第 31 卷第 5 期
2006 年 9 月

声 学 学 报
ACTA ACUSTICA

Vol. 31, No. 5
Sept., 2006

Mel 子带谱质心和高斯混合相关性在鲁棒话者识别中的应用

邓 菁 郑 方 刘 建 吴文虎
(清华大学计算机科学与技术系 北京 100084)

2005 年 4 月 15 日收到
2005 年 7 月 12 日定稿

摘要 提出了两种方法以克服背景噪音的干扰并提高说话人识别系统的鲁棒性：一种方法是基于频谱峰值位置受背景噪音影响相对较小的考虑，将子带幅度信息和子带 Mel 频谱质心 (SMSC) 相结合；另一种方法是通过计算类转移概率矩阵来对隐藏于高斯混合相关 (GMC) 中的说话人高层信息进行建模。实验表明 SMSC 和 GMC 都能够在平稳噪音环境下提高说话人识别系统的鲁棒性，并且采用 SMSC 和 GMC 的 GMM-UBM 系统跟使用传统 MFCC 的 GMM-UBM 基准系统相比，平均错误率下降了 11.7%。

PACS 数： 43.60, 43.70

Using subband Mel-spectrum centroid and Gaussian mixture correlation for robust speaker identification

DENG Jing ZHENG Fang LIU Jian WU Wenhu
(*Department of Computer Science and Technology, Tsinghua University* Beijing 100084)

Received Apr. 15, 2005
Revised Jul. 12, 2005

Abstract In order to overcome the influence of background noises and improve the robustness of speaker identification systems, two methods were proposed: One is to incorporate subband amplitude information with subband Mel-spectrum centroid (SMSC) because spectral peak positions remain practically unaffected in presence of additive noise. The other is to use a class transition probability matrix to model the high-level information hidden in Gaussian mixture correlation (GMC). Experiments showed that SMSC and GMC could improve the robustness of a speaker identification system in stationary noises, respectively. The average error rate of GMM-UBM system using SMSC and GMC can be reduced by 11.7% compared to conventional GMM-UBM system using MFCC.

引言

在实际应用环境中，由于语音信号受各种背景噪音的影响，说话人识别系统的性能将会大大下降。噪音鲁棒性的研究大致可以分为特征级和模型级两种，前者主要针对特征提取的过程研究去除噪音干扰的方法，后者主要是通过模型补偿的手段研究去除噪音干扰的方法。特征级上的噪音鲁棒方法主要有：倒谱均值减、倒谱归一化[1,2]、谱减法[3,4]、非线性谱减法[5]和维纳滤波[6,7]，正则相关分析的谱变换补偿法[8]，多种特征综合法[9]，等等。这些方法都是在所提特征的基础上进行噪音补偿和噪音抑制，因而需要对相应的背景噪音进行估计。本文试图利用语音中受噪音干扰小的频谱峰值位置信息和高层说话人信息来提高说话人识别系统的鲁棒性。

通常说话人识别系统使用的特征提取方法大多是基于语音短时幅度谱估计的。这些方法利用了幅度信息却忽略了频谱峰值位置信息。Sönmez[10,11]等人的研究表明，频谱峰值位置信息可以用来提高说话人识别系统的性能。近些年来，许多研究者把描述频谱峰值位置信息的子带频谱质心 (Subband Spectrum Centroid, SSC) 作为 MFCC 的附加特征或作为基于

SSC 的新特征矢量[12−14]，一定程度上提高了语音识别或说话人识别系统的性能。Paliwal[12] 的研究表明 SSC 非常接近于频谱中的峰值位置。由于频谱峰值位置受背景噪音的影响相对较小，因此基于 SSC 的语音前端处理能够提高说话人识别系统对背景噪音的鲁棒性。但是这些方法中，频谱峰值位置信息与幅度信息的利用是相互独立，互不相关的。本文提出了一种新的方法，在 Mel 子带上将子带峰值信息和基于 MFCC 的语音前端处理结合起来，提高了说话人识别系统的抗噪性。实验表明，基于子带 Mel 频谱质心 (Subband Mel Spectrum Centroid, SMSC) 的语音前端处理能够提高说话人识别系统的鲁棒性，并且与基于传统 MFCC 的说话人识别系统相比，在各种信噪比等级的高斯白噪音环境下平均错误率下降了 2.3%。

近年来，基于高斯混合模型和通用背景模型 (Gaussian Mixture Model-Universal Background Model, GMM-UBM)[15,16] 的说话人识别方法已成为说话人识别系统的主要方法之一。基于 GMM-UBM 的系统所用的特征 (如 MFCC) 大多基于短时、低层的声学信息。虽然这些系统在训练语音和测试语音环境匹配较好的情况下，能够取得较好的识别率，但是在不匹配的情况下，系统性能急剧下降。而高层声学信息，如习惯用语或方言等[17]，受环境干扰相对较小，因而能够提高说话人识别系统在噪音和跨信道方面的鲁棒性[18,19]。但是高层声学信息相对于低层声学信息来说，一般较难提取或建模。基于隐马尔可夫模型 (HMM) 的语音识别利用了高斯混合间的相关性 (Gaussian Mixture Correlation, GMC)，并在孤立词识别上取得了较好的效果[20−22]。但是多状态的 HMM 很难应用于文本无关的说话人识别应用中[23]。鉴于基于 HMM 的语音识别中 GMC 能够提高语音识别系统的性能，本文设想 GMC 在一定程度上反映了某些说话人相关的高层信息。在 GMM-UBM 的基础上，本文提出了一种用类转移概率矩阵来对 GMC 建模的方法。实验表明，该方法能够提高说话人识别系统的鲁棒性，并且与使用传统 MFCC 的说话人识别系统相比，在各种信噪比等级的高斯白噪音环境下平均错误率下降 8.1%。而采用 SMSC 和 GMC 的系统，与传统说话人识别系统相比，平均错误率下降 11.7%。

1 提出的方法

1.1 基于 SMSC 的语音前端处理

假设频段 $[0, F_s/2]$ 被分成 M 个子带，其中 F_s 是语音信号的采样频率。对第 m 个子带，假设它的最低和最高频率边界分别为 l_m 和 h_m。设子带滤波器为 $w_m(f)$，频率 f 处的能量为 $P(f)$。根据 Paliwal[12] 的研究，第 m 个子带的质心由下式计算：

$$C_m = \frac{\int_{l_m}^{h_m} f w_m(f) P^\gamma(f) \mathrm{d}f}{\int_{l_m}^{h_m} w_m(f) P^\gamma(f) \mathrm{d}f}, \tag{1}$$

其中 γ 是一个经验值。根据 Bojana[14] 等人的实验表明，当 γ 设为 1 时，系统可以取得较好的性能，因此在实验中使用相同的数值。

为了将子带频谱质心信息和传统 MFCC 语音前端处理结合起来，在本方法中，f 采用的是Mel频率。$P(f)$ 为第 m 个 Mel 滤波器的输出，而 $w_m(f)$ 为传统 MFCC 提取方法中使用的三角滤波器。通过公式 (1)，可以得到子带 Mel 频谱质心序列 $\{C_m, 1 \le m \le M\}$，其中 M 是 Mel 频率滤波器组的个数。新的 Mel 滤波器组输出由下式计算得到：

$$O'(m) = \frac{O(m)(C_m - l_m)}{h_m - l_m}, \tag{2}$$

其中 $O(m)$ 是第 m 个滤波器的初始输出，而 $O'(m)$ 为新的输出。公式 (2) 能够在不同语音在同一个 Mel 滤波器上输出相同的情况下，根据该子带 Mel 频谱质心位置的不同对输入语音加以区别。

$\{O'(m), 1 \le m \le M\}$ 经过对数压缩、DCT 和倒谱系数提升后，可以得到基于 SMSC 的倒谱系数，其提取过程参看图 1。

1.2 使用高斯混合相关的 GMM-UBM

传统的 GMM-UBM 系统认为高斯混合间都是相互独立的，也就是说忽略了它们之间的相关性。而本文认为这种相关性包含着某些说话人相关的高层信息并能够用来提高说话人识别系统的性能。基于此种考虑，本文将 UBM 中的 M 个高斯混合根据其均值聚成 K 类 ($K < M$)，并用 K 类间的转移概率矩阵来表示高斯混合的相关性。

算法详细描述如下：给定一个 UBM $ubm = \{g_m, \boldsymbol{\mu}_m^{ubm}, \boldsymbol{\Sigma}_m | m = 1, 2, \cdots, M\}$，其中 $\boldsymbol{\mu}_m^{ubm}$，$\boldsymbol{\Sigma}_m$ 和 g_m $(m = 1, 2, \cdots, M)$ 分别是均值向量，协方差矩阵和第 m 个高斯混合的权重。根据 K- 均值法将 M 个高斯混合聚成 K 类，即 $\{C_k^{ubm}, w_k^{ubm} | k = 1, 2, \cdots, K\}$，其中 w_k^{ubm} 是第 k 类的权重系数，即第 k 类中高斯混合数与 M 的比值。同时记录每个高斯混合索引号与类的对应信息。图 2 为 UBM 中高斯混合聚类示意图。

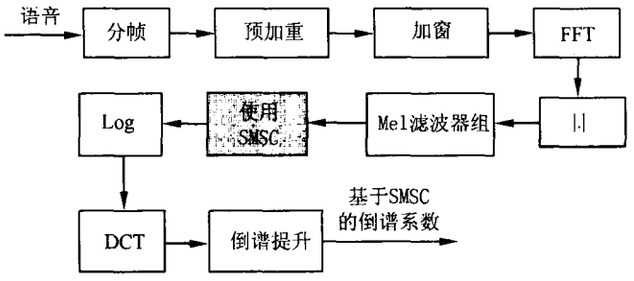

图 1　基于 SMSC 的倒谱系数提取示意图 (阴影部分为与传统 MFCC 提取方法不同的地方)

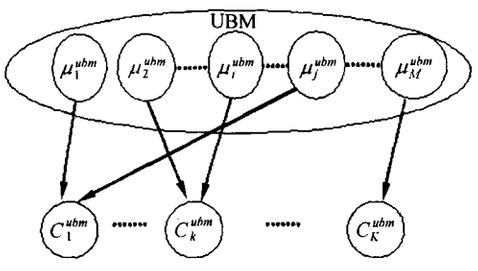

图 2　UBM 聚类示意图

假设 UBM 的训练特征序列为 $X=\{x_1,x_2,\cdots,x_T\}$，用一个通用类转移概率矩阵，即 $\{a_{ij}^{ubm}|1\le i,j\le K\}$，来描述大多数人发音的习惯，其计算步骤如下：

(1) 对每帧特征向量 x_t，计算 $g_m f(x_t|\mu_m^{ubm},\Sigma_m)$ 中最大 N 值的对数，其中 $f(\cdot,\cdot)$ 是高斯密度函数。这样可以得到由高斯混合索引号构成的 N 个序列 (这里 N 是一个经验值)，表示如下：

$$\left\{IG_n^{ubm}(t)|t=1,2,\cdots,T\right\},\quad 1\le n\le N \tag{3}$$

其中 $IG_n^{ubm}(t)$ 为第 t 帧特征在 UBM 上打分的第 n 极大值对应的高斯混合索引号。

(2) 通过查询高斯混合索引号与类的对应表可以得到由类的序号构成的 N 个序列，表示为：

$$\left\{IC_n^{ubm}(t)|t=1,2,\cdots,T\right\},\quad 1\le n\le N \tag{4}$$

其中 $IC_n^{ubm}(t)$ 为第 t 帧特征在 UBM 上打分的第 n 极大值对应的类号。

(3) 类转移概率矩阵由下式计算得出：

$$a_{ij}^{ubm}=\Pr\left[IC_n^{ubm}(t)=j|IC_n^{ubm}(t-1)=i\right]\triangleq \frac{\underset{1\le n\le N,1<t\le T}{\text{Count}}\left[IC_n^{ubm}(t-1),IC_n^{ubm}(t)\right]}{\underset{1\le n\le N,1<t\le T}{\text{Count}}\left[IC_n^{ubm}(t-1)\right]},\quad (1\le i,j\le K) \tag{5}$$

其中，$\text{Count}^{ubm}(i,j)$ 为从类 i 到类 j 的跳转次数，$\text{Count}^{ubm}(i)$ 为类 i 出现的次数。为了避免类转移概率矩阵中出现 0 值的情况，本文采用了一种简单的方法，即在初始化时，对每个 $\text{Count}^{ubm}(i,j)$ 赋值为 1，而 $\text{Count}^{ubm}(i)$ 赋值为 K。当然也可以采用其他更复杂的方法。

在训练阶段，假设说话人模型的训练特征序列为 $X=\{x_1,x_2,\cdots,x_T\}$。说话人模型通过 MAP 方法从 UBM上自适应得到，并被记为 $\lambda=\{g_m,\mu_m^{\lambda},\Sigma_m|m=1,2,\cdots,M\}$，其中高斯混合权重 g_m 和协方差矩阵 Σ_m 保持不变。通过相似的方法可以得到 $\{IG_n^{\lambda}(t)|t=1,2,\cdots,T\}$，$1\le n\le N$(训练特征序列在说话人模型 λ 上打分的最大 N 个值对应的高斯混合索引号序列)，和 $\{IC_n^{\lambda}(t)|t=1,2,\cdots,T\}$，$1\le n\le N$(训练特征序列在说话人模型 λ 上打分的最大 N 个值对应的类号序列)。说话人相关的类转移概率矩阵按下式得到：

$$a_{ij}^{\lambda}=\frac{\beta a_{ij}^{\lambda'}+a_{ij}^{ubm}}{\beta+1}, \tag{6}$$

$$a_{ij}^{\lambda'}=\Pr\left[IC_n^{\lambda}(t)=j|IC_n^{\lambda}(t-1)=i\right]\triangleq \frac{\underset{1\le n\le N,1<t\le T}{\text{Count}}\left[IC_n^{\lambda}(t-1),IC_n^{\lambda}(t)\right]}{\underset{1\le n\le N,1<t\le T}{\text{Count}}\left[IC_n^{\lambda}(t-1)\right]} \tag{7}$$

其中，β 是一个经验值，在实验中被设为 0.8。说话人相关的类权重 w_k^{λ} 保持不变。

识别阶段的前两步与训练阶段的前两步类似。在计算待测特征序列 $X=\{x_1,x_2,\cdots,x_T\}$ 与 UBM 或说话人模型 λ 的得分时，首先按照传统 GMM-UBM 方法对每帧进行打分，可以得到每帧的似然分序列 $p(x_t|ubm)$(背景模型的似然分序列) 和 $p(x_t|\lambda)(1\le t\le T)$(说话人模型 λ 的似然分序列) 及 N 个类序号的序列 $\{IC_n^{ubm}(t)|1\le t\le T\}$，$1\le n\le N$(训练特征序列在 UBM 上打分的最大 N 个值对应的类号序列) 和 $\{IC_n^{\lambda}(t)|1\le t\le T\}$，$1\le n\le N$(训练特征序列在说话人模型 λ 上打分的最大 N 个值对应的类号序列)。使用说话人相关类转移概率矩阵对说话人模型 λ 重新计算得分：

$$S_n(X|\lambda)=\frac{1}{T}\sum_{t=1}^{T}\log\left[p(x_t|\lambda)w_{IC_n(t)}^{\lambda}a_{IC_n(t-1),IC_n(t)}^{\lambda}\right]. \tag{8}$$

按照同样方法对UBM重新打分得到 $S_n(X|ubm)$ $(1\leq n\leq N)$, 其中使用的是通用类转移概率矩阵。最后，待测语音的得分由下式计算得到:

$$\mathrm{LLR}(X|\lambda)=\max_{1\leq n\leq N}[S_n(X|\lambda)-S_n(X|ubm)]. \quad (9)$$

本文将对高斯混合相关性建模的这种方法简称为 GMC 。

2 实验

实验中使用的数据库由 522 人 (347 男, 175 女) 在实验室环境下录制的语音组成, 其中语音文件的格式为 8 kHz, 16-bits 格式。每个说话人有三段语音, 其中一段用来训练说话人模型, 其余两段用于测试。训练语音的平均长度为 32 s, 而测试语音的平均长度为 8 s 。在此数据库的基础上，对每个测试语音文件加入高斯白噪音，其信噪比从 20 dB 到 0 dB, 间隔为 5 dB 。

实验中采用的语音帧长为24 ms, 帧移为12 ms。对每一帧语音数据, 其预加重系数为 0.97 。经过 Hamming 窗后，使用 256 点的 FFT 来计算每帧语音的频谱。实验使用的 Mel 滤波器组的个数为 30, 滤波器组的输出经过对数压缩和 DCT 之后，得到相应的 16 维特征系数，并提取相应的 16 维差分系数。所有特征经过倒谱均值减 (CMS) 后得到最后的语音特征序列。

实验中的 UBM 由 90 人 (45 男, 45 女) 的语音训练得到，其训练语音总长度大约为 2 个小时。UBM 和说话人模型都是由 $M=1,024$ 个高斯混合组成的，其中 UBM 被聚为 $K=32$ 个类。在训练与测试阶段，取最优的 4 个得分序列 $(N=4)$ 。

为了评测不同系统的鲁棒性 (参看表 1), 实验中使用 24s 的有效语音用来训练， 3s 的有效语音用于测试。其中用干净语音来训练说话人模型, 用不同信噪比下的语音分别进行比较测试。实验使用的基准系统 (Baseline) 为使用 16 维传统 MFCC 和 16 维差分系数的 GMM-UBM 说话人识别系统。表中每列的最大值用粗体给出，并给出 GMC+SMS 与 Baseline 的错误下降率 (Error Rate Reduction, ERR) 比较。关于 ERR 的定义如下:

$$\mathrm{ERR}=\frac{\mathrm{ER_{old}}-\mathrm{ER_{new}}}{\mathrm{ER_{old}}}\times 100\%, \quad (10)$$

其中， $\mathrm{ER_{old}}$ 为旧方法的错误率， $\mathrm{ER_{new}}$ 为新方法的错误率。

表 1 不同系统在不同测试语音信噪比下的识别率比较

系统 (%)	信噪比 (dB)						平均
	干净	20	15	10	5	0	
Baseline	93.9	88.1	78.9	48.5	31.4	22.8	60.6
SMSC	95.6	89.3	77.9	49.6	33.1	23.6	61.5
GMC	94.8	90.0	80.3	52.3	36.0	29.3	63.8
GMC+SMSC	**96.2**	**91.1**	**81.2**	**53.4**	**37.6**	**31.8**	**65.2**
ERR	37.7	25.2	10.9	9.5	9.0	11.7	11.7

3 结论

本文提出了两种噪音鲁棒性的方法, 一种是将子带 Mel 频谱质心 (SMSC) 和基于 MFCC 的语音前端处理相结合。另一种是通过类转移概率矩阵来表示隐藏在高斯混合相关性 (GMC) 中的高层说话人信息。实验表明，使用 SMSC 的系统与使用传统 MFCC 的说话人识别系统相比，平均错误率下降了 2.3% 。使用 GMC 的系统与使用传统 MFCC 的说话人识别系统相比，平均错误率下降 8.1% 。而采用 SMSC 和 GMC 的系统，平均错误率下降 11.7% 。实验表明，SMSC 和 GMC 都可以提高说话人识别系统的噪音鲁棒性。

由于 SMSC 和 GMC 的提取都不需要对背景噪音进行相应的估计或者建模，因而本文预想它们可以应用到各种噪音环境中。在高斯白噪音环境下的实验表明，这两种方法能够有效地提高说话人识别系统的鲁棒性。在其他噪音环境下，使用这两种方法的系统鲁棒性提高还需进一步的实验验证。

参 考 文 献

1 Rosenberg A E, Lee C H, Soong F K. Cepstral channel normalization techniques for HMM-based speaker verification. In: ICSLP, 1994: 1835—1838

2 Viikki O, Laurila K. Noise robust HMM-based speech recognition using Segmental Cepstral Feature vector normalization. In: ESCA NATO Workshop on robust speech recognition for unknown communication channels, Pont-a-Mousson, France, 1997: 107—101

3 Boll S F. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. on Acoustic. Speech, Signal Processing*, 1979; **27**(ASSP-33): 113—120

4 Hirsch H G, Ehrlicher C. Noise estimation techniques for robust speech recognition. In: ICASSP, 1995: 153—156

5 Bcrouti M, Schwartz R, Makhoul J. Enhancement of speech corrupted by additive noise. *Proceedings of the IEEE Conference on Acoustics. Speech. and Signal Processing*, 1979: 208—211

6 Lim J S, Oppenheim A V. All-pole modeling of degraded speech. *IEEE Trans. Acoust., Speech, Signal Processing*, 1978; **26**(3): 197—210

7 Moon S Y, Hwang J N. Noisy speech recognition via wavelet coefficient enhancement. In: Proc. IEEE 26th Asilomar Conf. Signals, Syst., Comput., Monterey, CA, 1992: 1086—1090

8 陈景车, 姚　磊, 黄泰翼. 几种高鲁棒性通信及说话人自适应语音识别算法研究. 声学学报, 1998; 23(6): 573—544

9 王成友, 汤叔祺, 梁甸农. 噪声对特征综合法语音识别性能的影响. 声学学报, 1997; **22**(3): 282—285

10 Sönmez M *et al.* Modeling dynamic prosodic variation for speaker verification. In: Proc. Int'l Conf. Spoken Language Processing, 1998; **7**: 3189—3192

11 Sönmez M K *et al.* A lognormal tied mixture model of pitch for prosody-based speaker recognition. In: Proc. Eurospeech, 1997; **3**: 1291—1394

12 Paliwal K K. Spectral subband centroid features for speech recognition. In: Proc. ICASSP, 1998; **2**: 617—620

13 Satoru Tsuge, Toshiaki Fukada, Harald Singer. Speaker normalized spectral subband parameters for noise robust speech recognition. In: Proc. ICASSP, 1999: 285—288

14 Bojana G, Paliwal K K. Robust feature extraction using subband spectral centroid histograms. In: Proc. ICASSP, 2001: 85—88

15 Reynolds D A. A Gaussian mixture modeling approach to Text-Independent Speaker Identification. Ph.D. thesis, Georgia Institute of Technology, 1992

16 Reynolds D A, Quatieri T F, Dunn R B, Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 2000; **10**: 19—41

17 Campbell J P, Reynolds D A, Dunn R B. Fusing high- and low-level features for speaker recognition. In: Eurospeech, ISCA, Geneva, Switzerland, 2003: 2665—2668

18 Weber F *et al.* Using prosodic and lexical information for speaker identification. In: ICASSP, 2002; **1**: 141—144

19 Andrews W *et al.* Gender-dependent phonetic refraction for speaker recognition. In: ICASSP, 2002; **1**: 149—153

20 Guo Q, Zheng F, Wu J *et al.* A new method used in HMM for modeling frame correlation. *International Conference on Acoustics, Speech and Signal Processing*, 1999: 169—172

21 Hu Zh-P, Satoshi I. Modeling improvement of the continuous hidden markov model for speech recognition. In: ICASSP, 1992: 373—376

22 Ostendorf M, Roukos S. A stochastic segment model for phoneme-based continuous speech recognition. *IEEE Trans. On Acoustics, Speech and Signal Processing*, 1989: 1857—1869

23 Reynolds D A. An overview of automatic speaker recognition technology. In: International Conference on Acoustics, Speech and Signal Processing, 2002: 4072—4075