几种改进的 MFCC 特征提取方法在说话人识别中的应用

许 鑫[1] 苏开娜[2] 胡起秀[3]

[1]（北京工业大学计算机学院，北京，100022）

[2]（北京工业大学计算机学院，北京，100022）

[3]（清华大学计算机科学与技术系，北京，100084）

摘 要：Mel 频率倒谱系数（MFCC）表征了人类的听觉特征。目前国内外提出了一些比较好的 MFCC 改进算法，可以提高语音特征提取的鲁棒性。本文介绍了一些在语音识别中取得一定效果的 Mel 倒谱提取的改进算法。将这些算法应用于文本无关的说话人识别，并在此基础上提出了四种改进方法。在 100 人和 200 人的电话语料库中，分别进行同信道和不同信道的实验，使识别率获得了不同程度的提高。尤其在不同信道上的识别效果更为显著。其中频率掩蔽滤波与 ExpoLog 尺度相结合的方法识别效果最好：在用座机语音建模手机语音测试的实验中，识别率从基准系统的 16.327%上升到 38.776%；在用手机语音建模座机语音测试的实验中，识别率从基准系统的 8%上升到 40%。可见，所提出的改进方法是非常有效的。

关键词：MFCC；说话人识别；特征提取；鲁棒性

1. 引言

研究人员发现在人类的听觉系统中存在着掩蔽效应。这种效应是指弱信号在强信号的附近会被掩蔽。同时人的耳蜗相当于一个非线性频率尺度的滤波器，使人耳对低频信号比对高频信号更敏感[1]。因此提出了利用 Mel 滤波器进行特征提取，它主要是模仿人的耳蜗对声音进行滤波，减小噪声对语音的影响。目前，MFCC 有较好的鲁棒性，所以在语音识别和说话人识别中得到了广泛的应用。但是由于噪声和信道的影响，在某些应用场合中仍然存在着不足。近年来，研究人员对此提出了改进 Mel 倒谱系数的计算方法，得到了一定的成效。本文介绍了国内外提出的针对语音识别的改进算法，将这些算法应用于说话人识别，并在此基础上提出了四种改进方法，使识别率获得了不同程度的提高。第 2 部分介绍传统的 MFCC 参数的提取算法，第 3 部分介绍国内外提出的改进的 MFCC 参数的提取算法，包括频率掩蔽滤波[2]、Mel 三角滤波器的能量加权滤波[3]、半升正弦函数倒谱提升[4]和 ExpoLog 尺度算法[5]，第 4 部分采用上面介绍的几种算法以及提出的四种改进方法进行实验分析，第 5 部分总结。

联系作者：许鑫，E-mail：xuxin@emails.bjut.edu.cn；苏开娜，E-mail：sukaina@bjut.edu.cn；胡起秀，E-mail：huqx8@sohu.com

2. 传统的 MFCC 参数的提取

MFCC 是着眼于人类的听觉机理而提出的。人耳对于不同频率信号可以引起不同的调节作用。在 MFCC 参数的提取过程中，Mel 三角滤波器就是模仿人耳的这种特性设计的。主要的算法流程如下[6]：对于某语音信号 $x(n)$

（1）预加重：目的在于对语音的高频部分进行加重，增加高频部分的分辨率。

$$x_n' = x_n - kx_{n-1} \quad k \in (0.9,1)$$ 公式（1）

（2）加窗：帧长为 N，一般采用的是汉明窗。目的在于进行短时信号的局部化分析，保持在窗边界处的信号可以平滑地衰减。

（3）离散傅里叶变换：

$$X_a(k) = \sum_{n=0}^{N-1} x(n) \mathrm{e}^{-\mathrm{j}2\pi k/N}$$ 公式（2）

其中，$x(n)$时输入的语音信号，N 是傅里叶变换的点数。

（4）用 Q 个带通三角滤波器，中心频率从 0~采样频率/2 间 Mel 频率分布，中心频率为 $f(q)$，$q=1,2,\cdots,Q$，三角滤波器设计如公式（3）所示。

$$H_q(k) = \begin{cases} 0 & k < f(q-1) \quad 或 \quad k > f(q+1) \\ \dfrac{k - f(q-1)}{f(q) - f(q-1)} & f(q-1) \leqslant k \leqslant f(q) \\ \dfrac{f(q+1) - k}{f(q+1) - f(q)} & f(q) < k \leqslant f(q+1) \end{cases}$$ 公式（3）

（5）计算每个滤波器组的输出能量，公式（4）表示第 q 个三角滤波器的输出能量。

$$S(q) = \ln\left(\sum_{k=0}^{N-1} \left| X_a^{(k)} \right|^2 H_q^{(k)}\right) = \ln(e(q)) \quad 0 \leqslant q < Q$$ 公式（4）

（6）利用离散余弦变换计算 MFCC 系数，第 n 个系数的计算如公式（5）所示。

$$C(n) = \sum_{q=0}^{Q-1} S(q) \cos(\pi n(q+0.5)/Q) \quad 0 \leqslant n < L$$ 公式（5）

其中，L 为 MFCC 系数的阶数。

3. 改进的 MFCC 特征提取算法

3.1　频率掩蔽滤波[2]

前面提到的掩蔽效应，实质上就是增强主要的信号抑制较弱的噪音信号。Weizhong Zhu 和 Douglas O'Shaughnessy[2]提出的频率掩蔽滤波（FMF，frequency masking filtering）算法，是用一个非线性的双向滤波器来模仿人耳的掩蔽机能，得到更加鲁棒的特征。FMF 由下面的方程具体给出。在使用公式（6）时，应初始化为 $y_n=x_n$；在使用公式（7）时，应初始化为 $x_0=y_0$。FMF 是在传统 MFCC 参数提取的第 3、4 步骤之间进行的。

$$y'_{i-1} = \alpha y_i$$
$$y_{i-1} = y'_{i-1} \quad \text{if } y'_{i-1} > x_{i-1} \qquad \text{公式（6）}$$
$$y_{i-1} = x_{i-1} \quad \text{if } y'_{i-1} \leqslant x_{i-1}$$

$$y'_{n\cdots} = \beta y_{n-1}$$
$$y_{n\cdots} = y'_n \quad \text{if } y'_{n\cdots} > x_n \qquad \text{公式（7）}$$
$$y_{n\cdots} = x_n \quad \text{if } y'_{n\cdots} \leqslant x_n$$

其中，α 为低频掩蔽阈值，β 为高频掩蔽阈值，x_i 为频率 i 的原始功率谱，y_i 为过滤频谱后的输出。

3.2 Mel 滤波器组加权分析[3]

Wei-Wen Hung 和 Hsiao-Chuan Wang[3]提出了 Mel 滤波器组加权分析（WFBA，weighted filter bank analysis）。主要通过提高对数滤波能量中高能量部分的权重及削弱低能量部分的权重，来提高 MFCC 的区分能力，这样可以使语音对环境的影响不敏感，具有较好的鲁棒性。根据公式（9）的滤波器加权分析因子，来计算公式（8）的倒谱系数。

$$C(n) = \sum_{q=0}^{Q-1} w(q) S(q) \cos(\pi n(q+0.5)/Q) \quad 0 \leqslant n < L \qquad \text{公式（8）}$$

$$w(q) = \log[e(q)+1] / \sum_{j=1}^{Q} \log[e(j)+1] \quad 0 \leqslant q < Q \qquad \text{公式（9）}$$

3.3 半升正弦函数倒谱提升[4]

通过大量实验发现，特征向量的各个分量对于识别率的贡献是不同的。在说话人识别中，高阶 MFCC 分量较低阶 MFCC 分量来讲，不易受到噪声的影响，具有很好的鲁棒性[7]。所以采用半升正弦函数（HRSF，half raised-sine function）进行倒谱提升，可以降低易受噪声干扰的低阶分量值，同时提高了数值相对小的中高阶分量值。权重公式（10）的前一部分 0.5 是为了保证倒谱分量不完全衰减，后一部分就是对高低阶分量进行了不同的加权。

$$r_i = 0.5 + 0.5\sin(\pi i/L) \qquad \text{公式（10）}$$

其中，i=0, 1, …, L–1；L 为特征阶数。

3.4 ExpoLog 尺度算法[5]

S Bou-Ghazale 和 J H L Hansen[5]提出了 ExpoLog 尺度算法（EFS，expolog frequency scale）。他们通过实验发现，在语音变异的情况下（包括心理紧张、情绪不稳定等），中频率段在语音识别中的识别率要高于低、高频率段识别率，也就是说中频率段的语音较为鲁棒。所以应该对中频率段进行加重，同时削弱其他频率段的权重。公式（11）为原始 Mel 尺度函数，公式（12）为 EFS 算法。观察两个公式，发现 EFS 对于 2000Hz 以上的高频率段没有变化，而低频率段在被相对削弱的同时，提高了中频率段的影响。

$$\text{Mel- scale} = 2595 \times \log(1 + f/700) \qquad \text{公式（11）}$$

$$\text{ExpoLog} = 700 \times (10^{f/3988} - 1) \quad \text{if } 0 \leqslant f \leqslant 2000\,\text{Hz}$$
$$\text{ExpoLog} = 2595 \times \log(1 + f/700) \quad \text{if } 2000 < f \leqslant 4000\,\text{Hz} \qquad \text{公式（12）}$$

4. 实验及结果分析

由于上面介绍的算法各有特点和优势，FMF 模拟了人的听觉掩蔽效应，WFBA 减少了易受噪声影响的谱谷部分，HRSF 提高了较鲁棒的倒谱系数中间阶的分量值，EFS 加大了识别率相对较高的中频率段的影响范围，考虑到优势互补，因此我们提出将 WFBA 与 HRSF 相结合、FMF（插补）与 EFS 相结合、EFS 与 HRSF 相结合、FMF（插补）与 HRSF 相结合的四种方法，来改进 MFCC 的特征提取过程。

我们采用 200 人的语料库，包括四组电话数据（信噪比约为 20dB），座机集合 *a*、座机集合 *b*、手机集合 *a* 以及手机集合 *b*。每个说话人都分别存在于四组数据中。我们使用座机集合 *a* 和手机集合 *a* 作为训练样本（训练音长为 24 秒），座机集合 *b* 和手机集合 *b* 作为测试样本（测试音长为 10s）。分别进行同信道的座机集合间、手机集合间及不同信道的座机和手机集合间的测试。所有语音样本采样率均为 8kHz，采用线性 PCM16bit 编码。特征参数采用 MFCC 系数的 *C*1~*C*16 的 16 阶系数。预加重系数为 0.95，窗长为 256 个点，窗移为 128 个点，使用 24 个三角滤波器进行 Mel 尺度的滤波。

我们分别对 100 人（参见表 1）和 200 人（参见表 2）集合进行训练，采用 50 人进行测试。测试样本是从 200 人语料库中随机抽取 50 人，每个说话人的语音样本均存在于不同规模训练集合中。

每个测试均在基于 VQ 的基准系统上进行如下实验：

（1）加入 FMF（阈值α=0.5、β=0.8）；

（2）加入 FMF（采用线性插补方法，阈值范围$\alpha \in [0.3,0.5]$ $\beta \in [0.6,0.8]$）；

（3）加入 WFBA；

（4）加入 HRSF；

（5）将原 Mel 尺度修改为 EFS；

（6）加入 WFBA 和 HRSF；

（7）加入 FMF（线性插补阈值范围$\alpha \in [0.3,0.5]$ $\beta \in [0.6,0.8]$）并修改为 EFS；

（8）加入 HRSF 并修改为 EFS；

（9）加入 FMF（线性插补阈值范围$\alpha \in [0.3,0.5]$ $\beta \in [0.6,0.8]$）和 HRSF。

表 1　采用上述几种方法的 100 人训练集合的识别率　%

100 人训练 50 人测试	同信道		不同信道	
	dha-dhb	sja-sjb	dha-sjb	sja-dhb
基准系统	92.000	97.959	16.327	16.000
FMF(α=0.5，β=0.8)	84.000	81.633	20.408	36.000
FMF（插补）	88.000	93.878	26.531	28.000
WFBA	94.000	95.918	18.367	22.000
HRSF	94.000	95.918	18.367	22.000
EFS	96.000	95.918	32.653	32.000
WFBA+ HRSF	88.000	95.918	22.449	20.000
FMF（插补）+ EFS	96.000	95.918	38.776	44.000
EFS + HRSF	96.000	97.959	38.776	32.000
FMF（插补）+ HRSF	92.000	95.918	24.490	40.000

表 2 采用上述几种方法的 200 人训练集合的识别率 %

200 人训练 50 人测试	同信道		不同信道	
	dha-dhb	sja-sjb	dha-sjb	sja-dhb
基准系统	92.000	97.959	16.327	8.000
FMF(α=0.5, β=0.8)	80.000	79.592	16.327	28.000
FMF（插补）	88.000	89.796	20.408	16.000
WFBA	94.000	95.918	18.367	12.000
HRSF	94.000	95.918	16.327	12.000
EFS	96.000	95.918	30.612	28.000
WFBA+ HRSF	88.000	95.918	20.408	14.000
FMF（插补）+ EFS	96.000	95.918	38.776	40.000
EFS + HRSF	96.000	97.959	32.653	28.000
FMF（插补）+ HRSF	90.000	93.878	20.408	26.000

注：dha—座机集合 a；dhb—座机集合 b；sja—手机集合 a；sjb—手机集合 b；dha-sjb—用座机集合 a 进行训练，用座机集合 b 进行测试。

结果分析：

（1）在 100 人训练 50 人测试的实验中，同信道的识别率有增有减，不同信道的识别率均在增加。所提出的四种改进方法识别效果都比较好，尤其对于不同信道表现很鲁棒。EFS、FMF（插补）与 EFS 相结合、EFS 与 HRSF、FMF（插补）与 HRSF 相结合这四种方法整体的识别性能提高较显著。

（2）在 200 人训练 50 人测试的实验中，EFS 尺度、FMF（插补）与 EFS 相结合、EFS 与 HRSF 的识别效果非常好，在保持同信道高识别率的同时，对于不同信道识别率的提高非常显著。在 dha-sjb 的测试中，FMF（插补）与 EFS 相结合的方法使识别率从 16.327% 上升到 38.776%；在 sja-dhb 的测试中，使识别率从 8%提高到 40%。

（3）通过不同规模训练集的实验，我们发现随着训练集合规模增大，识别率有所下降。但是所提出的四种改进方法对于不同信道的贡献是非常明显的。尤其 EFS 与 HRSF 相结合的方法和 FMF（插补）与 EFS 相结合的方法整体识别效果突出，虽然后者的识别效果最佳，前者次之，但是前者相对于基准系统增加的计算量是非常小的。

5. 结论

本文介绍了几种国内外提出的针对语音识别的 MFCC 的改进算法，将这些算法应用于说话人识别，并对所提出的 WFBA 与 HRSF 相结合、FMF（插补）与 EFS 相结合、EFS 与 HRSF 相结合、FMF（插补）与 HRSF 相结合的四种方法分别进行了实验。实验发现，这四种方法在不同规模训练集的不同信道的测试中，识别率提高很显著。尤其 FMF（插补）与 EFS 相结合、EFS 与 HRSF 相结合这两种方法对同信道和不同信道具有很好的鲁棒性。由于 FMF（插补）与 EFS 相结合的方法更符合人的听觉机理，所以识别效果最佳，但是计算开销稍大于 EFS 与 HRSF 相结合的方法。所以根据实际情况，权衡识别速度与准确度，选择适当的方法应用于说话人识别系统。

通过实验发现，FMF 算法只对于不同信道的识别有一定提高，其中α、β值的设定起着举足轻重的作用，所以今后可以进一步研究不同的α、β值对说话人识别率的影响。虽然 EFS 算法表现了较好的识别效果，但是相信通过不同频率对说话人识别的相对重要性的研究，来设计和改进更符合说话人特征的 Mel 尺度函数，将会获得更加鲁棒的 MFCC 特征参数。

参考文献

[1] 章熙春等. 语音MFCC特征计算的改进方法. 数据采集与处理, 2005, 20(2): 161~165

[2] Weizhong Zhu, Douglas O'Shaughnessy. Incorporating Frequency Masking Filtering in a Standard MFCC Feature Extraction Algorithm. Proc. 7th International Conference on Signal Processing, ICSP 2004, Aug. 31-Sept. 4 2004, Beijing, China: 617~620

[3] Wei-Wen Hung, Hsiao-Chuan Wang. On the Use of Weighted Filter Bank Analysis for the Derivation of Robust MFCCs. Signal Processing Letters, IEEE, 2001, 8(3): 70~73

[4] 马志友等. 二次特征提取及其在说话人识别中的应用. 电路与系统学报, 2005, 8(4): 130~133

[5] S Bou-Ghazale, J H L Hansen. A Comparative Study of Traditional and Newly Proposed Features for Recognition of Speech Under Stress. IEEE Trans Speech and Audio Processing, 2000, 8(4): 429~442

[6] 王让定等. 语音倒谱特征的研究. 计算机工程, 2003, 29(13): 31~33

[7] 甄斌等. 语音识别和说话人识别中各倒谱分量的相对重要性. 北京大学学报（自然科学版）, 2001, 37(3): 371~378

A Comparative Study of Some Improved MFCC Algorithms for Speaker Recognition

Xu Xin[1+], Su Kai-na[2], Hu Qi-xiu[3]

[1]（College of Computer, Beijing University of Technology, 100022, China）

[2]（College of Computer, Beijing University of Technology, 100022, China）

[3]（Department of Computer Science and Technology, Tsinghua University, 100084, China）

[+]Corresponding author: Phn: +86-10-6279-7001(804), E-mail: xuxin@emails.bjut.edu.cn

Key words: MFCC, speaker recognition, feature extraction, robust

Abstract: MFCC symbolizes the property of human auditory system, and it is the key feature parameter in speaker recognition and speech recognition. The researchers proposed some improved algorithms for MFCC feature extraction, which succeeded in speech recognition in some cases. Those algorithms that we introduce to text-independent speaker recognition are Frequency Masking Filtering (FMF), Weighted Filter Bank Analysis (WFBA), Half Raised-Sine Function (HRSF) and ExpoLog Frequency Scale (EFS). Due to their advantages, we consider combining these algorithms and proposing four combined methods, including WFBA and HRSF, FMF and EFS, EFS and HRFS, FMF and HRFS. Combined WFBA and HRSF could decrease

the influence by noise and emphasize the important middle MFCC terms; combined FMF and EFS could make MFCC more suitable to human auditory mechanism; combined EFS and HRFS could emphasize the more important mid-frequencies and lifter the more useful coefficients; combined FMF and HRFS could mimic a human masking mechanism to get more robust features. The speaker recognition system is based on Vector Quantization models. The speech database used in these experiments is telephone speech. These experiments are carried out between the same types or different types of handset. We train 100 people models and 200 people models respectively, and use speech from 50 people to test. With the proposed methods, the experiments reveal high robustness, especially in the different types of handset. In four proposed methods, combined FMF (linear interpolation) and EFS, combined EFS and HRSF, show higher robust than any other in both the same and different types of handset. In the different types of handset tests, combined FMF (linear interpolation) and EFS gets the correct recognition rate 38.776% and 40% respectively compared with the result of 16.327% and 8% in the baseline system. Although combined FMF (linear interpolation) and EFS gets the best result, it should require more extra computation than combined EFS and HRSF. By making a tradeoff between recognition speed and correct recognition rate with our needs, we can choose the right method for speaker recognition system.