基于语音谐波结构的鲁棒特征参数及其在说话人识别中的应用

张玲华 郑宝玉 杨 震
(南京邮电大学通信与信息工程学院 南京 210003)

摘 要 通过对纯净语音及含噪语音短时谱的分析比较，提出了一种基于基音频率及其谐波结构的新的语音特征参数。实验表明，与传统的倒谱特征相比，新特征对加性白噪声相对较不敏感，在闭集文本无关说话人识别中，新特征可以在加性白高斯噪声环境下提高系统的说话人识别率。

关键词 说话人识别，短时谱，谐波特征，基音频率
中图分类号：TN912.3 **文献标识码**：A **文章编号**：1009-5896(2006)10-1786-04

Robust Feature Based on Speech Harmonic Structure for Speaker Identification

Zhang Ling-hua Zheng Bao-yu Yang Zhen
(*College of Communication and Information Engineering, Nanjing Univ. of Posts & Telecomms., Nanjing* 210003, *China*)

Abstract An effective and robust speech feature extraction method based on pitch frequency and harmonic structure is proposed by means of short-time spectrum analysis of clear and noisy speech. Experimental results indicate that the new feature is relatively insensitive to Additive White Gaussian Noise (AWGN). Compared to conventional cepstrums, the new feature can give outstanding improvement for closed-set text-independent speaker identification under noisy environments corrupted by AWGN.

Key words Speaker identification, Short-time spectrum, Harmonic feature, Pitch frequency

1 引言

说话人识别系统通常对在特定环境中采集的大量语音数据进行训练，识别时系统需要同样的环境以保证好的性能。实用中，由于人声道特征、发音方式随时间变动，特别是电话和移动通信环境下，话带以外说话人信息的丢失，包括话机在内的传输线路特性的变化，来自不同干线的话音质量存在差异以及通话环境的噪声等，都严重影响说话人识别系统的性能，使得很多在实验室里性能很好的说话人识别系统，在实用环境下的识别性能却显著降低[1]。可以说，如何提高训练条件与测试条件不匹配情况下系统的性能，是说话人识别领域的研究热点和难点。

在说话人识别中，说话人样本模型的训练和对说话人身份的识别都是基于所选取的语音特征参数进行的。说话人识别系统中所选择的语音特征参数应尽量突出说话人的个性特征。该特征对不同说话人应有较大变化，而对同一说话人则变化很小，使不同说话人能在特征空间尽量分离，便于区分[2]。

另一方面，为了在特征域提高系统的鲁棒性，特征参数除了必须具备上述特点之外，还必须具有一定的抗噪声能力。

在鲁棒特征参数提取方面，人们已经做了大量的工作。例如，文献[3]基于小波变换的时频多分辨率分析提出了一种有效的鲁棒特征参数提取方法，取得了较好的效果。近年来许多文献[4–6]提出，自动说话人识别系统能够使用高层次声学信息，以提高系统的准确度、增强鲁棒性。但这些方法的分析和运算都较为复杂。

当前的说话人识别领域依然被使用短时、低层次声学信息(如倒谱特征)的系统所统治。得到广泛应用的线性预测倒谱系数(LPCC)和 Mel 频率倒谱系数(MFCC)在无噪声环境下性能很好，但在实用环境下，由于噪声的影响，系统的性能却会严重下降。虽然在 SNR 变差时可以采用 MFCC+ΔMFCC，甚至加上 ΔΔMFCC 来改进系统的性能，但特征矢量维数的增加将会使系统的运算量明显增加。

本文通过对语音信号短时谱的分析，提出了一种新的基于基音频率及其谐波结构的语音特征参数，该方法具有概念明确、运算简单的优点。基于 GMM 的说话人识别实验表明，新特征对加性白噪声较不敏感，在闭集文本无关说话人识别中，新特征具有较强的抗噪声性，可以提高系统在噪声环境下的说话人识别率。

2 语音信号的短时谱分析

说话人发音器官的先天差异主要表现在语音的频率结构上，语音的短时谱中包含有激励源和声道的特性，可以反映说话人的生理差别。这说明语音信号的短时谱能展示说话人的个性特征，可以作为说话人识别应用的特征参数。

对于人的听觉来说，浊音(Voice)是最重要的语音信号。由于来自背景的随机噪声一般可以假定为加性白噪声(Additive White Noise, AWN)[7]，下面我们分别对纯净浊音及

2005-02-21 收到, 2005-10-31 改回
江苏省“青蓝工程”跨世纪学术带头人专项基金（QL003YZ）和南京邮电大学科研发展基金（2001 院 17）资助课题

含加性白高斯噪声(Additive White Gaussian Noise, AWGN)的浊音进行短时傅里叶变换，并分析它们短时谱的异同。

图 1 和图 2 所示分别是男声和女声浊音帧的短时谱分析，语音的采样率是 11.025kHz，窗长为 30ms(相应的样点数为 330)，窗函数为 Hamming 窗。其中图 1(a) 是无噪声时的时域波形；图 1(b) 是无噪声时的短时谱(对数谱)；图 1(c) 是混有加性白高斯噪声(AWGN)，在信噪比(SNR)为 10dB 时的时域波形；图 1(d) 是在信噪比(SNR)为 10dB 时的短时谱(对数谱)。

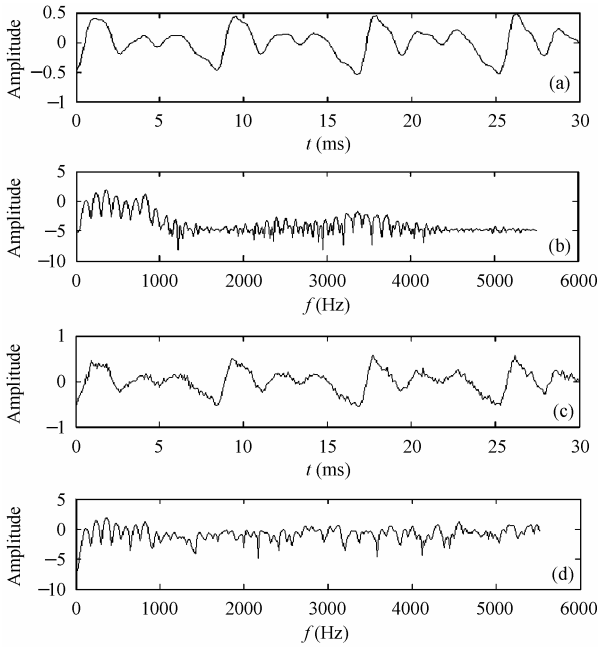

图 1　男声浊音帧的时域波形和短时谱　(a) 无噪声时的时域波形　(b) 无噪声时的对数谱　(c) SNR=10dB 时的时域波形　(d) SNR=10dB 时的对数谱

Fig.1 Voiced sound waveform and short-time spectrum from a male speaker　(a) Clean sound waveform　(b) Clean sound log-spectrum　(c) Noisy sound waveform with SNR=10dB　(d) Noisy sound log-spectrum with SNR=10dB

从图 1 中可以看出，图 1(a)与图 1(c)之间有较大差异，这主要表现在图 1(c)有较明显的毛刺，但图 1(b)与图 1(d)在低频段反映出来的基频及其谐波特性(包括谱峰的位置和幅度)却基本上保持不变。表 1 所示的数据是对图 1 所示的男声和图 2 所示的女声在无噪声和 SNR=10dB 两种情况下测出的低次谐波谱峰的幅度。

表 1　浊音帧短时谱低次谐波的幅度

Tab.1 The amplitudes of lower harmonics of voiced sound

低次谐波的幅度	男声 无噪声	男声 SNR=10dB	女声 无噪声	女声 SNR=10dB
一次谐波	1.0631	1.0660	0.3729	0.3790
二次谐波	2.4756	2.4829	0.7432	0.7565
三次谐波	2.8968	3.0484	1.3688	1.3280
四次谐波	1.0064	1.1418	1.6409	1.6505
五次谐波	0.3034	0.3527	0.3979	0.3949
六次谐波	0.2082	0.2473	0.1969	0.1843

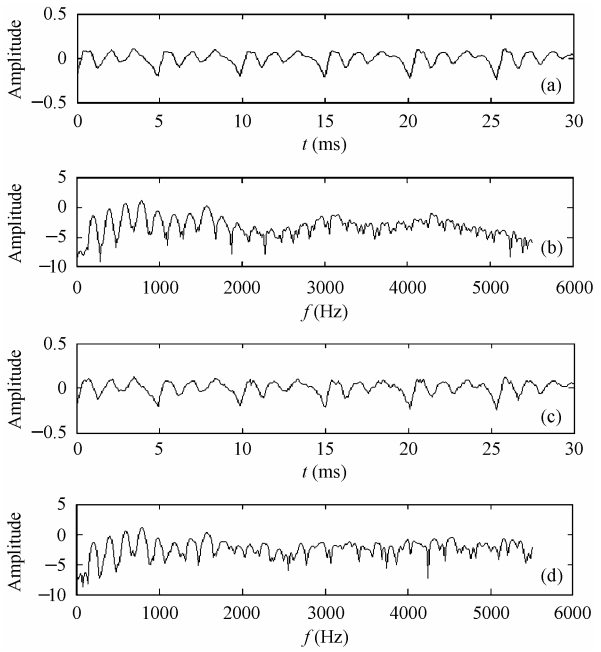

图 2　女声浊音帧的时域波形和短时谱　(a) 无噪声时的时域波形　(b) 无噪声时的对数谱　(c) SNR=10dB 时的时域波形　(d) SNR=10dB 时的对数谱

Fig.2 Voiced sound waveform and short-time spectrum from a female speaker　(a) Clean sound waveform　(b) Clean sound log-spectrum　(c) Noisy sound waveform with SNR=10dB　(d) Noisy sound log-spectrum with SNR=10dB

从表 1 中的数据可以看出，无论是男声还是女声，其短时谱的较低次谐波的幅度受噪声的影响相对较小，也就是说，基频及其低次谐波的幅度对加性白噪声不太敏感。下面我们将尝试利用这些特性构造特征矢量，并将其应用于基于 GMM 的说话人辨认系统，对其在 AWGN 环境下的性能进行测试。

3　一种新的基于基音频率及其谐波结构的语音特征参数

假定语音信号 $x(n)$ 的短时谱为 $X(f)$，基音频率为 f_0，则可以将 $f_0, 2f_0, \cdots, Nf_0$ 分别代入 $X(f)$，求出 $X(f_0)$，$X(2f_0), \cdots, X(Nf_0)$，由它们构成特征矢量，其中 N 表示谐波次数。

这里有两个问题需要考虑，首先，谐波次数 N 怎样选取，其次，$X(f_0)$，$X(2f_0), \cdots, X(Nf_0)$ 表示的是基波及其各次谐波的幅度，其中没有对基音频率 f_0 的描述。

先讨论前一个问题。语音信号包含十分丰富的谐波分量，而其基波分量往往不是最强的分量。因为语音的第 1 共振峰(Formant)通常在 300 ~ 1000Hz范围内，基音频率的有效范围通常是 60 ~ 400Hz[8]，也就是说，10 次谐波之内的谐波成分常常会有一些比基波分量还强，它们的位置和幅度与第 1 共振峰的位置(频率)和强度是密切相关的。而共振峰是反应声道谐振特性的重要特征，人们对语音的感知也利用了共振峰提供的信息[9]，而且在共振峰中第 1 共振峰和第 2 共振峰

是尤其重要的。因此，在建立特征矢量时，所选取的谐波成分应该不低于 10 次谐波，以保证至少能保留与第 1 共振峰有关的信息。

至于后一个问题，可以考虑将 f_0 值作为特征矢量的一个分量，与基波分量和各次谐波分量一起构成组合参数。

需要指出的是，在进行短时傅里叶分析时所加的窗函数不同，对各谐波幅度值的影响是不一样的。但只要在训练和识别时选用的是同一种窗函数，就不会造成不利影响。

根据上面的分析，可以建立基于基音频率及其谐波特性的语音特征参数为

$$\boldsymbol{V}=[X(f_0)\quad X(2f_0)\quad \cdots\quad X(Nf_0)\quad f_0] \tag{1}$$

式(1)中 $\boldsymbol{V}$ 为 N+1 维特征矢量，$\boldsymbol{V}$ 的前 N 个分量是短时谱 $X(f)$ 在基音频率 f_0 及其谐波频率上的幅度值，第 N+1 个分量是基音频率 f_0。

特征参数$\boldsymbol{V}$的提取过程如下：(1) 利用端点检测算法对输入语音信号进行端点检测，将语音信号中的空白音及能量较小的清音部分去掉。本文采用效果较好的双门限前端检测算法实现端点检测[10]。(2) 用窗长为 30ms的汉明(Hamming)窗对语音信号分帧，帧移为 15ms。(3) 对每帧语音信号 $x(n)$ 进行傅里叶分析，得短时谱 $X(f)$。(4) 利用短时谱检测出基音频率 f_0 并求出基波和各次谐波的幅度。

上述处理的 4 个步骤中最为关键的是基音频率检测。虽然无论在时域还是在频域，都已经有许多基音检测算法(Pitch Detection Algorithms, PDA's)被提出[11]，但如何精确可靠地判断浊音、检测基音频率依然没有得到很好的解决。困难来自语音信号的非平稳性和准周期性以及声门激励和声道之间的相互作用。

从减少运算量的角度考虑，可以采用“选峰法”，直接从已经获得的短时谱通过搜索局部最大值来确定基音频率。但由于基波往往不是最强的分量，这给基音检测带来了困难，经常发生基频估计结果为实际基音频率的二三次倍频或二次分频的情况。

综合考虑检测精度和方便实现两方面因素，本文采用频域基音检测方法中一种有效的方法——谐波积谱法[12]。

谐波积谱法是利用语音信号的短时谱 $X_n(e^{j\omega})$ 检测基音频率 f_0。信号 $x(n)$ 的谐波积谱定义为

$$P_n(e^{j\omega})=\prod_{r=1}^{k}\left|X_n(e^{jr\omega})\right| \tag{2}$$

对式(2)两边取对数，可得 $x(n)$ 的对数谐波积谱为

$$\log P_n(e^{j\omega})=\sum_{r=1}^{k}\log\left|X_n(e^{jr\omega})\right| \tag{3}$$

其中 $X_n(e^{jr\omega})$ 的频谱结构是 $X_n(e^{j\omega})$ 在频域压缩 r 倍的结果，它的 r 次谐波的峰值位置总是与 $X_n(e^{j\omega})$ 的基波频率对齐的。因此，无论原来的频谱中基波成分是否具有最高的峰，谐波积谱 $P_n(e^{j\omega})$ 都会在基频处出现最高的峰，即使所分析的语音信号是截去了低频分量的电话带宽语音，它也仍可在基频处获得最高的峰[12]。这就可以有效地避免基音频率估值落在倍频或分频上的可能性。这种技术特别能对抗加性噪声，因为噪声频谱中没有相干的结构。

检测出基音频率 f_0 之后，将 f_0，$2f_0,\cdots,Nf_0$ 分别代入信号 $x(n)$ 的短时谱 $X(f)$，即可得到特征矢量 $\boldsymbol{V}=[X(f_0)\ X(2f_0)\cdots X(Nf_0)\ f_0]$。

图 3 所示是用谐波积谱法对图 1 所示的男声和图 2 所示的女声在无噪声及信噪比(SNR)为 10dB 两种情况下检测基音频率的结果。在无噪声和有噪声两种情况下，实验检测到的男声基音频率都是 118.43Hz，女声基音频率都是 193.80Hz。对其他说话人的实验也可以达到同样的效果。

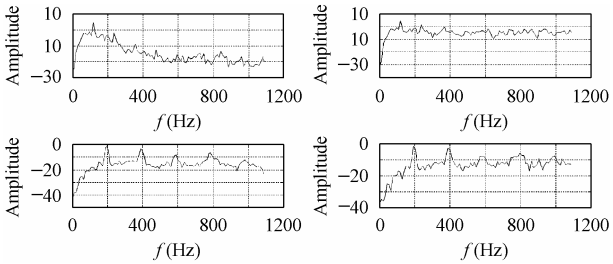

图 3 谐波积谱法检测基音频率(r=1~5) (a)男声在无噪声时的对数谐波积谱 (b)男声在 SNR=10dB 时的对数谐波积谱 (c)女声在无噪声时的对数谐波积谱 (d)女声在 SNR=10dB 时的对数谐波积谱

Fig.3 Pitch frequency detection using harmonic summation (r=1~5) (a)Clean sound log harmonic summation from a male speaker (b)Noisy sound log harmonic summation with SNR=10dB from a male speaker (c)Clean sound log harmonic summation from a female speaker (d)Noisy sound log harmonic summation with SNR=10dB from a female speaker

图 3 所示的基音频率检测结果是针对浊音帧进行的。实际上，在语音信号中，除了有基频特征十分明显的浊音外，还有基频有变化的短时帧，基音不显著的短时帧以及无基音特征的清音帧。虽然通过端点检测预处理可以去除部分能量较小的清音，但依然存在遇到非浊音帧、测不准基频的可能性。对于这种情况，处理方法是：在检测基音频率时，如果检测出的数值超出基音频率的有效范围(根据文献[8]，本文取为 60~400Hz)，则认为该数据无效而予以放弃，否则则按文中提出的方法加以利用。当然，不排除检测到的不是真正的基音频率，但数值却恰好在基音频率有效范围内的可能。这种情况下，如果依然按照其谐波特征构造特征参数会对识别性能产生一定的影响。但从后面的实验结果可以看出，这种现象并不严重。

4 新的语音特征参数在噪声环境下的说话人识别性能

为测试新特征矢量在说话人辨认系统中的性能，本节将该特征矢量与传统的线性预测倒谱系数(LPCC)和 Mel 频率倒谱系数(MFCC)相比较。3 种特征参数都取为 12 阶，识别对象为 20 人(10 男 10 女)，为每一说话人建立 12 阶 GMM。语音库是 20 个说话人分别阅读不同报刊时录取的每人 5min

左右的语音数据，读取前 60s 作为训练音，测试时从库中逐段读取 4s 的数据作为测试音，按信噪比(SNR)分别为 30dB，25dB，20dB，15dB，10dB 加入 AWGN，直至 20 个库全部读完，统计识别率，并与无噪声(SNR=∞)时的识别率进行比较。测试结果如表 2 和图 4 所示。

从实验结果可以看出，在无噪声情况下，LPCC 和 MFCC 的识别率接近 100%，而文中提出的新特征的识别率只有 90%左右。随着信噪比的降低，LPCC 和 MFCC 的识别率迅速减小，而新特征识别率的变化则相对平缓得多。也就是说，新特征在基于 GMM 的说话人识别系统中对 AWGN 相对较不敏感。

表 2　噪声环境下说话人识别率(%)

Tab.2 Speaker identification rate under noise environment (%)

信噪比 (SNR)	∞	30dB	25dB	20dB	15dB	10dB
LPCC	98.75	92.90	80.58	37.58	21.29	10.68
MFCC	99.37	94.99	80.77	44.26	26.70	15.03
新特征	90.33	89.39	89.02	85.85	83.49	62.26

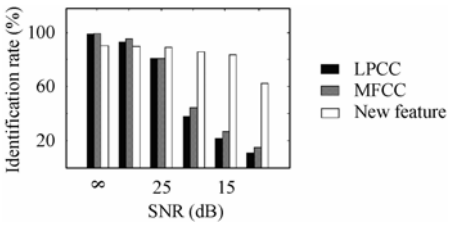

图 4　与文本无关说话人识别

Fig.4 Text-independent speaker identification

5　结束语

为了在特征域解决说话人识别系统的鲁棒性问题，文章通过对纯净浊音及含噪浊音短时谱的分析比较，提出了一种基于基音频率及其谐波结构的语音特征参数。该方法概念清楚、运算简单。基于 GMM 的说话人识别实验表明，在闭集文本无关说话人识别中，新特征具有较强的抗噪声性能，与传统的 LPCC 和 MFCC 相比，可以提高系统在低信噪比环境下的说话人识别率。但是，在无噪声(SNR=∞)或信噪比很高(SNR≥30dB)的情况下，新特征的性能不如传统的 LPCC 和 MFCC，这与基音这类参数中所含的话者个性特征不很充分有关。

需要指出的是，本文的工作只针对 AWGN 情况，在色噪声、卷积噪声等其它噪声情况以及实际噪声环境下的性能尚有待进一步研究。

参 考 文 献

[1] Murthy H A, Beaufays F, Heck L P, *et al*.. Robust text-independent speaker identification over telephone channels. *IEEE Trans. on Speech and Audio Processing*, 1999, 7(5): 554–568.

[2] Assaleh K T, Mammone R J. New LP-derived features for speaker identification. *IEEE Trans. on Speech and Audio Processing*, 1994, 2(4): 630–638.

[3] Hsieh C T, Lai E, Wang Y C. Robust speech features based on wavelet transform with application to speaker identification. *IEE Proc.-Vis. Image Signal Process*., 2002, 149(2): 108–114.

[4] Doddington G. Speaker recognition based on idiolectal differences between speakers. Eurospeech, 2001, 4: 2517–2520.

[5] Adami A, Mihaescu R, *et al*.. Modeling prosodic dynamics for speaker recognition. Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003 (ICASSP'03), Hong Kong, 2003, 4: 788–791.

[6] Peskin B, Navrati J, Abramson J,*et al*.. Using prosodic and conversational features for high performance speaker recognition: Report from JHU WS'02. Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003 (ICASSP'03), Hong Kong, 2003, 4: 792–795.

[7] Mammone R J, Zhang X, Pamachandran R P. Robust speaker recognition: A feature-based approach. *IEEE Signal Processing Magazine*, 1996, 13(5): 58–71.

[8] Ahmadi S, Spanias A S. Cepstrum-based pitch detection using a new statistical V/UV classification algorithm. *IEEE Trans. on Speech and Audio Processing*, 1999, 7(3): 333–338.

[9] 姚天任. 语音信号处理. 武汉：华中理工大学出版社，1992 年 4 月，第 6 章 6.8 节.

[10] 易克初, 田斌, 付强. 语音信号处理. 北京：国防工业出版社，2000 年 6 月，第 3 章 3.5 节.

[11] Hess W. Pitch Determination of Speech Signals. Springer-Verlag, Berlin, Germany: 1983.

[12] 易克初, 田斌, 付强. 语音信号处理. 北京：国防工业出版社，2000 年 6 月，第 4 章 4.4 节.

张玲华：　女，1964 年生，博士，副教授, 硕士生导师，主要研究方向为语音信号处理、智能信号处理等.

郑宝玉：　男，1945 年生，副校长, 教授, 博士生导师，上海交通大学兼职教授, 博士生导师，主要研究方向为智能信号处理及其在通信中的应用等.

杨　震：　男，1961 年生，副校长, 教授, 博士生导师，主要研究方向为语音信号处理及其在通信中的应用等.