一种基于小波变换和隐 Markov 模型的声调识别方法 [1]

程 俊 易克初 李兵兵

(西安电子科技大学综合业务网国家重点实验室 西安 710071)

摘 要 本文给出了一种基于小波变换和隐 Markov 模型 (HMM) 的声调识别方法。根据小波变换检测信号突变的性质，充分利用多分辨率分析，准确可靠地实现了基音检测；采用分划 Gauss 混合 (PGM) 概率密度函数的 HMM 进行汉语声调识别，推导出用 PGM 函数的 Viterbi 算法的简化递推式。在匹配计算量大大减小的情况下，特定人的四声识别率为 97.22%，非特定人达到 94.47% 。

关键词 基音检测，声调识别，小波变换，隐 Markov 模型

中图号 TN912.3

1 引 言

全汉语普通话音节如不考虑轻声可归为阴平、阳平、上声、去声四种声调。声调基本上由语音的基频包络决定。基频 (音) 检测和识别判决是汉语声调识别的两个阶段. 自相关法、倒谱法等经典的基音检测方法存在固有不足，如只能估计出帧内基音周期的平均值、分析帧长与基音周期长短有关等 [1] . 本文采用一种基于小波变换的基音检测方法，求得的基音值准确可靠；HMM 可用于四声识别 [2] 。文中采用 PGM 概率密度函数 (pdf) 的 HMM，与传统的 Gauss 混合 (GM)pdf 的 HMM 相比，计算量大大减少，而识别精度仍在可接受范围内。

2 小波变换和基音检测

小波变换的一些性质如检测信号的突变已在图象处理中得到应用 [3] . 本文应用这条性质来进行语音的基音检测. 选择光滑函数 $^{[3]}\theta(x)$ 的一阶导数作小波 $\psi(x)=\mathrm{d}\theta(x)/\mathrm{d}x$，则信号 $f(x)$ 的小波变换式可记为

$$W_sf(x)=f(x)*\psi_s(x)=f(x)*\left[s\frac{\mathrm{d}\theta_s(x)}{\mathrm{d}x}\right]=s\frac{\mathrm{d}}{\mathrm{d}x}(f(x)*\theta_s(x)). \tag{1}$$

式中 $\psi_s(x)=(1/s)\psi(x/s)$， $\theta_s(x)=(1/s)\theta(x/s)$，尺度 s 一般取 $2^j,(1\le j\le J)$. $*$ 表示卷积。小波变换 $W_sf(x)$ 可表示成信号 $f(x)$ 在尺度 s 被 $\theta_s(x)$ 平滑后的一阶导数. 因而有这样一条结论 [3,4] ：如果选择小波函数为某一光滑函数的一阶导数，则由小波变换的幅度极大点可以检测信号的突变点，即小波变换可以用于检测信号的突变.

根据语音产生理论，发音过程中声门闭合瞬间声道被强烈激励，表现在波形上就是此瞬间幅度剧增产生突变 (图 1(a) 箭头处)，两相邻声门闭合瞬间的时长就是该处的基音周期. 由小

[1] 1995-04-05 收到， 1995-09-21 定稿
国家自然科学基金资助项目

波变换检测信号突变的性质可知，选择光滑函数 $\theta(x)$ 的一阶导数 $\psi(x)$ 作小波，通过确定语音信号 $f(x)$ 的小波变换 $W_sf(x)$ 的幅值局部极大点位置 (图 1(b) 或 1(c) 箭头处)，就可精确地检测到因声门闭合产生的语音波形的突变 (图 1(a) 箭头处)，相邻突变点间的时长就是所求的基音周期 [5,6] 。

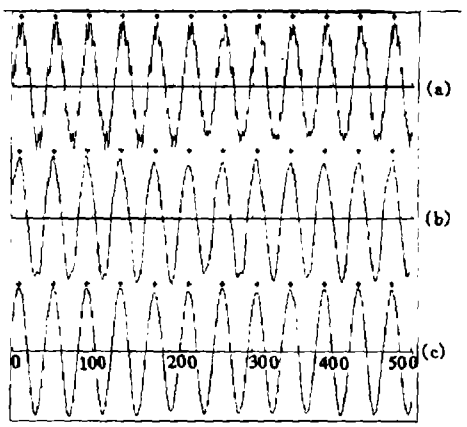

图 1 (a) 512 点语音信号，(b) $W_{2^3}f$，(c) $W_{2^4}f$

文中采用快速算法—— Mallat 算法求得各尺度上的小波变换。这里使用的 Mallat 算法与常用的 Mallat 算法略有不同。常用算法采用系数相同的 FIR 滤波器进行级联的结构，信号经每级滤波后均亚采样，且长度减少一半；而在以信号检测为目的的应用中，如果对信号进行亚采样，由于每级小波变换信号长度都成倍递减，就无法从变换后的信号中直接精确地找到与原始信号在位置上的对应关系。故而用于信号检测的 Mallat 算法应略作变化：取消亚采样，但仍采用滤波器级联的结构，只是每一级的滤波器系数不同，即第 j, $(2 \le j \le J)$ 级滤波器的系数是由第 1 级滤波器的非零系数间插 $2^{(j-1)}-1$ 个零得到的 [3,4] 。这种用于信号检测的 Mallat 算法与常用的 Mallat 算法原理上是一致的，仅仅是形式上不同而已。

图 2 给出了求 L 点 (一帧) 数字音段内各基音周期的流程图. 对 L 点的数字语音 d_n，先用上述的 Mallat 算法求得各尺度上的小波变换 $W_{2^j}f(x)$, $(1 \le j \le J)$. 根据文献 [4] 中的方法确定分析尺度的上下界. 本文中语音信号 10kHz 采样，则实验中尺度 s 下界取 2^3 、上界取 2^5，即图 2 中 $j_1 = 3$ 、 $j_u = 5$ 。在分析尺度范围内对 $W_{2^j}f(x)$ 进行如下处理：如果 $W_{2^j}f(x)$ 的最大值小于浊音门限 T ，则此语音段为清音，基音周期 P 置为 0 ；否则，可能为浊音段. 这时再用一削波门限 T_c 对 $W_{2^j}f(x)$ 进行削波，以便于检测到 $W_{2^j}f(x)$ 的各极大值点，两相邻极大点间的时长就是可能的基音周期值 P_j 。在相继的两个分析尺度下比较相应位置的基音周期 P_j 与 P_{j-1} ，如果相同，确认该基音周期 $P = P_j$ ；如果不同，则 $j = j+1$ ，在下一尺度重复上述工作. 直至上界 2^5 时，如两相继尺度下对应位置的基音周期仍不吻合，则设此处基音周期为 0 . 这一算法充分利用了小波理论多分辨率分析这一性质，在不同尺度 (不同分辨率) 下检测出基音并将结果进行比较，从而保证了检测出的基音周期准确可靠。这正是该方法优于传统方法的本质所在。

在实际中为实时实现这一检测算法，语音信号流应进行分帧处理，即用 L 点滑动窗对语音流进行截取后，用上述基音检测算法求出这 L 点中的各基音周期. 考虑到小波变换快速算法中

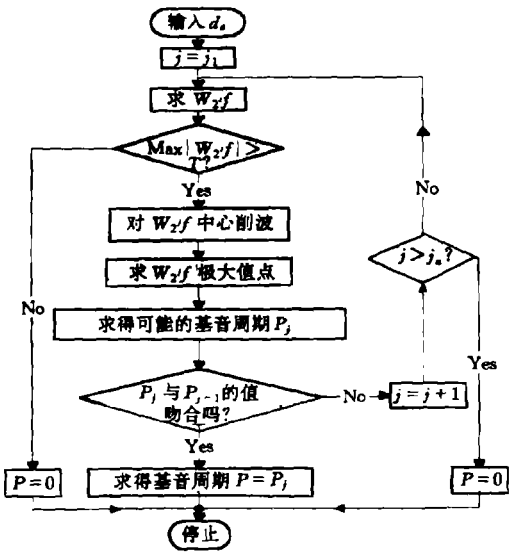

图 2　小波变换用于基音检测的流程图

因离散卷积产生的帧边界效应，窗的滑动应使两相继语音帧间有足够的叠接. 对一语音流检测后，可得基音周期序列 $\{P_t;\ t=1,2,\cdots,T\}$，相应基音频率序列 $\{f_t=1/P_t; t=1,2,\cdots,T\}$。

3　隐 Markov 模型和声调识别

由于基音包络是区分汉语四声的关键性参数，很自然我们定义以下特征矢量：

$$\boldsymbol{x}_t=[\log f_t+\log f_{t+1}, \log f_t-\log f_{t+1}]. \tag{2}$$

其中 f_t 表示 t 时刻基音频率. 矢量 $\boldsymbol{x}_t$ 的第一分量表示基频曲线在 t 时刻的局部幅度，第二分量则表示 t 时刻的局部斜率. 连续参数 HMM 时，一般大多采用 GM-pdf，即状态 j 时输出 pdf 为

$$b_{\mathrm{jGM}}(\boldsymbol{x}_t)=\sum_{m=1}^{M}c_{jm}b_{jm}(\boldsymbol{x}_t), \qquad \sum_{m=1}^{M}c_{jm}=1. \tag{3}$$

这里 $\boldsymbol{x}_t$ 是特征矢量，M 是总混合数，$b_{jm}(\boldsymbol{x}_t)$ 是状态 j 时第 m 个混合的 pdf，c_{jm} 是一非负实数。

本文采用 PGM-pdf 的 HMM[7,8]，即状态 j 时输出 pdf 为

$$b_j(\boldsymbol{x}_t)=\frac{1}{M}\max_{m=1,2,\cdots,M}\{b_{jm}(\boldsymbol{x}_t)\}. \tag{4}$$

文中选 $b_{jm}(\boldsymbol{x}_t)$ 为一个多维的 Gauss 分布 $N(\boldsymbol{x}_t,\boldsymbol{\mu}_{jm},\boldsymbol{C}_{jm})$，即

$$b_{jm}(\boldsymbol{x}_t)=N(\boldsymbol{x}_t,\boldsymbol{\mu}_{jm},\boldsymbol{C}_{jm})=\frac{1}{(2\pi)^{K/2}|\boldsymbol{C}_{jm}|^{1/2}}\exp\left\{-\frac{1}{2}[(\boldsymbol{x}_t-\boldsymbol{\mu}_{jm})^T\boldsymbol{C}_{jm}^{-1}(\boldsymbol{x}_t-\boldsymbol{\mu}_{jm})]\right\}. \tag{5}$$

这里 μ_{jm} 是均值矢量，C_{jm} 是协方差矩阵，上标 T 表示矩阵转置，K=2 为特征矢量的维数。对于 PGM 模型，特征空间可以被认为分划成几个聚类，每类由 Gauss-pdf 定义。(5) 式意味着任一矢量它属于特征空间中的哪一类，可由最邻近准则确定，这与矢量量化 [9] 是一致的.

识别用 Viterbi 算法。对于给定特征矢量序列 $X=(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_T)$，每个模型对此序列作最佳状态标注，求出该序列对此模型的似然值，达到最大似然值的模型对应的声调就是识别结果. 下面介绍 PGM 模型中 Viterbi 算法的一种实现方法。

对于 N 个状态的 HMM，矩阵 $\{a_{ij}\}_{n\times n}$ 为状态转移概率分布。设 $\delta_t(j)$ 为沿着单一路径，经前 t 个观察矢量，在 t 时刻为 j 状态时的最大概率的对数，则 Viterbi 算法中有如下递推式：

$$\delta_t(j)=\max_{1\le i\le N}[\delta_{t-1}(i)+\log a_{ij}]+\log b_j(\boldsymbol{x}_t),\quad 2\le t\le T,\quad 1\le j\le N. \tag{6}$$

假定矢量 $\boldsymbol{x}_t=\{x_{tk},k=1,2,\cdots,K\}$ 各维不相关，则 C_{jm} 为对角阵。对 PGM 模型，由 (5)、(6) 式得

$$\log b_j(\boldsymbol{x}_t)=\frac{1}{M}\max_{m=1,2,\cdots,M}\left\{\log\frac{1}{(2\pi)^{K/2}}+\log\prod_{k=1}^{K}\left[\frac{1}{\sigma_{jmk}}\right]-\frac{1}{2}\sum_{k=1}^{K}\left[\frac{x_{tk}-\mu_{jmk}}{\sigma_{jmk}}\right]^2\right\}. \tag{7}$$

上式中，大括号中第一项为常数，第二项可以事先求得，仅仅需要计算第三项. 如果不考虑前面的负号，第三项可以看成用方差倒数加权的均方距离. 将 (7) 式代入 (6) 式中，由于 $\log a_{ij}$ 也可事先求得，因此 PGM 模型的 Viterbi 算法中，不需任何对数运算. 这一简化计算的递推式，对于降低 HMM 计算代价有一定意义。

4 实验

采用基于小波变换的基音检测算法提取基音包络，选择紧支集二次样条作小波函数，它是一光滑函数的一阶导数。Mallat 算法中该小波相应低通、高通滤波器的有限脉冲响应为 [3] H=[0.125,0.375,0.375,0.125]、G=[-2,2]；PGM 模型的 Viterbi 算法作汉语单音节四声识别。声调识别实验分特定人和非特定人两类。HMM 选择自左至右单跳转结构。

实验 1 *一帧语音的小波变换与基音检测* 图 1 给出了 L 点数字语音信号及其在尺度 $s=2^3$、2^4 的小波变换 $W_{2^j}f,(3\le j\le 4)$ 的信号波形，箭头所示为声门闭合瞬间。从图 1 中可以看到，$W_{2^j}f,(3\le j\le 4)$ 的极大点位置与声门闭合瞬间吻合，特别是在尺度 2^3、2^4 时信号极大点的位置要比原始信号中容易判断得多。两相邻声门闭合瞬间的时长就是可能的基音周期。如果对相继尺度 (如：2^3 和 2^4) 上检测到的值进行比较，就可进一步确认所检测到的基音周期值。经检测该帧语音共有 11 个完整的基音周期段，基音周期约为 46 样点，即 4.6ms，相应基音频率为 217Hz。由这一实验可以看到：这种基于小波变换的基音检测算法能精确地定位每个基音周期的起止点，并能通过相邻尺度进行确认，这使得基音检测准确、可靠。该算法不受基音动态范围限制，分析帧长的选择与基音周期长短无关。

实验 2 *单音节基音检测* 单音节语音流应分帧处理，随着窗的滑动在各滑动窗内用基音检测算法求得各基音周期。图 3 是“胀 /Zhang(4)”音节的基音频率曲线。

实验 3 *特定人四声识别实验* 采用国家科委智能计算中心提供的特定人汉语孤立音节识别数据库。采样频率为 16kHz，精度 16bit，截止频率约 7kHz. 训练和测试限于一男性语音样本共两遍，每遍包括全汉语音节计 1264 个. 实验中除去了一些非常用音节，只采用了 1258 个音节。两遍数据，一遍作训练，另一遍作识别. 由于要与实时语音识别系统兼容，实验中全部数

据均经 16kHz 降速率处理至 10kHz．表 1 中给出了特定人四声识别结果．其中 N 表示 HMM 的状态数，M 是混合数．

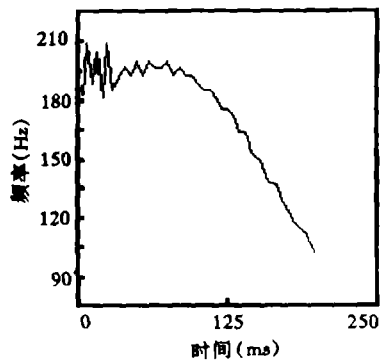

图 3　“胀 /Zhang(4)”音节的基音频率曲线

表 1　特定人四声识别实验统计表

参数	四声声调	1	2	3	4	合计
	识别音节数	336	255	319	348	1258
N=5	正确数	321	254	311	337	1223
M=1	正确率 (%)	95.54	99.61	97.49	96.84	97.22
N=5	正确数	320	251	307	341	1219
M=2	正确率 (%)	95.24	98.43	96.24	97.99	96.90
N=8	正确数	330	124	314	258	1026
M=1	正确率 (%)	98.21	48.63	98.43	74.14	81.56

从表 1 中统计数据可以看到，对于特定人四声识别，其识别率与 HMM 的参数密切相关．当 N=5 、M=1 时，识别性能最佳，达到 97.22%．对于相同的状态数 (N=5)，混合数 M=1 时较 M=2 时略好．其原因可能在于：对于汉语单音基频的特征参数分布的建模，用单 Gauss 模型 (M=1) 要比混合 Gauss 模型 (M=2) 略好；对于相同混合数 (M=1)，状态数 N=5 要比 N=8 好得多．这是因为汉语单音基频曲线结构相对简单，用 8 个状态的这一较复杂的模型来描述它，必然导致四声识别率下降．

实验 4　*非特定人四声识别实验*　由实时数据采集系统采集实验数据，采样频率 10kHz，精度 14bit，截止频率约 4.5kHz．训练和识别用音节均自汉语全音节 (1258) 中随机产生．HMM 中状态数 N=5，混合数 M=1．

训练集：男 6 人，共 480 个音节．女 6 人，共 480 个音节．

测试集：男 3 人，共 240 个音节．识别正确数：226；识别正确率：94.17%．

5　结　束　语

本文给出了一个运用小波变换和隐 Markov 模型技术识别汉语四声的方法。由语音小波变换相邻极大值间的时长来获得基音周期，并在下一尺度进行确认，从而保证了基频检测的准确性；文中首次尝试将 PGM-pdf 的 HMM 应用于汉语四声识别，并根据 PGM 函数式的特点，在识别过程的 Viterbi 算法中推导出一个简化的计算匹配打分的递推公式．该方法与传统的 GM-pdf 的 HMM 相比，计算量大大减少，特定人四声识别率为 97.22%，非特定人达到 94.47%．这一递推式的导出，对于解决 HMM 在语音识别中的计算代价问题具有参考价值。

参　考　文　献

[1]　拉宾纳 L R, 谢弗 R W. 语音信号数字处理. 北京：科学出版社，1983, 第 7.3 节

[2]　Yang W, Lee J, Chang Y, Wang H. Hidden Markov Model for mandarin lexical tone recognition. IEEE Trans. on ASSP, 1988, ASSP-36(7): 988-992.

[3]　Mallat S, Zhong S. Characterization of signals from multiscale edges. IEEE Trans. on PAMI, 1992, PAMI-14(7): 710-732.

[4]　程　俊，张　璞，戴善荣，易克初. 小波变换用于信号突变的检测. 通信学报，1995, 16(3): 96-104.

[5] Kadambe S, Boudreaux-Bartels G F. Application of the wavelet transform for pitch detection of speech signals. IEEE Trans. on IT, 1992, IT-38(2): 917–924.

[6] Cheng J, Zhang P, Dai S R, Hu Z. An event based pitch detector using fast wavelet transform. In ed. Yuan B Z. Proceeding of International Conference on Signal Processing(vol.1). Beijing: International Academic Publishers, 1993, 683–686.

[7] Juang B H, Rabiner L R. Mixture autoregressive hidden Markov models for speech signals. IEEE Trans. on ASSP, 1985, ASSP-33(6): 1404–1413.

[8] Lee Y, Lee L. Continuous hidden Markov models integrating transitional and instantaneous features for Mandarin syllable recognition. Computer Speech and Language, 1993, 7: 247–263.

[9] 胡 征，杨有为. 矢量量化原理与应用. 西安：西安电子科技大学出版社， 1988, 第二章.

A TONE RECOGNIZER USING WAVELET TRANSFORM AND HIDDEN MARKOV MODEL

Cheng Jun Yi kechu Li Bingbing

(*National Key Lab. on ISN, Xidian University, Xi'an* 710071)

Abstract This paper presents a tone recognizer for Mandarin speech using a combination of wavelet transforming and hidden Markov modeling techniques. The evaluation of pitch periods is exactly performed by a pitch detector which is based on the singularity detection of signal and multiresolution analysis with wavelet transform. The hidden Markov models with partitioned Gaussian mixtures(PGM) are used for tone recognition. In implementing the Viterbi algorithm for HMM's, a recursive relation is derived to improve the computation effieiency, where the accuracy is 97.22%, 94.47% for speaker-dependent and speaker-independent tone recognition respectively.

Key words Pitch detection, Tone recognition, Wavelet transform, Hidden Markov model

程 俊： 男， 1964 年生，硕士，副教授，现从事语音识别、合成，信号的时域表示和小波理论等方面的研究工作.

易克初： 男， 1943 年生，博士，教授，博士生导师，现从事语音处理与卫星通信等方面的教学与研究工作.

李兵兵： 男， 1955 年生，博士，副教授，现从事图象编码，小波理论等方面的研究工作.