基于时频分布与 MFCC 的说话人识别①

金银燕，于凤芹，何 艳

(江南大学 物联网工程学院，无锡 214122)

摘 要：针对 MFCC 不能得到高效的说话人识别性能的问题，提出了将时频特征与 MFCC 相结合的说话人特征提取方法。首先得到语音信号的时频分布，然后将时频域转换到频域再提取 MFCC+MFCC 作为特征参数，最后通过支持向量机来进行说话人识别研究。仿真实验比较了 MFCC、MFCC+MFCC 分别作为特征参数时语音信号与各种时频分布的识别性能，结果表明基于 CWD 分布的 MFCC 和 MFCC 的识别率可提高到 95.7%。

关键词：短时傅里叶变换；Wigner-Ville 分布；Choi-Williams 分布；Mel 频率倒谱系数；说话人识别

Speaker Recognition Based on Time-Frequency Distribution and MFCC

JIN Yin-Yan, YU Feng-Qin, HE Yan

(School of Internet of Things Engineering, Jiangnan University, Wuxi 214122, China)

Abstract: Because MFCC can't reflect the dynamic characteristics of speech signal and their own non-stationary, a feature extraction method by combining time-frequency distribution with MFCC is proposed. First get time-frequency distribution of speech signal, and convert time-frequency domain into frequency domain, then extract MFCC+MFCC as characteristic parameters. Finally speaker recognition uses the support vector machine. The simulation experiment compares recognition performance when MFCC and MFCC+MFCC are respectively as characteristic parameters by speech signal and all kinds of time-frequency distribution. Results show that the speaker recognition performance using MFCC+MFCC based on the CWD time-frequency distribution can be improved to 95.7%.

Key words: STFT;WVD;CWD; MFCC; speaker recognition

说话人识别是根据说话人所发的语音来确定说话人身份的过程。说话人识别注重的是语音信号中的个性特征，不同的特征参数描述了不同说话人的物理特性和声学特性，直接影响说话人识别的性能，因此有效地提取语音特征是说话人识别的关键。Mel 频率倒谱系数（Mel Frequency Cepstral Coefficients，MFCC）是目前普遍采用的声学特征参数，它根据人耳的听觉特性对卷积性信道失真有补偿的能力[1, 2]。文献[3]使音节和 MFCC 作为特征参数进行说话人识别，识别率为 84.11%；文献[4]提出了一种基于动态 MFCC 参数的识别方法，其识别率达到了 87.32%。动态的 MFCC 参数因考虑语音的动态特征而提高了识别率，但其识别性能还有待进一步提高。另外，由于语音信号是一种典型的非平稳信号，而时频分析方法是分析非平稳信号的有力工具[5]，它同时在时域和频域对信号进行分析，是近年来的研究热点。文献[6]是从图像处理的角度对时频分布进行处理，其识别率达到 86.52%。

针对以上情况，提出将时频特征和 MFCC 相结合进行特征提取，并使用支持向量机作为识别器来进行说话人识别。为了能够同时考虑语音的静态和动态特征，将 MFCC 和 MFCC 相结合来提取特征参数。仿真实验表明，MFCC 和 MFCC 相结合的特征参数在说话人识别中识别性能明显优于只使用 MFCC 参数的情况。而且在时频分析基础上再提取 MFCC 及 MFCC+

① 基金项目:国家自然科学基金(61075008)
收稿时间:2011-07-14;收到修改稿时间:2011-09-07

MFCC 参数，可以更好的提高说话人的识别性能。

1 基本理论

1.1 短时傅里叶变换

语音信号是一个非平稳的过程，其特性是随时间缓慢变换的，因此可假设它在一小段时间内是保持不变的，将短时分析的方法应用于傅里叶分析中，即有限时间的傅里叶变换称为短时傅里叶变换（Short Time Fourier Transform, STFT），相应的频谱称为短时谱[5]。

STFT 的基本思想是：把非平稳信号看成是一系列短时平稳信号的叠加，用窗函数将信号划分成许多小的时间间隔，对每个时间间隔进行傅里叶分析来确定那个时间间隔存在的频率。STFT 的定义为

$$STFT_x(t,f)=\int_{-\infty}^{\infty}x(\tau)h(\tau-t)e^{-j2\pi f\tau}d\tau \tag{1}$$

式中是 t 时间，f 是频率，$x(t)$ 是语音信号，$h(t)$ 是窗函数。

1.2 Wigner-Ville 分布和 Choi-Williams 分布

20 世纪 60 年代中期，Cohen 发现众多的时频分布可以用统一的形式表示，习惯称之为 Cohen.类时频分布，那么其表达式为：

$$P(t,f)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}z(u+\frac{\tau}{2})z^*(u-\frac{\tau}{2})\phi(\tau,v)e^{-j2\pi(tv+tf-uv)}dud\tau dv \tag{2}$$

式中 $z(t)$ 是信号 $x(t)$ 的解析信号，$\phi(\tau,v)$ 称为核函数。若 $\phi(\tau,v)=1$，即不加核函数，则 Cohen 类分布可退化为 Wigner-Ville 分布(WVD: Wigner-Ville Distribution)。

$$W_z(t,f)=\int_{-\infty}^{\infty}z(t+\frac{\tau}{2})z^*(t-\frac{\tau}{2})e^{-j2\pi\tau f}d\tau \tag{3}$$

WVD 是一种最基本，也是应用最多的时频分布，其重要特点之一是可以被看作信号能量在时域和频域中的分布[4]。用时频分析方法分析非平稳信号时，希望它在时频面上是高度集中，即具有良好的时频聚集性。WVD 对单分量信号具有比其他时频分布更好的时频聚集性，但对于多分量，WVD 存在的主要缺陷是出现交叉项，产生“虚假信号”。通常交叉项是振荡的，严重时，幅度可以达到自主项的两倍，造成信号的时频特征模糊不清。因此对于时频分析来说，抑制交叉项是非常重要的[5]。交叉项的抑制可以通过设计核函数来实现，这类时频分布称为 Cohen 类时频分布，其中一种是平滑伪 Wigner-Ville 分布(SPWD:Smoothed Pseudo Wigner-Ville Distribution)

$$SPWD_z(t,f)=\int_{-\infty}^{\infty}z(t-u+\frac{\tau}{2})z^*(t-u-\frac{\tau}{2})g(u)h(\tau)e^{-j2\pi\tau f}d\tau \tag{4}$$

其中 $z(t)$ 是 $x(t)$ 信号的解析信号，$g(u)$，$h(\tau)$ 是两个实的偶窗函数，且 $h(0)=G(0)=1$。

Choi 和 Willianms 提出了时频分析核，其表现形式为:

$$\phi(\tau,v)=\exp[-\frac{(2\pi\tau v)^2}{\sigma}] \tag{5}$$

Choi 和 Williams 核函数与时间和频率无关，是时延和频延的函数，具有时频移不变性。其中 σ 为衰减系数，它与交叉项的幅值成比例关系。

相应的分布为：

$$P(t,f)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}z(u+\frac{\tau}{2})z^*(u-\frac{\tau}{2})e^{-\frac{(2\pi\tau v)2}{\sigma}}e^{-j2\pi(tv+tf-uv)}dud\tau dv \tag{6}$$

式中 $z(t)$ 是信号 $x(t)$ 的解析信号。该分布称为 Choi-WilliamSSs 分布。注意当 $\sigma\to\infty$ 时，此时核函数趋近于 1，得到 WVD 分布，相反地，越小，相干项的衰减就越大。

1.3 Mel 频率倒谱系数

MFCC 是从人耳的听觉机理角度出发，利用 Mel 频率的弯折度来模拟声音频率与声音高低之间不成线性正比的非线性关系，针对听觉实验来分析语音信号的频谱[4]。生理学的研究结果表明，人耳具有在嘈杂的环境中以及各种变异情况下仍能正常地分辨出各种语音的功能，这是耳蜗起了关键的作用，由于人的内耳基础膜对外来信号会产生调节作用，所以它实质上充当了一个滤波器组。耳蜗的滤波器是根据主观音高在对数频率刻度上划分的一种合理刻度。其在 lKHz 以下为线性刻度，而在 lKHz 以上为对数刻度。根据这个原理，人们按临界带宽的大小设计了从低频到高频的一组由密到稀的带通滤波器组，实际频率与频率之间的转换公式为：

$$Mel(f)=2595\lg(1+\frac{f}{700}) \tag{7}$$

MFCC 利用了听觉原理和倒谱的解相关特性，还利用了 Mel 倒谱对卷积性信道失真有补偿的能力。这些特征不依赖于信号的性质，对输入的信号不做任何

假设和限制，降低信噪比时它仍有较好的识别性能。因此，MFCC 是语音识别中最成功的特征描述之一。

2 算法流程

本文主要是将时频分布与 MFCC 方法相结合来提取说话人的特征，并用支持向量机对说话人进行识别。具体流程如图 1 所示：

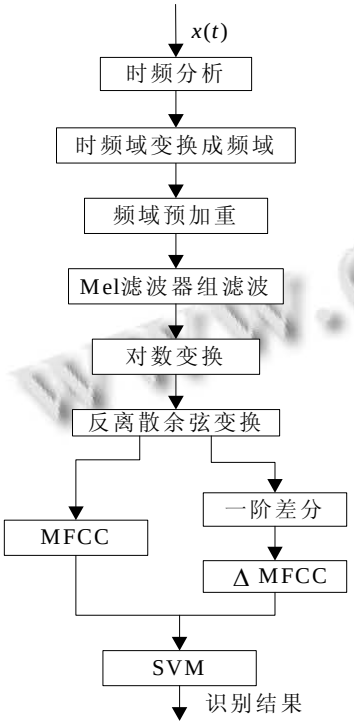


图 1 基于时频分布与 MFCC 的说话人识别流程图

1) 读取说话人的语音信号 $x(t)$，并进行时频分析得到语音的 STFT、WVD 分布和 CWD 分布。

2) 对时频图进行时间上的积分得到相应的频谱，并对频谱进行频域预加重处理，预加重滤波器为 $1-0.9375z^{-1}$。这里频域预加重的作用是对语音信号的高频成分进行提升，使频谱变得平坦；

3) 将预加重过的语音信号频谱用传递函数为 $H_l(k)$ 的 L 个 Mel 带通滤波器对功率谱 $|X(k)|^2$ 进行滤波，然后将每个滤波器的输出取对数；

4) 将对数功率谱 $M(l)$ 进行反离散余弦变换得到 N 阶 MFCC 系数；

5) 由于 MFCC 参数主要反映了语音信号的静态特性，而人耳对语音信号的动态特征更为敏感，因此其动态特征可以通过对 MFCC 参数求取一阶差分得到，即 Δ MFCC；

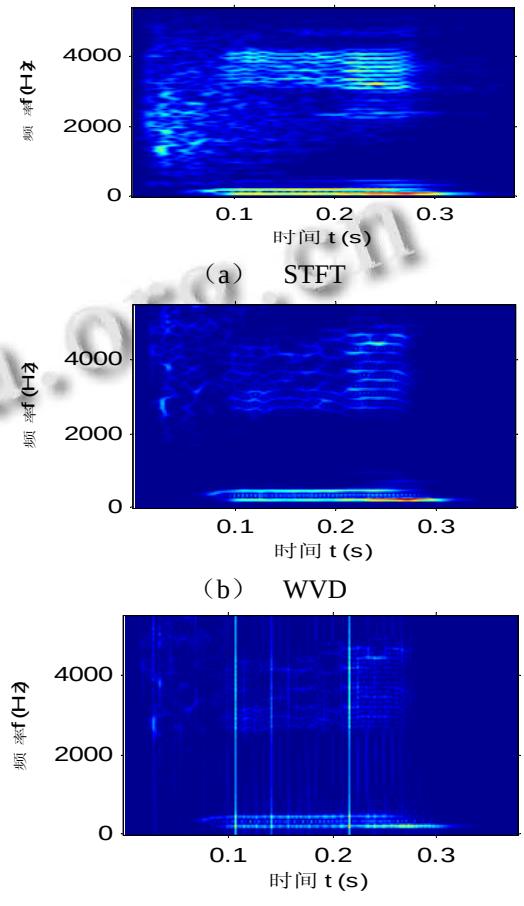
6) 最后将 MFCC 参数与一阶差分 MFCC 参数的组合作为语音信号的特征参数，送入支持向量机进行说话人识别。

3 仿真实验及结果分析

仿真实验中所用的语音库由 10 人录制，所有语音被分为训练集和测试集，其中训练集包括 10 个人的语音，每人各 3 段共 30 段；而测试集中包含了 10 个人的语音，每人各 1 段共 10 段。本实验先提取语音信号的 MFCC 参数，并求得其一阶差分系数 MFCC，然后将 MFCC 参数和一阶差分 Δ MFCC 参数组合起来作为语音的特征参数，最后将其送入支持向量机进行说话人识别。

3.1 时频分析

本实验中采用的语音信号是女生朗读“1”，采用频率为 11025HZ，分别用 STFT、WVD 和 CWD 对该语音信号进行时频分析，各种时频分布图如图 2 所示。



（a） STFT

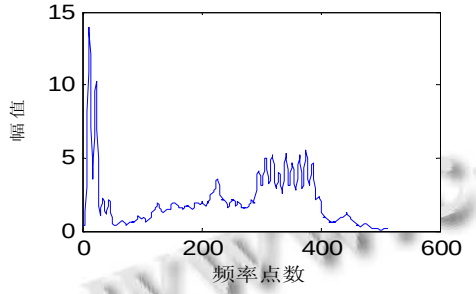
（b） WVD

（c） CWD

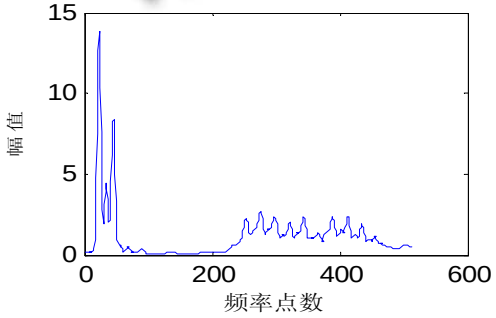
图 2 语音信号“1”的时频分布图

短时傅立叶变换是最早最简单的时频分析形式，而且是线性时频分布，从图 2（a）中可以看出短时傅立叶变换有很好的频域特征，能够区分出各个频率而不产生交叉项干扰。从图 2（b）可以看出 WVD 有很好的时频聚集性，但会产生交叉项干扰，而图 2（c）中的 CWD 虽然消除了交叉项，但是降低了时频聚集性。

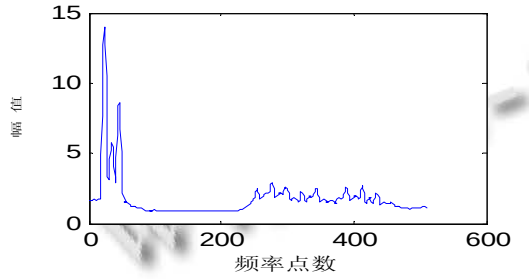
然后分别将图 2 中的各种时频分布图转换到频域上，得到相应的频谱，如图 3 所示。



（a） STFT 从时频域变换到频域



（b） WVD 从时频域变换到频域

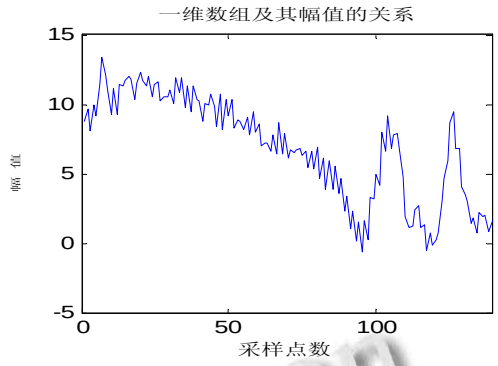


（c） CWD 从时频域变换到频域

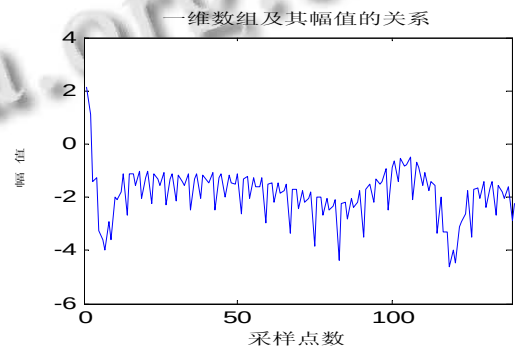
图 3　各种时频分布图从时频域转换到频域

3.2 MFCC+MFCC 特征提取

为了比较，图 4(a)是用传统方法提取的 MFCC+MFCC 特征参数。图 4(b)是将时频分布从时频域变换到频域后再提取 MFCC+MFCC 特征参数所得到的幅值关系，从图中可以全面的了解 MFCC 的静态及动态特性。



(a)　MFCC+MFCC



(b)　时频分布+ MFCC+MFCC

图 4　一维数组及其幅值的关系

3.3 识别性能的比较

本实验用上述方法提取语音的 MFCC 参数和 MFCC+MFCC 参数，分别作为支持向量机的输入，进行说话人识别。其性能比较如表 1 所示：

表 1　各时频分析方法识别性能的比较

	原始语音	STFT	WVD	CWD
MFCC	85%	86.64%	83.33%	86.64%
MFCC+ΔMFCC	92.4%	95%	90%	95.7%

从表 1 中可以看出，MFCC+MFCC 的识别性能明显优于 MFCC，这是主要是因为标准的 MFCC 只能反映语音信号的静态特性，不能精确地反映语音信号的动态特性，而一阶差分倒谱系数 ΔMFCC 能够较为精确地反映信号的动态特性，从而将 MFCC 和 ΔMFCC 相结合可以弥补各自的不足，使得 MFCC+ΔMFCC 具有很好的识别性能。由于 WVD 含有交叉项，掩盖了部分语音信号的特征信息，从而导致其识别性能明显不如其他方法；而 STFT 是线性时频表示，本身就不

(下转第 178 页)

活动涉及的变更集。项目采用三层流模式，即集成流，测试流，开发流，由配置管理员定期提交测试流通过的代码到集成流。生产变更到项目的同步采用 Clear quest 活动人工同步的方式。项目开发测试完成后，上线生产的时候，通过将原生产流隐藏的方式，将项目的流变成新的生产的流，版本发布在新的生产的测试流上进行。

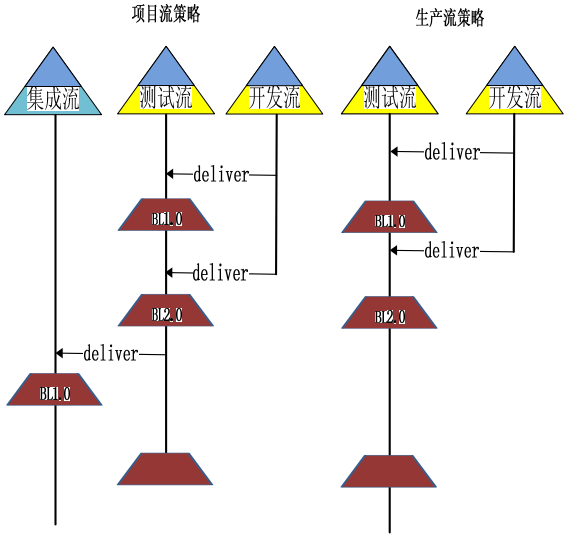


图 3　UCM 流策略示例 2

② 基线策略

在软件升级项目工程中配置管理员根据项目计划阶段性的创建基线和发布基线，在软件维护期的生产系统中根据活动的发布需求每个活动创建基线和取基线差异发布。

ClearQuest 的集成策略

在 ClearQuest 中定制开发流程，包括开发、测试、入库、缺陷、活动同步管理，通过固化的流程保证生产和项目的统一管理。

6　总结

持续改进型软件各个软件工程间如何同步管理变更是个复杂而艰巨的任务，使用 IBM Rational 提出的基于活动对软件进行统一变更管理（UCM），可以有效解决同步管理的问题。同时 UCM 是个灵活的工具，实施人员需要对 UCM 的思想进行深入的掌握，才能根据项目细节定制的对应方案和策略，达到最佳的实践。

参考文献

1 IBM.软件配置管理.IBM Rational 技术白皮书.V1.1.

(上接第 192 页)

含有交叉项，可以很好的体现语音信号的特征；CWD 是通过核函数来抑制 WVD 所存在的交叉项，故其识别性能优于 WVD。

4　结论

由于 MFCC 不能体现语音信号的动态特性，以及语音信号自身的非平稳性，提出了将时频分布与 MFCC 相结合的说话人特征提取方法。对语音信号进行时频分析得到相应的时频分布然后将时频域转换到频域再提取 MFCC+ΔMFCC 作为特征参数，通过支持向量机进行说话人识别研究。实验表明，说话人识别性能得到了很大的提高。

参考文献

1 刘丽岩.基于 MFCC 与 IMFCC 的说话人识别研究.哈尔滨:哈尔滨工程大学,2008.

2 余建潮,张瑞林.基于 MFCC 和 LPCC 的说话人识别.计算机工程与设计,2009,30(5):1189–1191.

3 Do CT, Pastor D, Goalic A. On the recognition of cochlear implant–like spectrally reduced speech with MFCC and HMM–based ASR. IEEE Trans. on Audio, Speech, and Language Processing, 2010,18(5):1065–1068.

4 张贤达,保铮.非平稳信号分析与处理.北京:国防工业出版社,1998:12–49.

5 Zhen B, Wu X, Liu Z. On the importance of components of the MFCC in speech and speaker recognition. Acta Scicentiarum Naturalum Universitis Pekinesis, 2001,37(3): 371–378.

6 Ferras M, Leung CC, Barras C, et al. Comparison of speaker adaptation methods as feature extraction for SVM-based speaker recognition. IEEE Trans. on Audio, Speech, and Language Processing, 2010,18(6):1366–1378.

7 Manikandan J, Venkataramani B. Design of a modified one-against-all SVM classifier. Proc. of the 2009 IEEE International Conference on Systems, Man, and Cybernetics. San Autonio,2009:1869–1874.