

文章编号:1001-9081(2007)08-2051-02

## 变帧长和变帧率在说话人确认中的应用

王 明, 肖 熙

(清华大学 电子工程系, 北京 100084)

(wangming00@mails.thu.edu.cn)

**摘要:**从变帧长、变帧率角度考虑提出一种新的提取 MFCC 的方法。该方法先将帧长和帧率都限制为基音周期的整数倍,即基音同步算法;然后基于变帧率算法的原理在语音特征变化缓慢的地方去除一些帧来降低帧率。在 NIST 99 说话人评测上进行的说话人确认实验表明,该方法不但提升了系统性能,而且降低了帧率,节省了特征文件的存储空间。

**关键词:**说话人确认;基音同步;变帧率算法

**中图分类号:** TP391.42    **文献标志码:**A

## Application of variable frame length and frame rate in speaker verification system

WANG Ming, XIAO Xi

(Department of Electronic Engineering, Tsinghua University, Beijing 100084, China)

**Abstract:** A new method for extracting Mel-Frequency Cepral Coefficients (MFCC) was proposed from the perspective of variable frame length and frame rate. The proposed method restricted the frame length and frame shift to multiples of pitch period, called pitch synchronous algorithm; then removed some frames where the acoustic feature changed slowly to decrease the frame rate according to the principle of variable frame rate algorithm. With speaker verification experiments on NIST 99 speaker recognition evaluation, the new approach not only improves the system performance but also decreases frame rate, which means saving the storage space of feature files.

**Key words:** speaker verification; pitch synchronization; variable frame rate algorithm

## 0 引言

一个典型的说话人确认系统包含前端的特征提取和后端的模型选择,前端的特征提取一般采用的是 Mel 频率倒谱系数 (Mel-Frequency Cepral Coefficients, MFCC) 特征,后端的模型一般选择全局背景—高斯混合模型 (UBM-GMM)<sup>[1]</sup>,本文所关注的是 MFCC 特征提取部分。MFCC 的核心思想是采用了人耳临界带分析的方法,考虑到人的听觉系统是一个特殊的非线性系统,它响应不同频率信号的灵敏度不同,其与线性频率的映射关系如式(1):

$$F_{\text{mel}} = 1127 \ln(1 + \frac{f}{700}) \quad (1)$$

传统的提取 MFCC 的方法大都是采用固定帧长和固定帧率,例如帧长取为 25 ms,帧移 10 ms,这么做是基于这样的认识:从整体时间段上看语音信号是非平稳的,而在短时域内可以认为是短时平稳的。不过在实际的语音处理中,有些帧内的信号却是非平稳的,采用固定帧长提取特征会导致频谱的模糊,采用固定帧率会导致信息的丢失,从而导致说话人确认系统的性能下降。为了解决这个问题,本文提出一种新的提取 MFCC 的方法,新方法分两步进行:1)利用基音周期来调整帧长和帧率,这一步完成之后系统性能已经有了提高,但是帧率相比原来的方法增大了许多;2)进行丢帧处理来降低帧率,将前一步计算得到的 MFCC 信息和语音的能量分布作为丢帧的依据。实验表明,丢帧运算在降低帧率的同时也能带来系统性能的提升。

## 1 算法介绍

### 1.1 基音同步算法

MFCC 作为频域参数,是一种基于对数功率谱的特征,因此 MFCC 对频谱中的小信号分量变化较敏感。语音浊音段的谐波分量都比较小,当分析窗的长度不是基音周期的整数倍时,谱分析中的谐波幅度容易受到强信号的基频分量信号的干扰,这种干扰在识别时会导致识别率下降<sup>[2]</sup>。当分析窗的长度为基音周期的整数倍时,分析窗频谱的零点正好对准谐波的整数倍位置。因此可以有效地抑制强基频信号对谐波分量的影响<sup>[3,4]</sup>。另外,传统方法对帧选取的位置信息也比较敏感<sup>[5]</sup>,而这种选取方法则基本可以忽略帧选取的位置。

这种算法的前提是要准确提取出语音信号的基音轮廓,本文采用 ESPS 语音工具包中的 get\_f0 算法<sup>[6]</sup>,这种算法的核心思想分为以下几步:

- 1) 将语音信号  $S$  经过 LPC 逆滤波器得到残差信号  $S_r$ 。
- 2) 将残差信号  $S_r$  低通滤波,去除非语音段的高频噪声,得到  $S_l$ 。
- 3) 计算  $S_l$  信号的归一化自相关函数,峰值点作为基音周期的候选点。
- 4) 利用动态规划 (Dynamic Programming, DP) 技术对上述候选点进行后处理。

实验表明,这种方法得到的基音轮廓抗噪性能较好。

整个基音同步算法的流程如图 1 所示,首先对输入的一整段语音提取基音频率  $f_0$  的轮廓,由  $f_0$  的轮廓可以判断出这段语音的浊音段和非浊音段,然后对浊音段和非浊音段分

收稿日期:2007-02-09;修回日期:2007-03-30。

作者简介:王明(1982-),男,江苏南京人,硕士研究生,主要研究方向:说话人识别; 肖熙(1967-),男,福建福州人,副研究员,主要研究方向:语音信号处理。

开进行处理,非浊音段保持原有的帧长和帧移,而浊音段的处理过程如下:

1)若 2 倍基音周期对应的采样点数小于 FFT 的点数,则帧长取为 2 倍基音周期,否则取为一倍基音周期。

2)把帧移取为一倍基音周期,然后做微调,令下一帧开始点为  $P$ ,在  $P$  点左右一定范围内搜索能量最小点  $Q$ ,把  $Q$  点置成下一帧的开始点,即保证每帧的开始点都是局部能量最小点。

最后按照传统的方法提取 MFCC,以下把这种方法提取的特征记为 PS-MFCC(Pitch Synchronization, PS)。

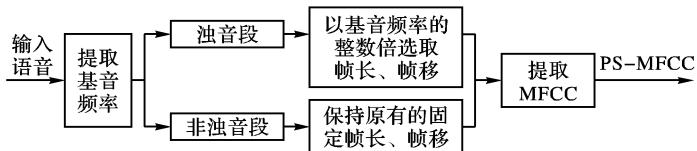


图 1 基音同步算法流程

## 1.2 丢帧算法

语音识别的 HMM 是基于特征的帧间独立性假设的,说话人确认的 GMM 可以认为是 HMM 的一种特例,令  $X = \{x_1, \dots, x_T\}$  为特征向量的一个序列,  $T$  为帧数,则  $X$  相对于模型  $\lambda$  的对数似然值计算如下:

$$\log p(X | \lambda) = \sum_{t=1}^T \log p(x_t | \lambda) \quad (2)$$

式(2)是在特征的帧间独立这个前提下得到的,而实际中帧间是不独立的,所以由式(2)计算出来的对数似然值小于真值,可以用变帧率算法来减小这种误差。在语音特征变化明显的地方,帧间信息的冗余度小,应当多取一些帧来保证采集到足够多的可用信息;而在语音特征变化缓慢的地方,帧间信息的冗余度大,就可以少取一些帧。这种算法在实现的时候通常是先把帧移取小,例如 2.5 ms,然后再依据一定的准则进行丢帧运算<sup>[7]</sup>。

本文算法的第一步是基音同步算法,帧移已经限制为一倍基音周期,为了保证帧移是基音周期的整数倍,在第二步只进行丢帧运算,一方面可以把帧率降下来,另外一方面,经过上面的分析,也能提高系统的性能。这里,丢帧运算的准则选取是关键,可以利用能量、信息熵<sup>[8]</sup>,也可以利用频域信息,如 MFCC,下面列举的算法选择了 MFCC 和能量。首先考虑 MFCC 的变化,MFCC 变化小的区域是丢帧的首选,不过有些能量低的地方 MFCC 的变化也很明显,这些可能是干扰信号,应该考虑去除影响,所以在计算相邻帧的 MFCC 差值时又加入了能量因子,最后得到的整个算法步骤如下:

1) 在第一步中已经得到了 PS-MFCC,计算第  $i - 1$  帧和第  $i$  帧的 PS-MFCC 向量的欧式距离  $d_m(i)$ ,再令  $d(i) = d_m(i) \cdot (E - \beta)$ ,这里  $E$  是第  $i$  帧的 log 能量,  $\beta$  为一常数。

2) 计算  $d(i)$  的平均值  $d$ 。

3) 计算下一步挑选帧时的门限值  $\theta$ ,  $\theta = \alpha \cdot d$ ,其中  $\alpha$  为控制帧率的一个可调参数。

4) 这一步挑选帧,从第  $n$  帧将  $d(i)$  值累加,如果在第  $n + k$  帧时累加值超过  $\theta$ ,则选取第  $n + k$  帧,然后从第  $n + k + 1$  帧继续选取过程,初始化时令  $n = 1$ 。

用以上挑选出来的帧数据组成新的 MFCC 向量,下面简记为 PS-TF-MFCC(Throw Frame, TF)。

## 2 说话人确认实验

### 2.1 实验设置

本文的实验数据基于 1999 年 NIST 的说话人评测<sup>[9]</sup>数据

集,选择其中的 One-Speaker Detection 任务,对每个说话人提供 2 min 左右的训练语料,分别从该说话人的两次通话中截取,总共有 230 个男性说话人,309 个女性说话人。测试语音共有 3420 段,长度从 0.5 ~ 60 s, 静音段已被去除,用来测试的每一段语音,都要对 11 个假想说话人进行说话人确认实验,这里,男性和女性说话人分别是两个测试集合。

实验的前端处理采用 14 维的 MFCC,后端模型采用 UBM-GMM 模型,UBM 和 GMM 的混合分量取 256 维,先用所有说话人的语音数据训练出 UBM 模型,再用 MAP 自适应训练出每个说话人的模型,共进行了三类实验:基线实验提取 MFCC、基音同步实验提取 PS-MFCC、基音同步加丢帧实验提取 PS-TF-MFCC。

### 2.2 结果分析

首先用 DET 曲线<sup>[10]</sup>来表征各个系统的性能,DET 曲线直观地表示出了虚警率和漏报率之间的关系,曲线越靠近零点,说明说话人确认系统的性能越好,图 2 用来表征男性说话人系统,图 3 表征女性说话人系统,从图上可以明显地看出,MFCC、PS-MFCC、PS-TF-MFCC 对于说话人确认系统,性能是逐渐提高的。

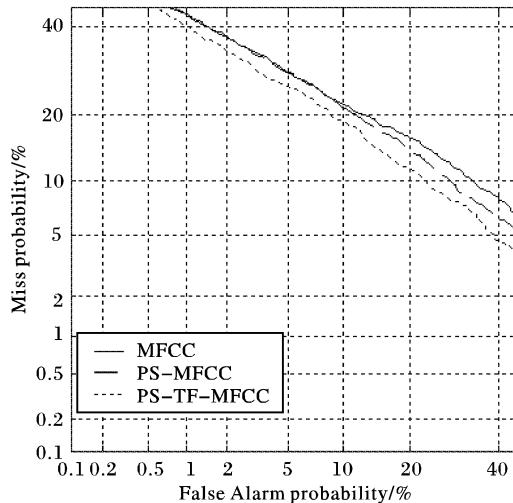


图 2 男性说话人三类系统的 DET 曲线

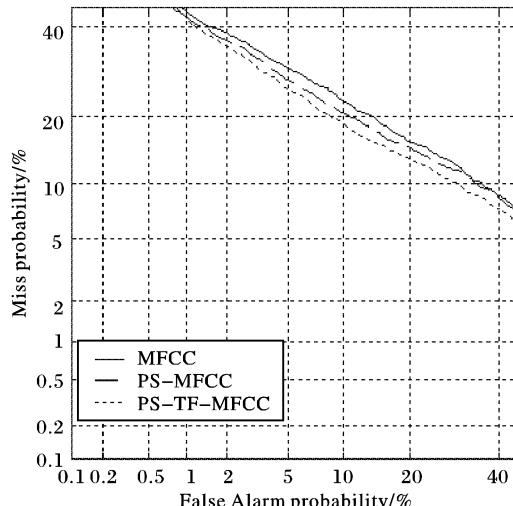


图 3 女性说话人三类系统的 DET 曲线

其次用如下三个指标来定性反应系统性能:

1)等错误率(Equal Error Rate, EER):虚警率和漏报率相等时的数值。

(下转第 2076 页)

- Management, 2006, 43(2):179 – 193.
- [3] ROSENBERG F, DUSTDAR S. Business rules integration in BPEL—a service-oriented approach[ C]// Proceedings of the 7th IEEE International Conference on E-Commerce Technology. Washington: IEEE Computer Society, 2005: 476 – 479.
- [4] WfMC, Workflow Management Coalition. XML Process Definition Language (XPDL) Version 1.0 [EB/OL]. (2002-07-31) [2006-12-19]. <http://www.wfmc.org/>.
- [5] WfMC, Workflow Management Coalition. XML Process Definition Language (XPDL) Version 2.0 [EB/OL]. (2005-10-03) [2006-12-19]. <http://www.wfmc.org/>.
- [6] RUTH SARA AGUILAR-SAV? N. Business process modeling: review and framework[ J]. Internal Journal of Production Economics, 2004, 90(2): 129 – 149.
- [7] BPEL4WS, Business process execution language for web services version 1.1[ EB/OL]. (2003-05-05) [2006-12-19]. <http://www-128.ibm.com/developerworks/library/specification/ws-bpel/>.
- [8] LEE S, KIM T-Y, KANG D, et al. Composition of executable business process models by combining business rules and process flows [J]. Expert Systems with Applications, 2007, 33(1):221 – 229.
- [9] The Business Rules Group. Defining business rules—what are they really?[ EB/OL]. (2000-07) [2006-12-19]. [http://www.businessrulesgroup.org/first\\_paper/bro1c1.htm/](http://www.businessrulesgroup.org/first_paper/bro1c1.htm/).

(上接第 2052 页)

2) 检测代价函数(Detection Cost Function, DCF): DCF 定义为虚警和漏报的后验概率加权<sup>[9]</sup>, 公式如下:

$$C_{\text{det}} = c_{\text{miss}} \cdot E_{\text{miss}} \cdot P_{\text{target}} + c_{fa} \cdot E_{fa} \cdot (1 - P_{\text{target}}) \quad (3)$$

SRE'99 中各参数一般取如下值:

$$c_{\text{miss}} = 10; c_{fa} = 1; P_{\text{target}} = 0.01$$

3) 帧率(Frame Rate, FR): 后两种算法的 FR 是不固定的, 只能考虑其平均值, 可以通过对所有语料(训练和识别)生成特征文件的大小来估算, 令基线系统特征文件的大小为  $M_0$ , 新算法特征文件的大小为  $M_1$ , 则新算法的平均帧率为:

$$FR = \frac{M_1}{M_0} \times 100 \text{ (frame/s)} \quad (4)$$

各类实验结果的指标见表 1、2。

表 1 男性说话人三类实验指标比较

	Male	MFCC	PS-MFCC	PS-TF-MFCC
EER/%	17.24	16.40	14.26	
DCF/%	5.20	5.27	4.97	
FR/frame · s <sup>-1</sup>	100	113	48.7	

表 2 女性说话人三类实验指标比较

	Female	MFCC	PS-MFCC	PS-TF-MFCC
EER/%	17.17	15.98	15.06	
DCF/%	5.36	5.22	5.14	
FR/frame · s <sup>-1</sup>	100	161	69.1	

EER 和 DCF 分别是 DET 曲线上的两个点, EER 是曲线上的中点, 而 DCF 则偏左上角, 对应虚警率较低, 这种考虑对应的是系统安全性要求较高的情况。首先考虑 EER, 无论是男性说话人还是女性说话人, 三类实验的 EER 都是逐渐降低的, 这从 DET 曲线上可以观察出来, 但是男性说话人的 PS-MFCC 相对于 MFCC 的 DCF 却升高了, 从 DET 曲线上观察, 男性说话人的 PS-MFCC 曲线只有在高虚警率的时候性能有提高, 在低虚警率的时候性能基本维持不变, 而女性说话人做基音同步效果则要好一些, 这是因为女性说话人的基音频率变化范围比较大<sup>[2]</sup>, 基音同步正好可以消除这个影响, 不过无论男性女性, 在做完丢帧运算以后 DCF 全都降低, 系统性能都有很大提高。

从表 1、2 还可以看出, 只做基音同步 FR 是增高的, 女性说话人由于基音频率高, 做基音同步的帧移就小, 算出来特征文件的存储量就很大, 增加了近 60%, 所以在基音同步基础上做丢帧运算是很有必要的, 在提高确认系统性能的同时也节约了存储空间。

### 3 结语

本文提出的 PS-TF-MFCC 从变帧长、变帧率的角度出发, 综合利用了基音信息、能量信息, 可以弥补传统 MFCC 的一些缺陷, 在提升系统性能的同时也节省了特征文件的存储空间, 可以作为说话人确认系统中一种实用的提取特征的方法。

#### 参考文献:

- [1] REYNOLDS D, QUATIERI T, DUNN R. Speaker verification using adapted mixture models[ J]. Digital Signal Processing, 2002, 10(1-3): 181 – 202.
- [2] QUATIERI T, DUNN B, REYNOLDS D. On the influence of rate, pitch, and spectrum on automatic speaker recognition performance [C]// ICSLP. Beijing: [ s. n. ], 2002: 491 – 494.
- [3] ZILCA R, NAVRATIL J, RAMASWAMY G. Depitch and the role of fundamental frequency in speaker recognition[ C]// ICASSP'03, Hong Kong. [ S. l. ]: IEEE Press, 2003, 2: 81 – 84.
- [4] ZILCA R, NAVRATIL J, RAMASWAMY G. Synepitch: A pseudo pitch synchronous algorithm for speaker recognition[ C/OL]// EUROSPEECH, Geneva, Switzerland, 2003 [2007-01-15]. [http://www.research.ibm.com/CBG/papers/eurospeech03\\_synepitch.pdf](http://www.research.ibm.com/CBG/papers/eurospeech03_synepitch.pdf).
- [5] KIM S, ERIKSSON T, KANG H-G, et al. A pitch synchronous feature extraction method for speaker recognition [ C]// ICASSP'04, Montreal, Canada. [ S. l. ]: IEEE Press, 2004, 1: 405 – 408.
- [6] SECREST B, DODDINGTON G. An integrated pitch tracking algorithm for speech systems[ C]// ICASSP'83, Boston, Massachusetts. [ S. l. ]: IEEE Press, 1983, 8: 1352 – 1355.
- [7] ZHU Q F, ALWAN A. On the use of variable frame rate analysis in speech recognition[ C]// ICASSP'00, Istanbul, Turkey. [ S. l. ]: IEEE Press, 2000: 1783 – 1786.
- [8] YOU H, ZHU Q F, ALWAN A. Entropy-based variable frame rate analysis of speech signals and its application to ASR [ C]// ICASSP'04, Montreal, Canada. [ S. l. ]: IEEE Press, 2004: 549 – 552.
- [9] MARTIN A, PRZYBOCKI M. The NIST 1999 speaker recognition evaluation an overview [ J]. Digital Signal Processing, 2000, 10(1): 1 – 18.
- [10] MARTIN A, DODDINGTON G, KAMM G, et al. The DET curve in assessment of detection task performance[ C]// Proceedings of 5th European Conference on Speech Communication and Technology ( Eurospeech'97). Rhodes, Greece: [ s. n. ], 1997: 1895 – 1898.