第三届全国计算机智能接口与智能应用学术会议（NCCIIIA），1997 年 8 月 11-16 日, 93-97，张家界

基于音调的特征提取在非特定人语音识别中的运用*

武　健，郑　方, 吴文虎, 方棣棠
（清华大学 计算机科学与技术系，北京 100084）
[jwu, fzheng]@sp.cs.tsinghua.edu.cn, fzheng@cenpok.net

摘要：本文介绍了一种新的特征提取方法，该方法利用了人类听觉对语音音高的感受，依据实验心理学的实验结果，对短时功率谱做频率弯折处理，使之更加符合人的听觉习惯，有助于提高非特定人语音识别的性能。文中详细解释了这种新方法的理论基础及实现方案，并与传统的线性预测技术作了比较，实验结果表明这种方法比较有效。

关键词：线性预测编码，短时功率谱，音调的区分域限，非特定人语音识别。

一、引言

特征提取是语音识别的一个关键步骤。如果提出的特征参数能够比较好地刻划语音的本质特征，那么在后续处理中就有可能取得理想的效果。目前较为常用的特征提取方法大都是基于声道的全极点模型的，最典型的就是线性预测编码(LPC)技术。

通常使用的线性预测技术实际上是使自回归全极点模型 A(ω)在各个频带上均等地逼近语音短时功率谱 P(ω)，然而人在辩听某个音时常常对各个不同的频率(或者说是音调)段有不同的反应。

实际上早已有许多人做过这方面的研究，如 Hermansky[1] 在研究语音的感知线性预测分析 (PLP 技术) 时就考虑用临界带宽 (Critical Band) 来对频谱 P(ω) 沿频率轴作一定的弯折，从而达到在不同的频段上以不同的精度逼近 P(ω)的目的，在此基础上再利用等响度曲线的原理进行处理，这个实验取得了很好的效果。

我们这里介绍的是另外一种利用音调特性来进行特征提取的方案，即用 MEL 刻度对频率轴进行弯折。

二、基本原理

语音学中为了描述听觉上分辨声音高低的感受，引入了音调这个概念。根据实验心理学的实验结果，心理上的主观音高主要与声音刺激的频率大小有关[2]，我们根据实验研究，规定音调的单位为美(Mel)，确定 1000 美的音调为 1000Hz 60dB 的声音刺激的主观感觉。于是，在声强不变的前提下便可以对不同频率的声音刺激进行音调判断，并探求出音调和频率间的关系。Stevens 和 Walkman 通过实验建立了一张音调－频率表[3](如图 1)，形象地阐述了在强度不变的情况下，

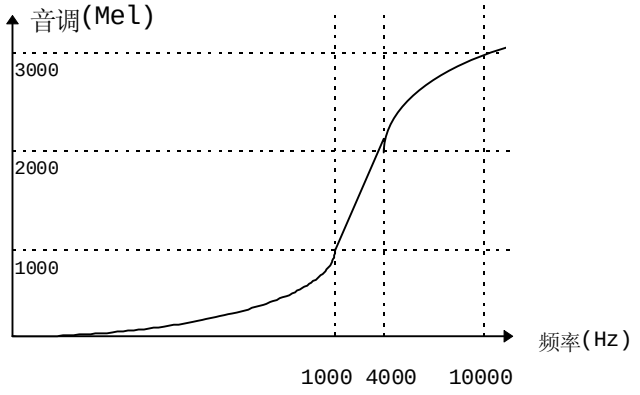

图 1　音调－频率表

* 本课题得到 863-306 资助，合同号为 863-306-03-02-4。

音调和频率之间的关系。

从图上可以看出两者之间并不是简单的线性关系。在 1,000~4,000Hz 范围内，两者之间基本上是线性相关的；当频率大于 4,000Hz 时，基本上呈对数关系；而当频率小于 1,000Hz 时，则近似于指数关系。

另外，实验心理学的实验结果表明：一般来说，音调的区分阈限 (DL，Distinct Limit) 随音调的频率的变化而变化。频率越低，人耳对频率的变化越敏感，即 DL 值较小。在强度为 40dB 时，2,000Hz 的音调只要改变 3Hz 即可被察觉；而当音调频率达到 10,000Hz 时，DL 值已上升到 30Hz。实验表明，只要音调频率高于 1,000Hz，能察觉到的频率差异所需频率变化是相对恒定的，大约是 0.3％。

这个结果说明：我们可以设计模型使其在音调轴 (Mel) 上均等地逼近语音短时功率谱，从而更好地符合人的听觉特性。

假设人耳的听觉范围是从 f_0 到 f_n ，相对应的音调从 M_0 到 M_n 。对 M_0 到 M_n 进行 n 等分，分点为 M_1， M_2，…， M_{n-2}， M_{n-1}；相对应的频率为 f_1， f_2，…， f_{n-2}， f_{n-1}。这样，所获得的 n 个频段实际上就是根据音调的特性，将频率轴作了合理的弯折。在我们现在所划分的频率轴上，每一个频段对即将提取的特征参数都具有相同的重要性。

根据以上原理，我们可以采用以下的设计来实现特征参数的提取。

三、特征提取的实现过程

1. 特征提取的过程简述 (见图 2)

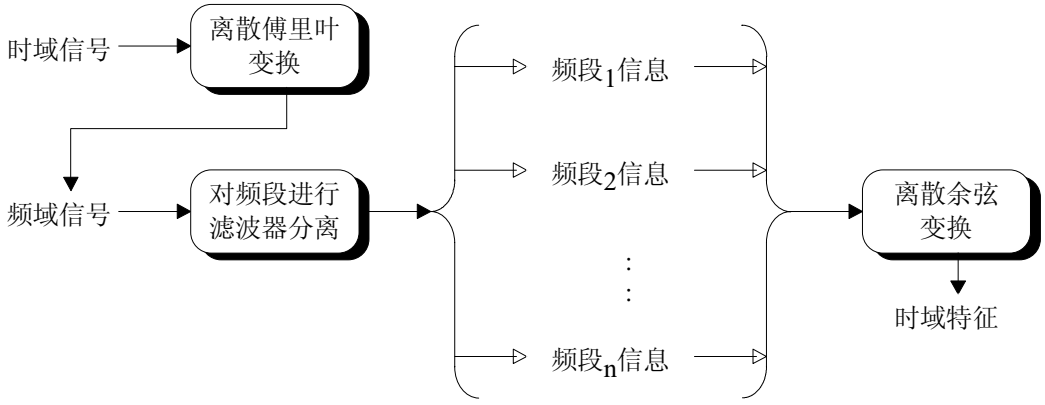

图 2　特征提取框图

2. 特征提取过程的具体设计

(1) 离散傅里叶变换 (DFT)

时域中的语音信号首先要经过 Hamming 窗加权，加权公式是：

$$H(n) = 0.54 + 0.46\cos\left(2\pi n / (N-1)\right) \qquad (1)$$

其中 N 是窗长。

通过 DFT 之后，再利用下式获取短时功率谱：

$$P(\omega) = \left[\left(\mathrm{Re}\,S(\omega)\right)^2 + \left(\mathrm{Im}\,S(\omega)\right)^2\right]^{1/2} \qquad (2)$$

(2) 对已划分好的频段进行滤波器分离

根据前述的音调-频率曲线，可以用如下的近似公式将短时功率谱 P(ω)沿频率轴弯折成 Q(m)，其中频率ω到 MEL 刻度 m（单位为美）的弯折函数关系为：

$$m = 3298.5 \cdot \ln\left(1 + \frac{\omega}{700}\right) \qquad (3)$$

弯折之后就可以根据原理部分的分析，利用 Mel 刻度对 Q(m)设计出一组滤波器来进行频带信息的分离。

另外，滤波器的设计可以是各种类型，最简单的设计当然是矩形滤波器 (**SF,** Square Filter)(如图 3)。它实际上是对各个频率上的分量同等对待。

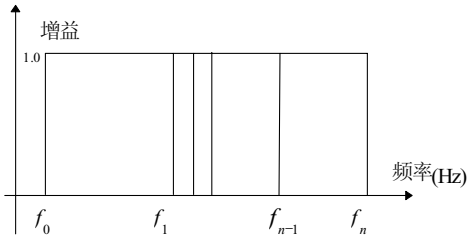

图 3 矩形滤波器（**SF**）

第二种滤波器的设计方法是三角型滤波器 (**TF,** Triangle Filter) (如图 4) ，这种设计考虑到同一频带内不同频率分量贡献不同，分别乘以不同的权重，但在频带分界点附近处的频谱信息并未得到充分利用。

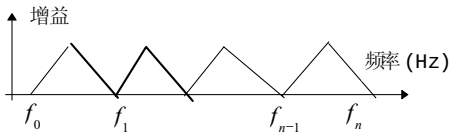

图 4 三角型滤波器（**TF**）

第三种滤波器的设计可以结合前两种的设计，构造一个频带交叉的滤波器组 (**CTF,** Cross-Triangle Filter) (如图 5)

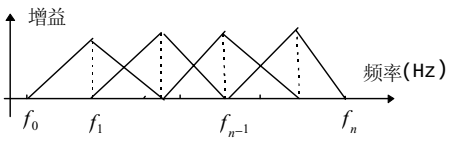

图 5 交叉三角型滤波器（**CTF**）

(3) 离散余弦变换(DCT)

在各个频带的信息分离出来了之后，就可以将频域上的特征反变换到时域上。这里由于处理后的短时功率谱已是实数域上，因而可以用离散余弦变换来实现。

另外一个需要注意的是：从实验数据上来看，变换之后的时域特征值具有很好的分布，呈较明显的集中趋势，而且高维的值与低维的值相比完全可以忽略(见表 1)，这是由汉语的发声特性决定的，因而可以只取较低的几维作为最后的特征向量,从而获得较高的压缩比。本实验中取的是 10 维。

1 维	2 维	3 维	4 维	5 维	6 维	7 维	8 维	9 维	10 维
292.9	3.300	–23.70	5.271	–17.63	6.584	3.260	2.688	–0.845	1.137

表 1 音 qiao 任取一帧的前 10 维特征值，可以看出前述的变化规律。

(4) 讨论

需要指出的是，我们采用的这个方案有一个前提条件，那就是假定音调的变化和语音

的强度无关。但实验表明，如果一个恒定的声源，当它与人耳的距离不同时，音高的感觉也要发生变化。Stevens 作了这方面的实验，并总结出如下的规律[3]：

当声音低于 1000Hz 时，音调的感觉随声音强度的增强而减弱；当声音高于 3000Hz 时，音调的感觉随声音强度的增强而增强；当声音介于 1,000Hz 和 3,000Hz 之间时，音调的感觉基本上不受声音强度的影响。

如果在确定音调的时候，不仅考虑频率的影响，而且也考虑强度的话，处理过程将相当复杂。不过，经过对汉语语音的声谱分析，汉语的主要信息都集中在 300Hz 到 3,400Hz 之间[4]，因而我们可以近似地认为音调只与声音的频率有关。

四、对 Mel 技术的评估

为了检验运用了 Mel 技术之后对汉语语音识别是否会有所帮助，我们设计了一套汉语的非特定人音节识别系统[5]。该系统是对训练集内的各个音节的特征向量进行预处理之后合理聚类，得出每个音节的分布，然后在识别时根据测试集内各个样本对训练结果的后验概率来决定归于何类。语音的识别基元是汉语的音节。

本论文所利用的数据来自电话网络。全部数据取自完全自然发音，环境噪音各种各样，口音南腔北调，年龄分布比较宽。

图 6 是本数据库的音节音长统计结果，从图上可以看出大部分音节长度只有 10 帧左右，这也从另一个侧面说明我们的数据是一个语速较快的现实世界数据库。

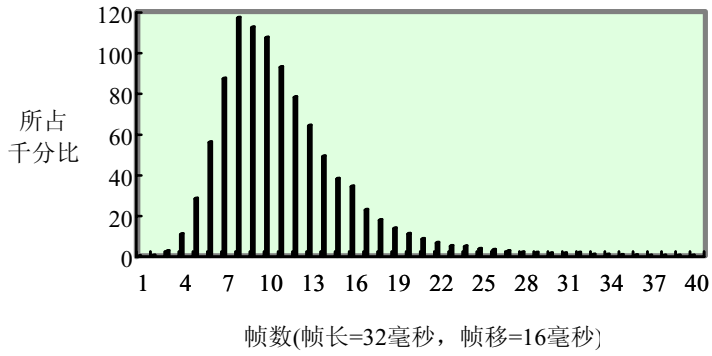

图 6　实验所用数据库中汉语音节音长统计直方图

下面是评估的结果：

(1) 相同条件下与传统的 LPC 参数相比较

测试总样本数 3110 个发音,以下的数据是每种方法测试的前十名命中率：

名次	一	二	三	四	五	六	七	八	九	十
LPC	85.95	89.94	91.22	91.74	92.19	92.38	92.51	92.70	92.86	92.86
MEL, 40 SFs	90.06	95.31	96.69	97.36	97.72	98.07	98.26	98.36	98.49	98.65

表 2　LPC 和 MEL 的比较

(2) 不同参数的比较

a) 三种滤波器设计方案的比较

测试总样本数 3110 个发音，滤波器个数均为 40 个。

名次	一	二	三	四	五	六	七	八	九	十
Mel, 40 TFs	91.03	95.31	96.27	97.14	97.56	97.78	98.04	98.14	98.23	98.39
Mel, 40 SFs	90.06	95.31	96.69	97.36	97.72	98.07	98.26	98.36	98.49	98.65
Mel, 40 CTFs	91.03	95.31	96.27	97.14	97.56	97.78	98.04	98.14	98.23	98.39

表 3　滤波器的选择

b) 滤波器个数的比较

测试总样本数 3110 个发音，滤波器种类均为三角型滤波器。

名次	一	二	三	四	五	六	七	八	九	十
MEL, 40 TFs	91.03	95.31	96.27	97.14	97.56	97.78	98.04	98.14	98.23	98.39
MEL, 20 TFs	90.42	94.79	96.17	96.85	97.40	97.62	97.85	98.01	98.14	98.26

表 4　滤波器个数的选择

五、应用的展望

从前面的结果来看，基于音调的特征提取技术取得了较好的效果，但是它也有一些不足之处，如：一.提取特征用的是最简单的滤波器组相加，显得有些粗糙；二.忽略了其他语音信号特征对音调的影响，除前面已提到的声音强度之外，周相[3]等因素对音调感觉也有一定的作用。如果再采取一些附加措施，可能会取得更好的效果。

参考文献

[1] **Hermansky, H., (1990)** "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Am.,* Apr. 1990，87(4):1738-1752,

[2] **杨治良主编，(1990)** 实验心理学。上海：华东师范大学出版社，1990 年 6 月第一版

[3] **Stevens, S.S. (Ed)**, Handbook of Experimental Psycgology, N.Y.: John Wiley and Sons, 1951

[4] **吴宗济，林茂灿等。(1989)** 实验语音学教程。 北京： 高等教育出版社，1989 年

[5] **郑方，吴文虎，方棣棠 (1996)** "CDCPM 及其在语音识别中的应用，" 《软件学报》，1996 年 10 月，863 专刊，7: 69-75