第 24 卷第 1 期　　声 学 学 报　　Vol. 24, No. 1
1999 年 1 月　　ACTA ACUSTICA　　Jan., 1999

基于小波变换的重叠语音基频提取及声调识别 *

赵鹤鸣　周旭东　金延庆　翁桂荣

(苏州大学电子工程系　江苏苏州　215006)

1997 年 7 月 14 日收到
1997 年 11 月 25 日定稿

摘要　提出一种基于小波变换的重叠语音基频提取及声调识别的方法。利用小波的伸缩和时移特性，通过对重叠语音做多尺度的小波变换，可以有效地提取重叠语音中各自的基音频率，并在此基础上实现声调的识别。实验表明，此方法是有效的，是重叠语音基频提取及声调识别的一种新途径。

PACS 数：　43.60, 43.70

Overlapping speech pitch extraction and tone recognition based on wavelet transform

ZHAO Heming　ZHOU Xudong　JIN Yanqing　WENG Guirong

(*Department of Electronic Engineering, Suzhou University* Suzhou　215006)

Received Jul. 14, 1997
Revised Nov. 25, 1997

Abstract　Based on wavelet transform, this paper submits a new approach to overlapping speech pitch extraction and tone recognition. Taking advantage of retractility and time-delay characteristics of wavelet, it can detect each pitch in overlapping speech effectively and implement tone recognition by multiscale wavelet transform on overlapping speech. Results of experiment showed the approach is efficient.

引言

汉语是声调语言，声调识别对汉语语音信号处理具有重要意义。由于声调与基音频率密切相关，因而一旦求得基音频率随时间变换的轨迹，就可用多种方法识别声调[1-2]，目前，对语音进行基频提取已有较为成熟的方法[3]。但是，在重叠语音 (例如两个不同话者语音的混叠) 情况下，要从中分别检测各自的基音频率就非常困难[4]，重叠语音基频提取及声调识别不仅是研究混叠语音分离的基础，而且对于语音识别技术走向实用具有实际意义。迄今为止，重叠语音基频提

* 江苏省自然科学基金资助项目

取的方法主要有两类[5-6]：一类基于信号处理方法 (例如 SHS 方法等)，另一类则建立在听觉感知特性的基础上。这两类方法目前均有局限性，前者检测基频的准确性依赖于两重叠语音的基频分布，当构成重叠语音的基频接近或成整数倍时，该类方法几乎接近失效；后一类则往往需要知道重叠语音基频的一些先验知识，因而应用受到限制。

我们知道，人耳很容易分辨重叠语音中的不同声音特征及语意，对于人耳处理复杂声信号的能力及内在机理已有不少学者对此做了深入研究，并已证明人耳蜗滤波器本质上是一个小波变换[7]，因此，运用小波变换分析重叠语音并进行特征提取是一个有价值的研究课题。小波变换是近年来发展起来的一种新的时频多分辨率分析方法，它将一维时间信号映射到二维时间尺度平面，其中尺度这个参数和频率相关，但又和频率不同，它从不同的角度刻划了信号特征，调整尺度因子，可使信号的低频成分有较高的频域分辨率，而对信号中的突变 (即高频部分) 能较好地定位，即有较高的时域分辨率。本文首先讨论小波变换检测重叠语音基频的原理及实现方法，并将检测结果用于声调识别，取得了较好的结果。

1 小波变换及基频检测原理

1.1 小波变换

设 $\psi(t) \in L^2(R)$，其傅里叶变换为 $\Psi(\omega)$，若 $\Psi(\omega)$ 满足：

$$\int_{-\infty}^{\infty} \frac{|\Psi(\omega)|^2}{|\omega|} \mathrm{d}\omega < \infty \tag{1}$$

则称 $\psi(t)$ 为一基本小波或母小波。将母小波函数 $\psi(t)$ 经伸缩和平移后得：

$$\psi_{m,n}(t) = 2^{m/2}\psi(2^m t - n) \qquad m, n \in Z \tag{2}$$

满足：

$$\int_{-\infty}^{\infty} \psi_{m,n}(t)\overline{\psi_{m',n'}(t)}\,\mathrm{d}t = \delta_{mm',nn'} = \begin{cases} 1 & m = m', n = n' \\ 0 & m \neq m', n \neq n' \end{cases} \tag{3}$$

构成 $L^2(R)$ 的一个正交完备集。此时我们称 $\psi_{m,n}(t)$ 为二进正交小波基，且对任意 $f \in L^2(R)$，其小波变换定义为：

$$Wf(m,n) = \langle f, \psi_{m,n} \rangle = \int_{-\infty}^{\infty} f(t)\psi(2^m t - n)\mathrm{d}t \tag{4}$$

小波变换系数 $Wf(m,n)$ 给出了 $f(t)$ 的尺度 2^m 位置 n 处的逼近。

小波变换一般由符合条件的有限长脉冲响应滤波器 (FIR) 实现，其实现算法为[8]：

$$W_{2^{m+1}} = \sum_k g(k-2n)S_{2^m}f \tag{5}$$

$$S_{2^{m+1}} = \sum_k h(k-2n)S_{2^m}f \tag{6}$$

图 1 表示了离散正交小波变换的过程。

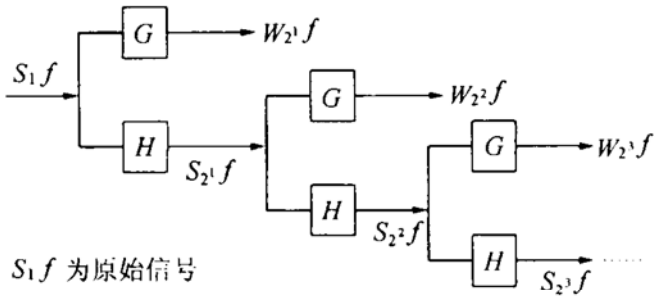

图 1　小波变换快速算法流图

该算法只需根据半带滤波器 H 、 G 即可实现信号的多尺度分解。本文采用具有紧支集且有线性相位的正交镜像 FIR 滤波器，其传递函数为[9]：

$$H(\omega)=e^{\mathrm{j}\omega/2}\left[\cos\left(\frac{\omega}{2}\right)\right]^3 \tag{7}$$

$$G(\omega)=4\,\mathrm{j}\,e^{\mathrm{j}\omega/2}\sin\left(\frac{\omega}{2}\right) \tag{8}$$

1.2 基音频率检测原理

基音频率 (与基音周期互为倒数) 即发浊音时声带振动的频率，人在发音过程中，由于声门瞬时闭合，声道被强烈激励，表现在语音波形上就是此瞬间幅度剧增，产生突变。相邻两个声门闭合之间的时间长度的倒数就是该处的基音频率。因此，只要能检测到声门闭合产生的语音突变就可以求出基频。小波变换是检测信号突变的有效工具。只要定位语音信号经小波变换 $W_{2^m}f$ 的幅度极大点位置，就可精确定位语音波形因声门闭合产生的突变点，求此相邻突变点间时长倒数就可得到基音频率[10]。

2　重叠语音基频提取

小波变换是一种多分辨率分析，它可在“放大”了的不同频带内分析信号，使本来不易察觉的信号特征在不同分辨率的若干子空间中显露出来，因此它可以用于重叠语音各自基频的提取。当两个语音混叠时，由于话者的基频一般总有差异，所以，对其用不同尺度 m 进行小波变换时，$W_{2^m}f$ 的极大点出现在不同时刻点上，它对应于不同话者声门闭合产生的语音波形突变点，这样，就可根据不同尺度的 $W_{2^m}f$ 检测不同话者的基频。

由于基音频率有一限定范围，因而只需对语音信号作有限尺度 ($m\le M$) 小波变换即可用于检测。一般来说，尺度 2^m 越大， $W_{2^m}f$ 中反映基频信息的极大点越明显，越易处理，但当语音信号采样频率 F_s 为 11 kHz 左右时，取 M 为 5 或 6 已能满足绝大多数基频检测要求，这是因为当分析尺度为 2^M 时， $W_{2^M}f$ 占据的频带为 $F_s/2^{M+1}\sim F_s/2^M$ ，音调较低话者的基频大多数落在该频带内。因此，可根据最大分析尺度 2^M 对应的小波变换 $W_{2^M}f$ 波形，通过求相邻极大点间的时长就可确定重叠语音中音调较低者的基频，见图 2(c)，重叠语音中另一话者的基频可通过 $W_{2^j}f(j=M-2$ 或 $j=M-3)$ 的极大点间隔计算确定。虽然构成重叠语音的基频分布是随机的，即另一话者的基频不一定正好落在 $W_{2^j}f$ 占据的频带内，但 $W_{2^j}f$ 波形中必定包含了重叠语音中另一话者声门闭合产生语音波形突变点的信息 (见图 2(b))，这是小波分析的多分辨率特性所决定的。

从重叠语音中提取各自的基频信息还需解决以下两个问题：一是由 $W_{2^j}f$ 确定极大点位置；二是要确定构成重叠语音的各自浊音段的起终点。对于第一个问题，由于 $W_{2^j}f$ 不象 $W_{2^M}f$ 波形

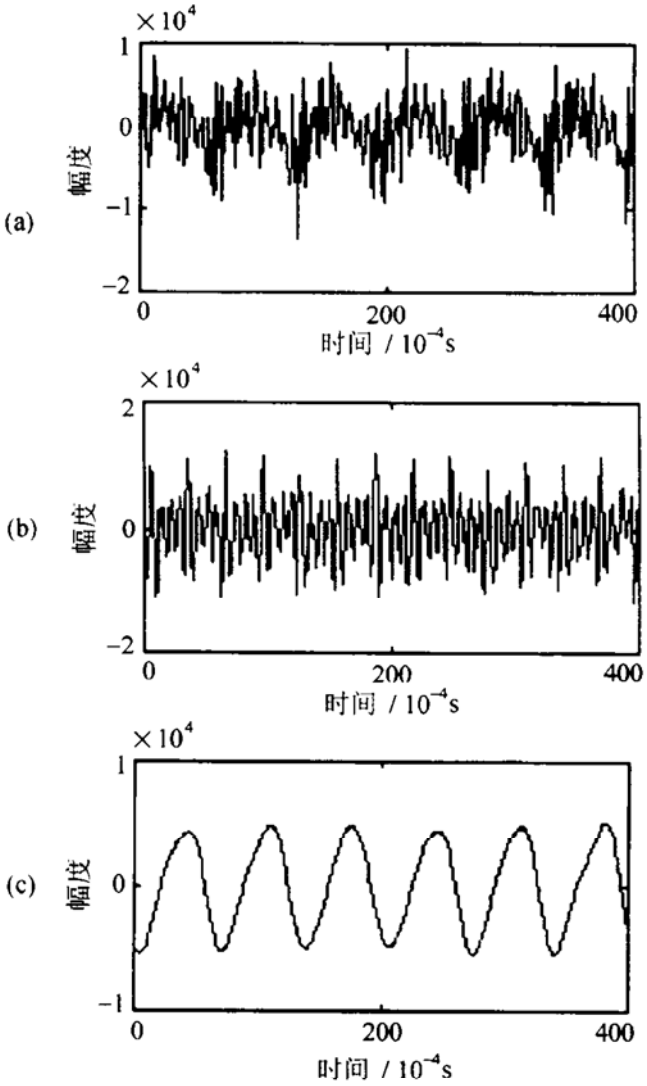

图 2 (a) 原始重叠语音 (b) $W_{2^j}f(j=3)$ (c) $W_{2^M}f(M=6)$

(对照图 2(b) 和 2(c)) 那样容易定位其极大点位置，因而我们根据声门瞬时闭合产生语音波形的幅度剧变的发音机理，由尺度为 2^j 小波变换波形的幅值及“局部变化量”来确定 $W_{2^j}f$ 的极大点位置，定义 t 时刻波形的局部变化量为：

$$D(t)=\sum_i A_{\max}^{(i)}(t)-A_{\min}^{(i)}(t) \tag{9}$$

式中 $A_{\max}^{(i)}(t)$ 与 $A_{\min}^{(i)}(t)$ 分别为 t 时刻在邻域 Δt 内 $W_{2^j}f$ 的第 i 个峰值与谷值。实验表明，合理选择 $W_{2^j}f$ 的幅度阈值和 $D(t)$ 阈值就能简单有效地定位 $W_{2^j}f$ 的极大点。

基频提取的第二个问题是指重叠语音中存在两个语音各自的噪声段 N 、声母段 C 和韵母段 V 的多种可能的组合，例如： N_1N_2 、 N_1C_2 、 N_1V_2 、 N_2C_1 、 N_2V_1 、 C_1C_2 、 C_1V_2 、 C_2V_1 、 V_1V_2 、 V_1N_2 、 V_2N_1 等等，如前所述，在两个韵母重叠区间，可根据 $W_{2^j}f$ 和 $W_{2^M}f$ 来提取两个不同的基频，但在 C_1V_2 、 C_2V_1 、 N_2V_1 、 N_1V_2 、 V_1N_2 、 V_2N_1 等可能段中，只有构成重叠语音的一个话音存在基频，但 $W_{2^j}f$ 与 $W_{2^M}f$ 中均含有对应的极大点。为此，根据由 $W_{2^j}f$ 和 $W_{2^M}f$ 所有极大点定位得到的候选基频序列 $\{P_{1k}\}$ 、 $\{P_{2k}\}(k$ 为离散时间点，当噪声 N 与声母 C 时， $P_k=0)$ 作如下处理：在对应相同时间点，当 P_{1k} 与 P_{2k} 相差较大时 (对应 V_1V_2 段)，则保留各自的 P_k 的值，当 P_{1k} 与 P_{2k} 连续一段相等或很接近 (计算误差所致) 时，则判断各自基频序列的连续性，由此保留 P_{1k}(或 P_{2k}) 且令 P_{2k}(或 P_{1k}) 为零，由于这种情况均发生在重叠语音的始、终端，因而可分区间判断，其算法框图见图 3 。

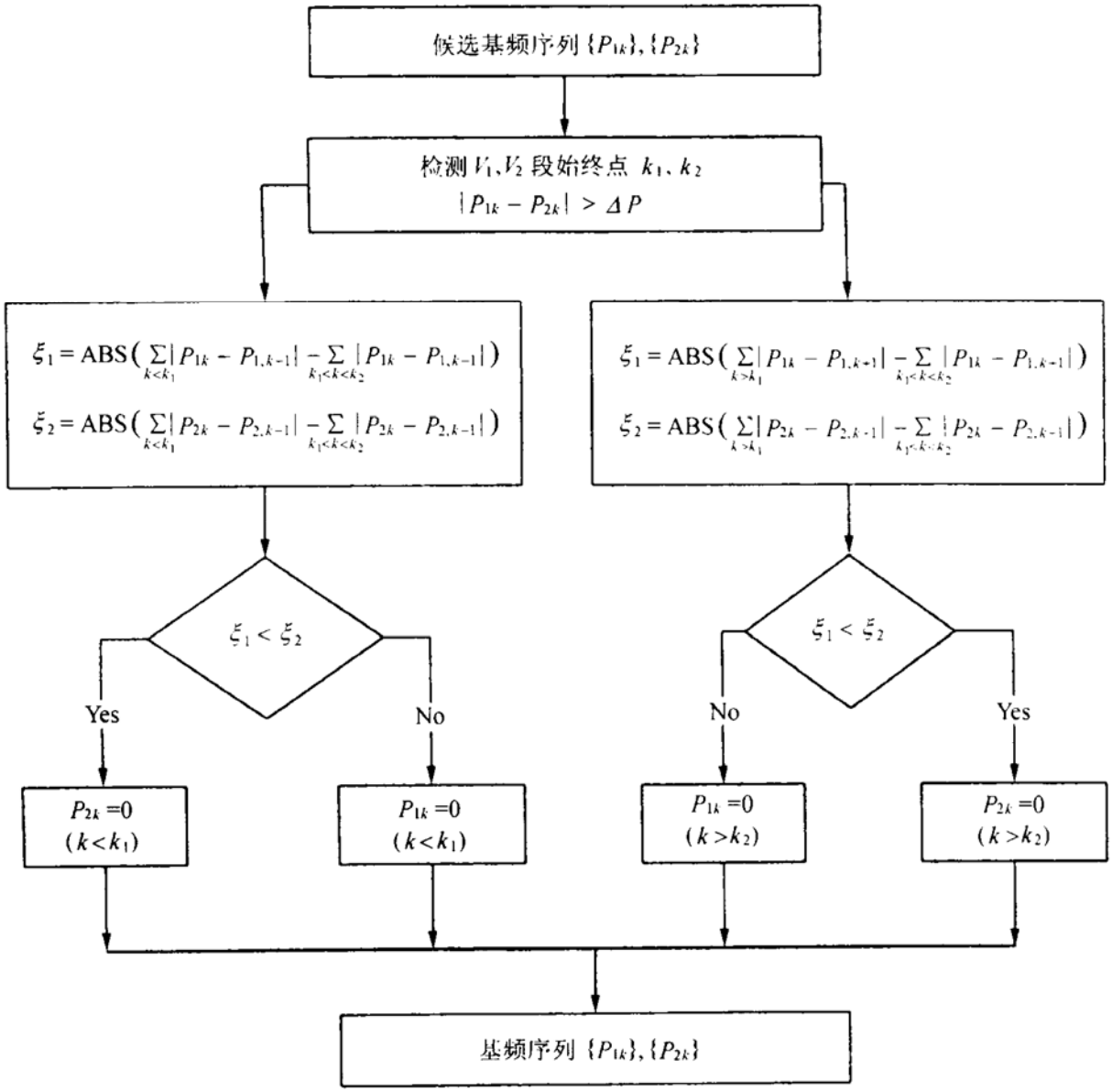

图 3　基频序列选取算法框图

3　声调识别

汉语声调信息载于韵母段的基频曲线上，为检验本文所述方法提取重叠语音基频的有效性，将检测结果用于四声的识别。一般来说汉语第一声调的基频轨迹很平坦，而其它三种声调则具有不同程度的弯曲和倾斜：第二声调上倾，第三声调是凹的，而第四声调明显下倾。因而可采用模式识别方法进行四声识别。训练样本为构成重叠语音的各单独发音，对每一类声调，根据基频轨迹计算曲线的斜率、谷点和平坦度等参数并形成声调模式，然后从重叠语音提取得到的基频曲线计算相应的轨迹描述参数，再计算出它与各声调类别的距离，具有最小距离的类别即为识别结果。有关各参数的定义及具体计算方法、样本训练方法可参见文献 [1]，此处不再赘述。

4　实验结果及讨论

根据上述方法我们对多种重叠语音样本 (由成年男女及儿童的孤立发音两两混叠而成) 进行了基频检测和声调识别，并与单独发音得到的结果相比较。图 4(a) ~ 图 4(d) 分别给出了一男声

单独发音“苏”和一女声单独发音“大”的语音波形及基频轨迹(检测方法也用小波变换法)，图 4(e) 为上面两语音混叠得到的重叠语音波形，图 4(f) 和图 4(g) 则为根据重叠语音波形经小波变换多尺度分析后提取得到的各自基频曲线。为对比实验结果，图 4 给出的所有基频曲线均未进行平滑处理，由图可见，从重叠语音提取的各自基频曲线与单独发音得到的曲线非常接近。另外，两组曲线若经相同算法平滑处理后，完全能从基频轨迹反映各自的音调模式。

为进一步检验本文提出方法的性能，对五十个重叠语音样本(含各种声调)进行了声调识别实验，识别结果正确率为 84 %，不能正确识别的样本多为重叠语音由两个第三声或一个第三声和一个第二声组成。这是因为构成重叠语音的两个信号若在某一时段频率非常接近(两个第三声的声调语音容易出现这种情况)时将受小波变换时频分辨率的限制而影响基频检测结果，况且本文采用的是二进离散小波变换，其时频分辨率性能一般，对时频特性的定位有一定影响。有关细化时频分辨率的小波变换方法有待进一步研究。

作为对比，我们将上述重叠语音样本用 SHS 算法也进行了基频分离和检测，并将所得结果用于声调识别。在基频曲线平滑算法与声调识别算法相同的前提下，用 SHS 算法所得结果进行声调识别的正确率为 72 %。实验结果对比表明：对于本文算法基频分离、检测效果较差的样本，应用 SHS 算法时效果并未改善；混叠语音样本中两基频接近倍数关系时，本文算法能将各自基频有效分离并检测，而 SHS 算法则不能；当两基频相差不多时，本文算法的分离检测效果好于 SHS 算法。另外，应用 SHS 算法时，因检测错误导致基频轨迹含较多零乱点，虽经平滑处理，但对声调识别仍有影响。在计算复杂度方面，本文算法由于采用小波变换快速算法，其乘法与加法次数为 $(5 \sim 6)(F_H + F_G)N$ 量级，其中 N 为每帧语音样点数，F_H、F_G 分别为 H、G 滤波器的非零点数，算法中无对数运算。而 SHS 算法首先经 FFT 计算频谱(乘、加法次数为 $N_1 \log_2 N_1$ 量级，且 $N_1 > N$，这是因为作 FFT 时需添零)，并进行对数压缩(N_1 次对数运算)，然后再对每倍频程进行 J 点(根据分辨率要求，J 一般取 45 ~ 60，本文取 48)三次样条插值(计算量为 $(7 \sim 8)J$ 点三次样条插值运算)。所以，SHS 算法的计算量和运算复杂度均高于本文方法。

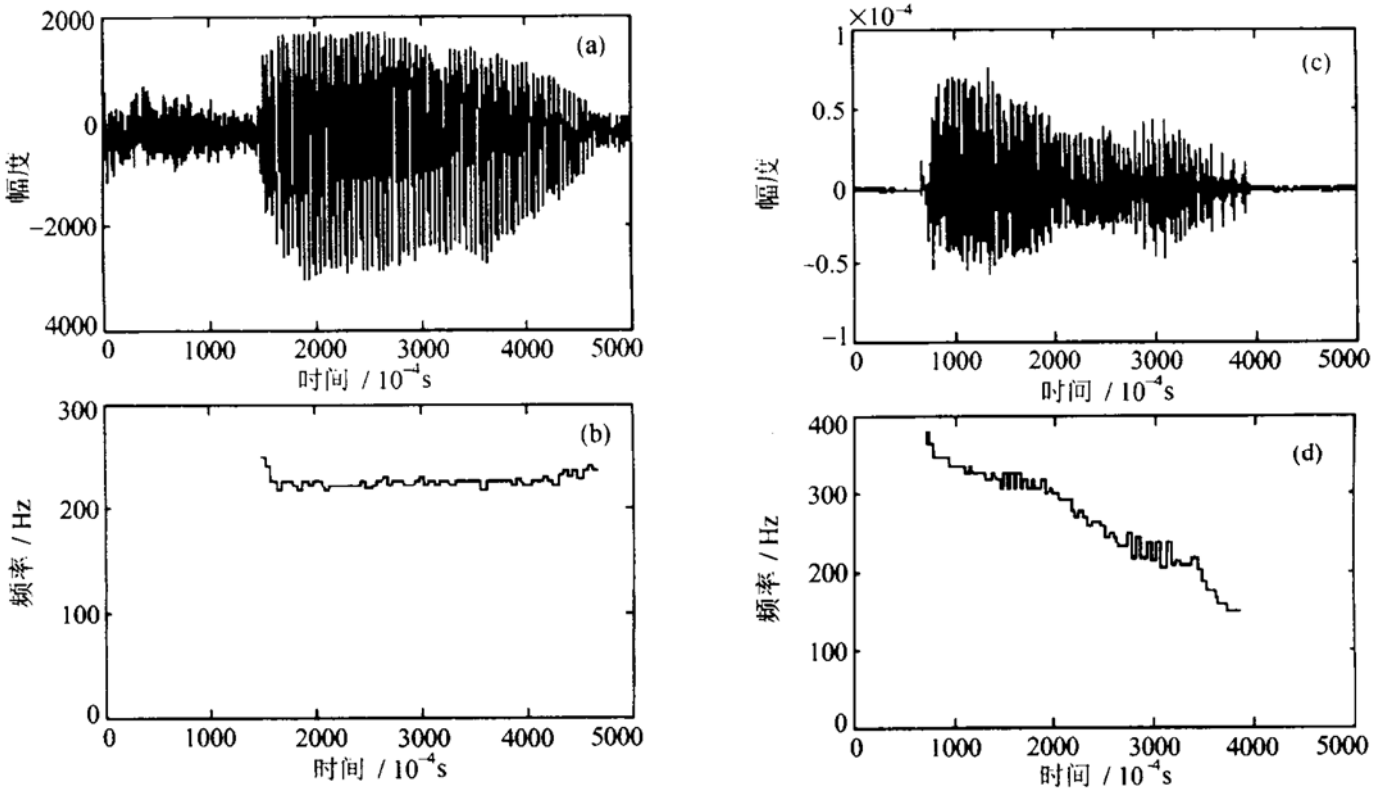

图 4 (a) 单个语音信号 S_1“苏”的原始波形 (b) 对应 (a) 的基频曲线
(c) 单个语音信号 S_2“大”的原始波形 (d) 对应 (c) 的基频曲线

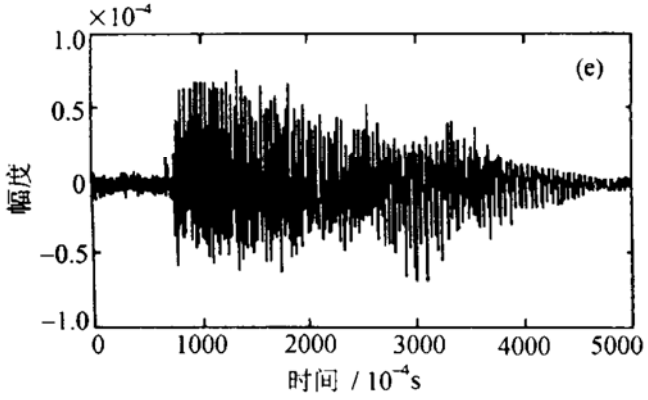

图 4(e) S_1 和 S_2 混叠语音信号

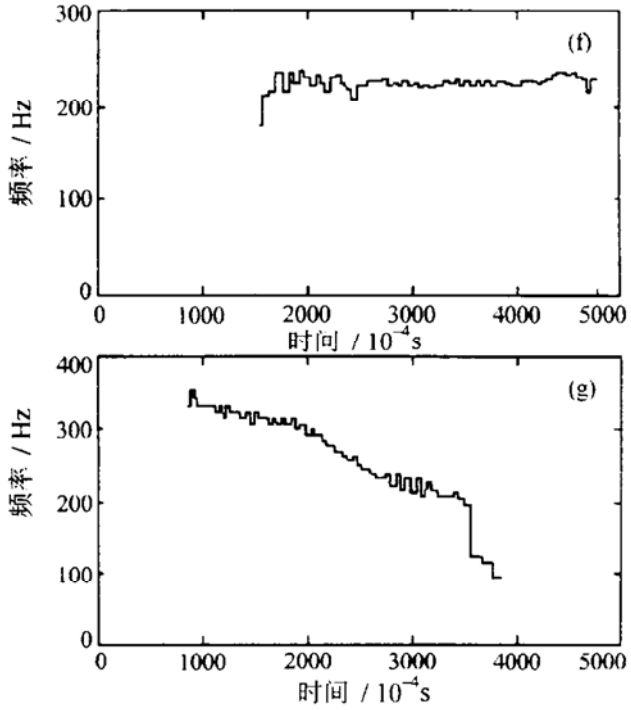

图 4(f) 从混叠语音信号中检测对应 S_1 的基频曲线

(g) 从混叠语音信号中检测对应 S_2 的基频曲线

5　结论

本文首次利用小波变换方法有效实现了重叠语音各自基音频率的提取，与重叠语音基频提取的其它方法比较，它具有以下特点：(1) 不需对语音信号作平稳假定，因而分析窗长任选。(2) 能精确定位基音周期段的开始与结束因而准确估计基频，而已有方法往往估计分析窗内的平均值。(3) 计算复杂度为 $O(N)$，比其它方法低。(4) 构成重叠语音的两信号基频相差越大，检测效果越好，但当两基频相差不太大时仍能取得较好结果。本文提出的方法主要适合于由孤立单字发音构成的重叠语音，有关连续语音混叠相应的基频提取方法正在进一步研究中。

参　考　文　献

1　黄泽镇，杨行峻．普通话孤立字四声的一种模式识别方法．声学学报，1990; **15**(1): 36—43

2　Ying Y, Xu S. A fast method of pitch detection for Chinese four tones recognition. Pro. of ICSP'93, Oct. 1993, Beijing

3　Hess W. Pitch determination of speech signals. Springer–Verlag, 1983

4　Stubbs R J, Summerfield O. Algorithms for separating the speech of interfering talkers. *J. Acoust. Soc. Am.*, 1990; **87**(1): 359—372

5　Luo H Y, Separation of overlapping speech. Ph D Thesis. Sussex University, 1994

6　Denbigh P N *et al.* Pitch estimation for two overlapping voices. *Proc. of Int. Conf. on Speech & Hearing.*, 1996, Windermere, 515—521

7　Yang X W *et al.* Auditory representations of acoustic signals. *IEEE Trans. on IT*, 1992; **38**(2): 824—839

8　Mallat S G. Atheory for multiresolution signal decomposition: wavelet representation. *IEEE Trans. on PAMI*, 1989; **11**(7): 674—692

9　程　俊等．小波变换用于信号突变的检测．通信学报，1995; **16**(3): 96—104

10　Kadambe S *et al.* Application of the wavelet transform for pitch detection of speech signals. *IEEE Trans. on IT*, 1992; **38**(2): 917—924