Mel 刻度上非均匀分布滤波器组在 MFCC 参数提取中的应用

温源 侯震 李明 王之禹 俞铁城

中科院声学所 5 室
wenyuan@mail.ioa.ac.cn

摘要

本文阐述了一种在 mel 刻度上应用非均匀分布滤波器组来求 MFCC 参数的方法。语音的频率分布在 mel 刻度上并非是均匀分布的：本文以大量统计的方法对一定的数据集求出语音各频率的幅度分布，并以此来确定新的三角滤波器组的分布。最后，本文通过对比识别试验来证明新方法提参数的有效性。

1. 引言

语音识别系统大致可以看成由三个部分组成：参数提取，训练和识别。其中在参数的提取，最重要的是要在声学层面上考虑特征对于语音描述的忠实程度和鲁棒性，最大程度的挖掘语音信息，并把他们体现到特征中去。这就要求从频谱分析到数学变换以及参数的最终确定都要考虑到语音的声学特点。

MFCC 作为一种具有鲁棒性且较能忠实的反映语音特性的参数，在语音识别中得到了广泛的应用。传统的计算的方法是基于 mel 刻度上均匀分布的若干三角滤波器来做频谱分析的。一般来说这些滤波器将覆盖 8000Hz 以下的语音频段。计算时首先要把语音的频谱变换到 mel 刻度域，利用三角滤波器来求通带输出，然后作 DCT 变换解相关，来求得 MFCC 参数。通带的分布主要考虑的是 mel 刻度对人耳听觉特性的描述，这种描述是针对所有可听声的，包括音乐、噪声、语音等等，它并不只针对人类语音。而事实上，尽管在求 MFCC 参数时考虑了人声主要集中于 8000Hz 以下这一特点，但语音在这一频带上并非均与分布，于是当我们在 mel 刻度上使用均匀分布的滤波器来求通带输出时，其长时统计结果必然是各通带输出不均等。而由信息论可知，只有当各个通道的输出统计均等时，它们所能表达的信息量最大。因此，我们考虑有必要在 mel 刻度上调整滤波器的相对位置来满足通道设计的准则——各通道统计输出均等。于是依据这一准则，本文讨论了如何通过先统计各频率语音在 mel 刻度上的幅值分布，然后由此推导滤波器组的分布。

频谱弯折在说话人自适应，声道归一化中也有应用，在[1][2]中，声道归一化是从声道因子的最大似然估计入手的；而本文直接从统计语料的频率分布入手，且本文的目的不是论述一种说话人自适应的方法，而是提出一种特征提取的改进方案。

本文仍然采用三角滤波器组，与传统方法相比，只对各个滤波器的相对位置进行调整，滤波器组所覆盖频带范围、DCT 变换，以及之后的训练和识别不作任何改动。本文最后通过对比试验反映两种 MFCC 参数提取对识别的影响。

2. 滤波器组的确定

采用三角滤波器的一个好处就是滤波器组可以由三角滤波器的顶点来方便的确定。本文将依据在 mel 刻度上通道输出应均等这一准则来定制通道位置，从而确定滤波器顶点。

2.1. 语音频率分布的统计方法

语音的频率分布是通过大量统计实现的，本文对一个 300 人的数据库进行统计。首先对于所有语料进行严格的端点检测，目的是只针对语音部分进行频率分布的统计。我们利用 FFT 作帧长 512 点、帧移 256 点的逐帧频谱分析。设第 i 帧信号经过高频预加重和汉明窗为 $s_i(n)$、其帧幅度谱为 $S_i(k)$，则：

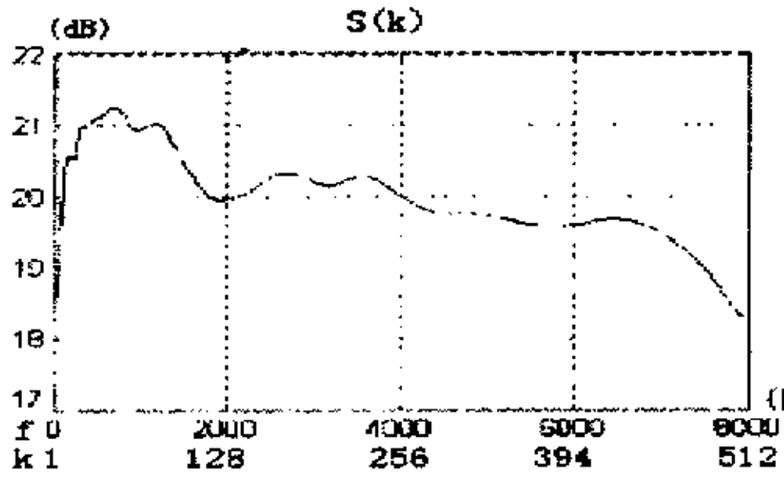
$$S_i(k) = |FFT[s_i(n)]| \qquad (1)$$

其中，k 为 FFT 序号、$1 \le k \le 512$；n 为帧信号 $s_i(n)$ 的点索引，显然 $1 \le n \le 512$；$|\cdot|$ 表示求模。

设数据库共有 N 帧语音信号参与统计，则统计幅度谱为：

$$S(k) = 20\log_{10}\sum_{i=1}^{N} S_i(k) \qquad (2)$$

下图一为统计得到的幅度谱。



图一 频域统计幅度谱

在上图中横坐标提供了频率以及 FFT 序号两种作标，而需注意的是纵坐标为 dB 刻度。

由于数据库语音的采样率为 $f_s = 16000Hz$，所以 512 点 FFT 的频谱精度为 $F = 8000/512Hz$；FFT 序号 1～512 线性对应频率范围 0～8000Hz，所以：

$$f_k = k \cdot F \qquad (3)$$

接下来需要将 FFT 序号 k 折射到 mel 刻度上。由[3]中的 mel 刻度变换公式：

$$M(f) = 2595.0 \cdot \log_{10}(1 + f/700) \quad (4)$$

所以序号 k 对应在 mel 刻度上的值为：

$$m_k = 2595.0 \cdot \log_{10}(1 + k \cdot F/700) \quad (5)$$

由上式(3)可将统计幅度谱 S(k)映射到 mel 刻度域，用 E(m)表示 mel 刻度上的幅度谱，则：

$$E(m_k) = S(k)$$

由于 S(k)是由 512 个离散点表示的，所以变换后的 E(m)，我们也只知道 512 个离散点 $E(m_k)$。为了对通道的位置能够较能准确的定位，我们对相邻点[$E(m_k)$，$E(m_{k-1})$]之间的曲线采用斜边梯形拟合。即：

则于，m $m_{k-1} \le m \le m_k$，有：

$$E(m) = E(m_k) + \frac{E(m_{k+1}) - E(m_k)}{m_{k+1} - m_k} \cdot (m - m_k) \qquad (6)$$

于是由上式可以得到 mel 刻度上语音统计的幅度分布图，如下：

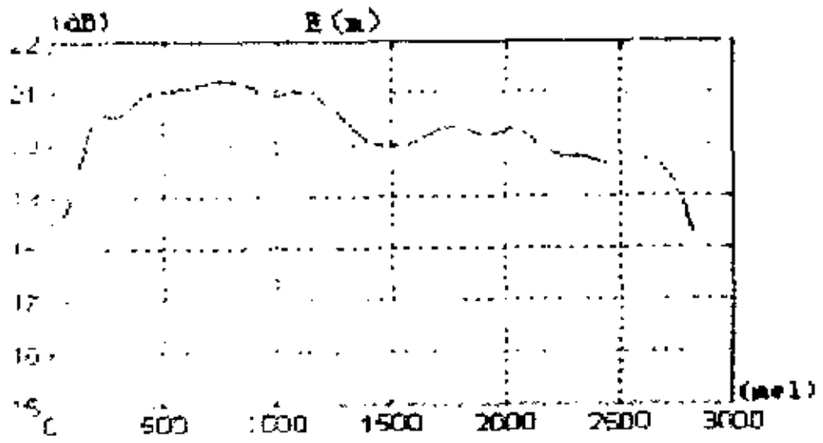


图 mel 刻度上语音统计的幅度分布图

2.2. 三角滤波器组的确定

假设需要确定 N 个三角滤波器的顶点位置，我们的目的是希望 N 个滤波器的统计输出均等。这里我们通过确定 $N+1$ 个通道来实现。通道的确定依据通道输出统计均等这一原则。也就是说我们希望在 0～8000Hz 对应的 mel 刻度上找出 N 个点，使得这些点所确定的通道将幅度统计图 E(m)曲线所包括的面积等分，也就是希望在 mel 刻度上确定 N 个点 $m_1, m_2, \ldots, m_N$，使得它们确定的 $N+1$ 个曲边梯形面积相等，即：

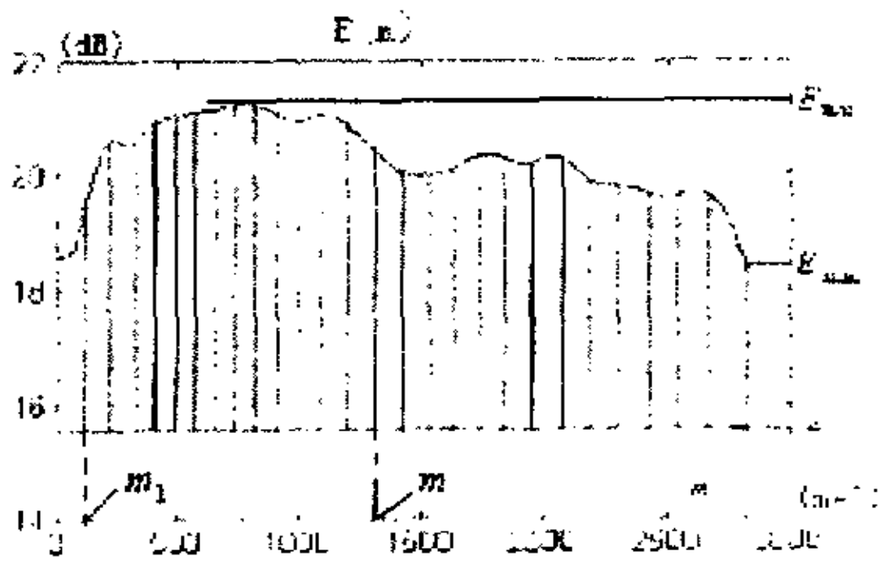
$$\frac{1}{N+1}\int_0^{m_{N+1}} E(m)dm = \int_{m_{i-1}}^{m_i} E(m)dm \qquad (7)$$

其中，$m_{N+1} = M(8000)$（参看(4)式）

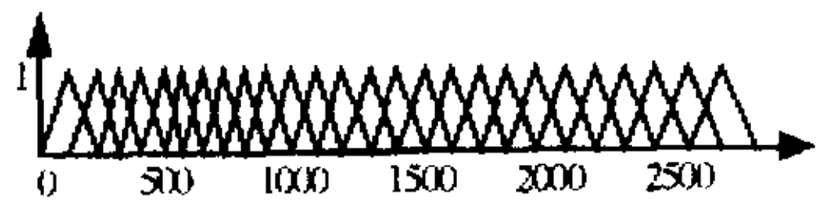
事实上由式（7）所确定的滤波器顶点是唯一的，也就是说它确定唯一的滤波器的疏密程度。但是在具体应用时，希望能加强或者减小滤波器的相对疏密程度，为此，可以引入参数 ε 来实现这一目的：使以 ε 为基准的 $N+1$ 个曲边梯形面积相等，如下式：

$$\frac{1}{N+1}\int_0^{m_{N+1}} |E(m) - \varepsilon| dm = \int_{m_{i-1}}^{m_i} |E(m) - \varepsilon| dm \qquad (8)$$

由（8）式，依次取 i=1,2,....N 即可确定 N 个分界点如图三所示。于是以 $m_1, m_2, \ldots, m_N$ 为三角滤波器的顶点，则可以确定滤波器组的分布：如图四所示



图三 确定通道



图四 非均匀分布滤波器组

对比图三、图四可以看到，在语音统计幅度较大的频段上对应三角滤波器分布也较密集；而幅度小的频带上，滤波器分布较稀疏。同时可以注意到，中频部分幅度较大，对应滤波器分布也较密。我们可以预测，这个频段上的语音将会得到相对较细致的分析。

ε 的选择和数据库语音以及所选的其它参数，如通道数有关。它的基本作用是控制滤波器的疏密程度，ε 可以由下式来确定：

$$\varepsilon = E_{min} - \theta \cdot (E_{max} - E_{min}) \qquad (9)$$

其中：$E_{min} = \min[E(m)]$ $0 < m < m_{N+1}$

$E_{max} = \max[E(m)]$ $0 < m < m_{N+1}$

$\theta \in [0.75, 2]$ 本文 $\theta = 1.25$

θ 越大则 ε 越小，滤波器之间分布的疏密差异也就越大，在 $\theta = 1$ 时，滤波器组的密集程度在 E_{max} 的位置上大约为在 E_{min} 位置上的两倍。

此外，我们还可以考察极限情况。变换（8）式，则：

$$\frac{1}{N+1}\int_{0}^{m_{N+1}} E(m)dm - \varepsilon \cdot \frac{m_{N+1}}{N+1}$$

$$= \int_{m_{i-1}}^{m_i} E(m)dm - \varepsilon \cdot (m_i - m_{i-1})$$

显然，当 $\varepsilon \to -\infty$ 时，$m_i - m_{i-1} \to \frac{m_{N+1}}{N+1}$；这相当于 N 个顶点在 $0 \sim m_{N+1}$ 之间均匀分布，也就是传统的三角滤波器组的分布方式。

2.3. MFCC 参数的计算

由获得的三角滤波器组被确定下来，MFCC 参数的计算步骤完全同于传统的计算方法[4]：

（1） 对帧语音作预加重；
（2） 加 hamming 窗；
（3） 求 FFT 变换；
（4） 求三角滤波器输出；
（5） 作 DCT 变换
（6） 倒谱加权

3. 试验

本文通过对比试验来考察均匀和非均匀分布滤波器组对识别的影响。试验数据采用一个 180 人的数据库，每人 200 句数字串语音，每句字长不同，平均字长为 4。用 130 人训练，50 人识别。试验系统采用 HMM 连续语音识别。26 维特征分三组，分别为：MFCC，ΔMFCC，以及 E 和 ΔE。试验共分三组对比试验，每组依次设置通道数为 20，26 以及 30，从而考察在三角滤波器组数不同的情况下，用均匀和非均匀分布滤波器组提取参数对识别的影响。

	20 维通道	26 维通道	30 维通道
均匀分布识别率	94．76%	95．12%	95．23%
非均匀分布识别率	95．47%	95．60%	95．63%
错误率下降	13．5%	9．8%	8．4%

表一 对比 试验结果

4. 讨论

由本文的三组对比试验可以看到，在求 MFCC 参数时改变三角滤波器组的疏密程度，使之成为非均匀分布可以改善识别率。对于 20，26 和 30 几种常用的通道数分别获得了 13．5%，9．8%和 8．4%的错误率下降。同时我们也注意到，对于较少的通道数，系统改善要较通道数较多的明显。

5. 总结

本文讨论了一种 MFCC 参数提取改进方案。此方案利用统计得到的语音的频率分布特性，以及通道输出应统计均等这一准则，引入了滤波器组的非均匀分布的计算方法；并用来提取新的 MFCC 参数。识别试验证明可以获得大约 10%的错误率下降。

6. 参考文献

[1] Li L. Richard Speaker normalization using efficient warping procedures Proc. ICASSP, 1996. 353-356

[2] Lincoln M, Cox S, Ringland. A fast method of speaker normalization using formant estimation. Proc. Eurospeech. 1997.2095-2098

[3] J. Picone, "Signal Modeling Techniques in Speech Recognition,", IEEE Proceedings, vol. 81, no. 9, pp. 1215-1247, Sept. 1993.

[4] Davis, S.B. and Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoustics, Speech and Signal Proc., 28, p357 366. 1980.

Mel刻度上非均匀分布滤波器组在MFCC参数提取中的应用

作者： 温源， 侯霞， 李明， 王之禹， 俞铁城
作者单位： 中科院声学所5室