[(f_1 - 20)/600]$. 式中 c_2 为规整因子,如何取值将在下节中说明.

进一步可以在频域定义 Bark 子波变换为

$$\begin{aligned} s_k(t) &= \int_{-\infty}^{\infty} S(f)\cdot W_k(f)\cdot \mathrm{e}^{j2\pi ft}\,dt \\ &= \int_{-\infty}^{\infty} c_2\cdot S(f)\cdot 2^{-4[6.7\,a\sinh((f-20)/600 - (b_1 + k\Delta b))]^2}\cdot \mathrm{e}^{j2\pi ft}\,df \\ &= c_2 \int_{-\infty}^{\infty} S(f)\exp\{-4\ln 2\cdot[6.7\,a\sinh((f-20)/600 \\ &\quad - (b_1 + k\Delta b))]^2 + j2\pi ft\}\,df,\ k = 0,1,\cdots,K-1 \end{aligned} \tag{5}$$

Bark 子波的时域形式,可以将式(4)用逆傅里叶变换变换到时域得到,并且若 $W_k(-f) = 0$ 则得到复数形式子波函数;若

$$\mathrm{Im}[W_k(f)] = 0;\ \mathrm{Re}[W_k(f)] = \mathrm{Re}[W_k(-f)]$$

或　　$$\mathrm{Re}[W_k(f)] = 0;\ \mathrm{Im}[W_k(f)] = \mathrm{Im}[-W_k(-f)] \tag{6}$$

则分别对应于实偶和实奇两种形式的子波函数. 这里所用的是实偶形式子波函数,以下也仅以该种情况为例讨论. 图 1 给出了这种情况下的 Bark 子波在 $k = 5$ 时的时域和频域波形.

3　有关性质讨论

从感知域滤波器组角度导出了 Bark 子波函数的定义及其变换式. 容易验证

$$\int_{-\infty}^{\infty} \frac{\left|\mathrm{e}^{-c_1 b^2}\right|^2}{|b|}\,db < \infty \tag{7}$$

而又由于式(1)是单调函数,所以 Bark 到线性频率的转换并不影响式(7)的成立,故此 Bark 子波是满足"可容许性条件"的,也就是说 Bark 子波变换是可完全重构的. 实际上如果从工程近似的角度出发,可以简单地验证这一点. 首先,认为式(3)中的 c_2 可近似认为是一个常数,并由下式给出

$$c_2 \cdot \sum_{k=0}^{K-1} W_k(b) = 1,\ 0 < b_1 \le b \le b_2 \tag{8}$$

并注意到线性频率与 Bark 频率之间的并不影响式(8)的成立,于是就有

$$\begin{aligned} \sum_{k=0}^{K-1} s_k(t) &= \sum_{K=0}^{K-1} \int_{-\infty}^{\infty} W_k(f)\cdot S(f)\cdot \mathrm{e}^{j2\pi ft}\,df = \int_{-\infty}^{\infty} \sum_{k=0}^{K-1} W_k(f) \\ &\quad \cdot S(f)\cdot \mathrm{e}^{j2\pi ft}\,df = \int_{-\infty}^{\infty} S(f)\,\mathrm{e}^{j2\pi ft}\,df = s(f) \end{aligned} \tag{9}$$

于是可以把式(8)称之为 Bark 子波在工程意义上的完全重构条件. 总之,Bark 子波变换是可逆的且是自反演的、非正交的但却是超完备的,也即 Bark 子波簇构成了紧框架[6]. 这一性质对 Bark 子波变换在语音编码和合成等领域的应用尤其是有益的.

图 2 是利用 Bark 子波对一段从汉语塞音/d/到元音/i/过渡的语音信号分解和重构图. 图 2(a)为原始语音时域波形,图 2(b)为该段语音在 21 个尺度下的时间-尺度图,可以看到 Bark 子波很好地表现出了该过渡音的从非平稳段到平稳段的过渡过程,这是短时傅里叶变换所无法实现的. 图 2(c)是将 Bark 子波的在 21 个尺度下的分解信号叠加而成的重构信号波形,不难看出,除起点处由于子波边界效应造成的误差外,它与原始信号非常接近.

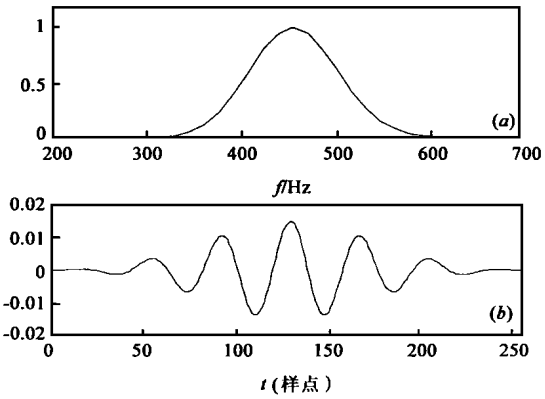

图 1　$k = 5$ 时的 Bark 子波时频域波形. (a)频域波形;(b)时域波形

4　Bark 子波变换在语音识别中的应用

4.1　特征参数的形成

在语音识别应用中,提高其前端处理的时间分辨率一直是极受关注的问题. 由于传统的特征参数一般都是在分析帧

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

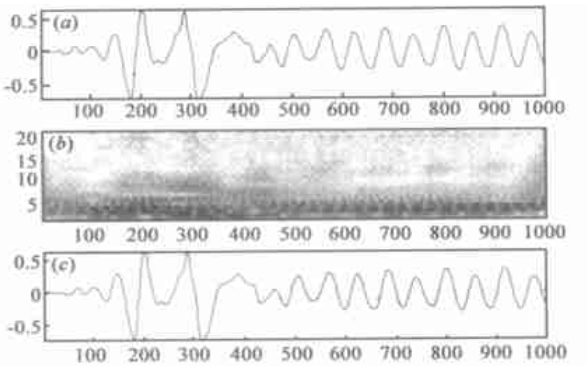

图 2 利用 Bark 子波对一段过渡音进行分解和重构

长范围内的时间平均值(例如 *MFCC* 参数[10]),所以实际上帧长的大小就决定了识别系统的时间分辨率,于是较小的帧长就意味着较高的时间分辨率. 傅里叶分析在帧长较小时受到窗函数的影响很大,无法提供足够的频域分辨率. 而子波变换由于无需窗函数以及具有局部化(有限支撑)的基函数,在排除边界效应的影响后,帧长较小时仍然能够给出很好的频域信息. 实际的特征参数由下式计算

$$c_n = \sum_{k=1}^{K} \{\log(E_{\mathrm{TEO}}^{k}) \cdot \cos[(k-0.5)n\pi/K]\},\ n=0,1,\cdots,M \tag{10}$$

M 一般可取 12～16. 注意,c_0 反映的是该帧语音 Teager 能量的和,可根据实际情况取舍.

图 3(*a*)示出的是一段男声"大地"/dadi/音的时域波形. 图 3(*b*)为用上述方法对该段信号分帧计算的特征再进行逆 DCT 变换所得到的动态 Teager 能量谱图,它能够比较可靠地反映用上述方法提取的特征参数(14 维)所真正含有的信息. 图 3(*c*)、(*d*)是用同样原理对 Mel 倒谱(倒谱系数 14 维)求逆 DCT 变换得到的动态谱图. 图 3(*b*)、(*c*)帧长为 2.2ms. 图 3(*d*)帧长为 25ms,是傅里叶变换分析语音的典型值.

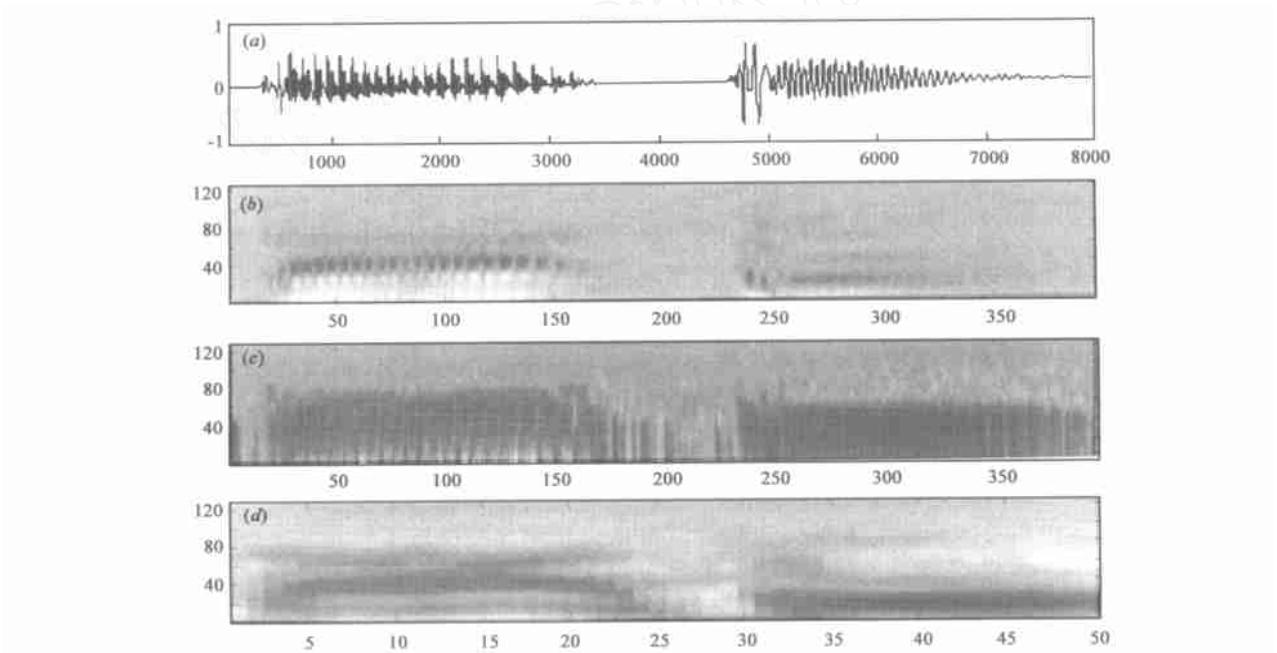

图 3 基于 Bark WT 的 Teager 能量表示与 Mel 倒谱瞬时能量表示的对比

由图 3 不难看出,在分析帧长很短的情况下,新特征的动态谱对清音的时域局部化特征及共振峰结构都能够给予比较清晰的反映. 而基于傅里叶分析的 Mel 倒谱所形成的动态谱,在帧长较短时,虽然其时域局部性较好,但共振峰却很模糊;帧长较长时可使其共振峰结构清晰起来,但清音的时域局部性也丢失了.

4.2 实验结果

作者分别用新的特征参数和在语音识别系统应用广泛的 MFCC 特征参数进行了孤立输入方式的语音识别对比实验. 均取 14 维静态特征,不含能量. 第一组识别对象为汉语易混集-/a/集的所有元素,即/a/、/jia/、/ba/、/fa/、/la/、/sha/、/ta/、/ya/、/zha/、/cha/、/ma/、/sa/、/ga/、/za/、/ha/、/ka/、/wa/、/da/、/ca/、/pa/、/na/,共 21 个单音节. 采样率 16kHz. 识别系统采用混合数为 2 的分划高斯密度连续 HMM 模型,用 MFCC 特征时状态数取 6,新特征时状态数取 20. 共训练 30 遍,男 15 人,女 15 人,每人一遍. 识别时采用 Viterbi 算法对所有训练集内数据进行识别. 表 1 是实验结果:

表 1 新特征与 MFCC 对汉语易混集-/a/集的识别实验结果

	男声平均识别率	女声平均识别率	总体平均识别率
MFCC	74.0 %	78.4 %	76.2 %
新特征	87.8 %	89.6 %	88.7 %

在表 1 中实验数据的差别主要是辅音的差异,而新特征的识别率较之 MFCC 同样有了明显提高,这就说明新特征参数在易混语音条件下的对辅音信息的区别能力比 MFCC 特征要强.

第二组识别对象是 57 个汉语词汇,包括 7 个单字词,4 个

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

双字词,10 个数字(0～9)和 26 个英文字母. 采样率 16kHz. 用 MFCC 特征时 HMM 模型的状态数为 8,用新特征时状态数 25. 共训练 36 遍,男 6 人,每人一遍. 所有训练集内数据进行识别,其结果是 MFCC 特征的平均识别率 87.7%,新特征为 92.9%.

5　结论

Bark 子波变换,由于设计思路上的原因,在感知特性上是与人耳的听觉系统特性(这里主要指 Bark 尺度特性)是十分吻合的;在数学上它是一种非正交的但却是超完备的,近似可逆且是自反演的子波变换. Bark 子波变换用于语音识别的特征提取过程. 所构造的新的特征与 MFCC 特征的对比实验表明,新特征对具有明显的非平稳特性的辅音的分辨能力比 MFCC 特征要强的多,这正是子波变换所应该具有的优势. 当然,这种特征由于其 HMM 模型的状态数比 MFCC 特征要多,将会使得搜索算法的复杂度有所增加.

参考文献:

[1] [美]L. 科恩. 时-频分析:理论与应用 [M]. 西安交通大学出版社,1998.

[2] D'Alessandro, C. Auditory-based wavelet representation [R]. Visual Representation of Speech Signals, John Wiley & Sons, London, 1993.

[3] Dryjalgo, A. Multiresolution time-frequency speech proceeding based on orthogonal wavelet packet pulse forms [C]. Proceeding of the Conference of Eurospeech'93, 21 - 23 September, Berlin, Germany, 1993: 147 - 150.

[4] Irino, T. & Kawahara, H. Signal reconstruction from modified auditory wavelet transform [J]. IEEE SP 41, 1993: 3549 - 3554.

[5] Pitton, J. W Wang, K, S Juang B, W. Time-Frequency Analysis and Auditory Modeling for Automatic of Speech [C]. Proc. of IEEE, 1996.

[6] Daubechies. I. Ten Lectures on Wavelets [M]. Philadelpha: SIAM, 1992.

[7] Evangelista, G. Cavaliere, S. Discrete frequency warped wavelets: theory and apllications [J]. IEEE Trans. on Signal Processing, April 1998, 46 (4).

[8] 实验语音学讲义 [M]. 中国社会科学院语言研究所语音研究室,1986.

[9] Traunmullar, H. Analytical expression for the tonotopic sensory scale [J]. Journal of the Acoustical Society of America, 1990, 88: 97 - 100.

[10] S. B. Davis & P. Mermelsten. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences [J]. IEEE Trans. on ASSP, August 1980, ASSP-28 (4).

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net