第35卷 第23期 计 算 机 工 程 2009年12月
Vol.35 No.23 Computer Engineering December 2009
·人工智能及识别技术· 文章编号：1000—3428(2009)23—0201—03 文献标识码：A 中图分类号：TN912

汉语普通话易混淆音素的识别

李晨冲[1,2]，董 滨[2]，潘复平[2]，曾兴雯[1]，颜永红[2]
(1. 西安电子科技大学通信工程学院，西安 710071；2. 中国科学院声学研究所中科信利语音实验室，北京 100190)

摘 要：针对汉语普通话语音识别中易混淆音素的声学特征，把小波包分解理论应用在感觉加权线性预测(PLP)特征中，提出一种新的特征参数提取算法，可以更精确地描述易混淆音素的频谱特征。使用高斯混合模型对新的声学特征进行分类，从而达到区分的目的。实验结果证明，新的特征参数识别结果优于使用传统PLP特征参数的识别结果，识别错误率下降30%以上。

关键词：小波包分解；感觉加权线性预测；语音识别

Recognition of Easily Confused Mandarin Phone

LI Chen-chong[1,2], DONG Bin[2], PAN Fu-ping[2], ZENG Xing-wen[1], YAN Yong-hong[2]
(1. School of Telecommunication Engineering, Xidian University, Xi'an 710071;
2. ThinkIT Speech Lab, Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190)

【Abstract】Aiming at the acoustic features of some easily confused mandarin speech recognition, this paper directs towards revising the Perceptual Linear Predictive(PLP) acoustic feature of these consonants by applying wavelet packet decomposition theory, in which a new feature extraction algorithm is proposed. The new feature can describe frequency spectrum of the easily confused phones more accurately. It uses Gaussian Mixture Modeling(GMM) to classify the new feature for phone discrimination. Experimental results show that the distinguishing error rates of those easily confused consonants are decreased greatly more than 30% compared with traditional PLP feature.

【Key words】wavelet packet decomposition; Perceptual Linear Predictive(PLP); speech recognition

1 概述

在目前汉语普通话语音识别研究中，人们提出了很多特征参数，如线性预测系数(LPC)、线性预测倒谱系数(LPCC)、梅尔倒谱系数(MFCC)和感觉加权线性预测(Perceptual Linear Predictive, PLP)系数等[1]。目前使用较为广泛的是MFCC和PLP，这2种特征参数虽然能够表现语音信号的基本特征，但是也有自身的缺点，它们均基于语音信号短时平稳的假设，在短时傅里叶变换的基础上提取语音特征。短时傅里叶变换以固定的滑动窗对信号进行分析，时域的滑动窗处理等效于频域以滤波器组将信号分频段滤波。滑动窗固定，即各个滤波器的频率形状相同，导致各中心频率沿分析的频带等间隔分布，这样会造成对信号频率成分描述不够充分的情况。

汉语普通话音素f和h、sh和s、ch和c由于发音方法[2]相似，容易混淆，要想对其进行区分，必须对其频谱进行精细的刻画，才能表现各自的声学特征[2]信息，而短时傅里叶变换不能精细地描述易混淆发音的频域信息，对相似发音的鉴别能力有限，导致易混淆音素识别正确率不高。

小波变换是时间和频率的局部变换，并且其时频窗口可以根据不同频率自适应地调节，具有多分辨分析的特点，从而能够精确地反映非平稳信号的瞬间变化，小波包变换是小波变换的推广，对小波变换没有分解的高频部分进行进一步细分，所以对语音信号进行小波包分解能细致地刻画出易混淆音素之间的时频差别，弥补了短时傅里叶变换对时频信息描述不够充分的缺点。

本文借鉴PLP的提取过程，结合小波包对语音频带的多层次划分，并根据人耳听觉特性，选择相应的小波包分解节点系数，提出一种基于小波包分解的特征参数(Wavelet Packet Perceptual Linear Predictive, WPPLP)，在识别算法中，采用混合高斯模型(Gaussian Mixture Modeling, GMM)进行分类。以f和h、sh和s、ch和c这3组汉语普通话易混淆声母为例，分别采用PLP和WPPLP这2种特征作对比实验，结果表明：在不同的高斯混合模型阶数下，使用新的特征参数的识别性能明显优于传统的PLP特征参数。

2 小波包分析的原理

2.1 小波包的定义

现代小波变换称为数学的显微镜，它通过有限个基函数在时间-频率域上对信号进行分析，在控制分辨率的同时，保留了时域信息，因此，在时域信号的处理上受到了极大关注。且小波变换在各分析频段的恒Q(品质因数)特性与人耳听觉对信号的加工特点一致，这一良好的特性为利用小波变换提取语音特征参数奠定了基础。近年来，在语音信号的特征提取中，小波分析已被尝试使用，例如DWT-MFC[3]。

小波变换中的多分辨率分析[4]最终目的是力求构造一个在频率上高度逼近$L^2(R)$空间的正交小波基，这些频率和分辨率不同的正交小波基相当于带通各异的带通滤波器。文献[4]

基金项目：国家“863”计划基金资助项目(2006AA010102, 2006AA01Z195)；国家“973”计划基金资助项目(2004CB318106)；国家自然科学基金资助项目(10574140, 60535030)

作者简介：李晨冲(1983－)，女，硕士研究生，主研方向：语音识别，计算机辅助语言学习；董 滨、潘复平，助理研究员、博士；曾兴雯，教授；颜永红，研究员、博士生导师

收稿日期：2009-04-09 **E-mail：**lichenchong@126.com

从空间的概念上形象地说明了小波的多分辨率特性，给出了正交小波包的构造方法及正交小波变换的快速算法——Mallat 算法。

定义子空间 u_j^n 是 $u_n(t)$ 的闭包空间，而 u_j^{2n} 是函数 $u_{2n}(t)$ 的闭包空间：

$$u_{2n}(t)=\sqrt{2}\sum_{k\in Z}h(k)u_n(2t-k)$$
$$u_{2n+1}(t)=\sqrt{2}\sum_{k\in Z}g(k)u_n(2t-k) \quad (1)$$

其中，$g(k)=(-1)^k h(1-k)$。当 n=0 时，式(1)直接给出：

$$u_0(t)=\sum_{k\in Z}h(k)u_0(2t-k)$$
$$u_1(t)=\sum_{k\in Z}g(k)u_0(2t-k) \quad (2)$$

在多分辨分析中，尺度函数 $\varphi(t)$ 和小波基函数 $\psi(t)$ 满足以下双尺度方程：

$$\varphi(t)=\sum_{k\in Z}h(k)\varphi(2t-k)$$
$$\psi(t)=\sum_{k\in Z}g(k)\psi(2t-k) \quad (3)$$

小波包定义为正交尺度函数 $u_0(t)=\varphi(t)$ 确定的函数族。小波包分解算法为

$$d_n^{k,2l}=\sum_m d_m^{k+1,l}\times h_{m-2n}$$
$$d_n^{k,2l+1}=\sum_m d_m^{k+1,l}\times g_{m-2n} \quad (4)$$

2.2 小波包的分解

小波包变换是多分辨分析的推广，能够为信号提供一种更为精细的分析方法，它同时对信号的低频和高频部分进行多层次划分，在满足 Hersenberg 测不准原理下，将信号按任意的时频分辨率分解到不同的频带，并将信号的时频成分相应地投影到代表不同频带的正交小波空间上。小波包变换可以根据被分析信号的特征，自适应地选择相应频带，使之与信号频谱匹配，进一步提高了时频分辨率，对于汉语普通话易混淆音素 f 和 h、sh 和 s、ch 和 c 的频谱描述，利用了小波包变换的上述特性，依据人耳听觉感知特性，自适应地选择小波包分解节点，细化了易混淆音素的频带特点。关于小波包变换的理解，本文以一个 3 层的分解进行说明，其小波包分解树如图 1 所示。

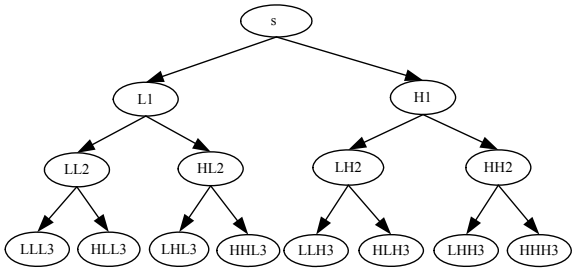

图 1　3 层小波包分解树

其中，L 和 H 分别表示信号的低高频，后面的数字为小波包分解的层数。则信号 S 可表示为 S=LLL3+HLL3+LHL3+HHL3+LLH3+HLH3+LHH3+HHH3。

3　WPPLP特征参数的提取

WPPLP 特征参数的提取过程如下：

(1)对语音信号进行预加重、分帧、加窗。

1)预加重：目的是提升高频部分，使信号的频谱变得平坦，便于进行频谱分析。通常使用一阶数字滤波器实现，即

$$H(z)=1-\mu z^{-1} \quad (5)$$

其中，μ 取值为 0.94。

2)分帧和加窗：目的是满足语音信号的短时平稳，采用 30 ms 为 1 帧，加 Hamming 窗。

$$w(n)=0.54-0.46\cos 2\pi n(N-1)$$
$$0\leqslant n\leqslant N-1 \quad (6)$$

(2)对每帧语音信号进行 6 层小波包分解，根据临界带宽 Z(单位为 Bark)与频率(f)关系公式：

$$Z=6\times\ln\left(\frac{f}{600}+\sqrt{\left(\frac{f}{600}\right)^2+1}\right) \quad (7)$$

其中，$0\leqslant Z\leqslant 19.7$ Bark，$0\leqslant f<8$ kHz；临界带 k 的中心频率 Z_k 位于 0.98k Bark(k=1, 2,…, 20)处。分析临界带宽在频带上的分布，选择 20 个小波包分析节点。6 层小波包分解中节点及其频带的选择如表 1 所示。

表 1　6 层小波包分解中的节点及频带

小波包节点	频带范围/Hz	中心频率/Hz
(6, 0)	0~125	62.5
(6, 1)	125~250	187.5
(6, 2)	250~375	312.5
(6, 3)	375~500	437.5
(6, 4)	500~625	562.5
(6, 5)	625~750	687.5
(5, 3)	750~1 000	875.0
(6, 8)	1 000~1 125	1 062.5
(6, 9)	1 125~1 250	1 187.5
(5, 5)	1 250~1 500	1 375.0
(4, 3)	1 500~2 000	1 750.0
(4, 4)	2 000~2 500	2 125.0
(5, 10)	2 500~2 750	2 562.5
(5, 11)	2 750~3 000	2 812.5
(3, 3)	3 000~4 000	3 500.0
(4, 8)	4 000~4 500	4 250.0
(4, 9)	4 500~5 000	4 750.0
(3, 5)	5 000~6 000	5 500.0
(3, 6)	6 000~7 000	6 500.0
(3, 7)	7 000~8 000	7 500.0

每个子带内的系数为 $SBC_{m,n}$，表示第 m 个子带的第 n 个小波系数，选择 Daubechies 小波族中的 db10 小波进行小波分解。每个子带内的系数取平方，拼接成整个频带内的系数能量谱。

(3)将每个频带内的系数能量谱再按 Bark 域划分标准划分为 20 个子带，每个子带内的能量谱与如下的加权重函数相乘，求和后得到临界带听觉谱。加权重公式为

$$C_k(Z)=\begin{cases}10^{(Z-Z_k+0.5)} & Z\leqslant Z_k-0.5\\ 1 & Z_k-0.5\leqslant Z\leqslant Z_k+0.5\\ 10^{-2.5(Z-Z_k-0.5)} & Z\geqslant Z_k+0.5\end{cases} \quad (8)$$

(4)由于在相同的声强下，人耳对不同的频率所感到的响度并不相同，为了模拟人耳的特点，将式(3)的每帧输出取反对数，做 lg40 dB 等响度曲线函数变换。等响度曲线函数为

$$e(w)=\frac{w^2\times(w^2+1.44\times10^6)}{(w^2+1.6\times10^5)\times(w^2+9.61\times10^6)} \quad (9)$$

(5)经过离散傅里叶反变换(IDFT)之后，用德宾(Durbin)算法计算 11 阶全极点模型，得到的系数再计算倒谱系数，即得到每帧的 12 维 WPPLP 系数。

(6)一个音素的语音段可以分为很多帧，每一帧都进行上面的处理，然后把所有帧的特征系数做平均作为该音素的 12 维特征矢量。

WPPLP 特征参数的提取流程如图 2 所示。

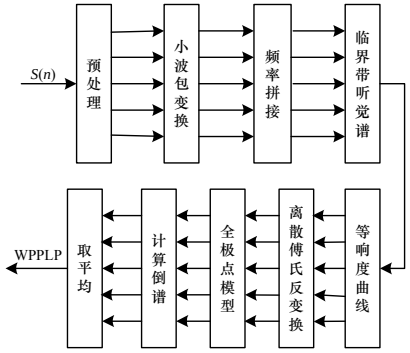

图 2　WPPLP 特征参数的提取流程

4　高斯混合模型

在语音的识别部分，选择 GMM 描述易混淆音素的特征分布。通过大量数据的训练，GMM 可以较好地逼近任意概率密度函数。GMM 本质上是一种多维概率密度函数，一个具有 M 个混合成分的 D 维 GMM，可以用 M 个高斯成员的加权和来表示，即

$$p(\boldsymbol{x}_t \mid \lambda)=\sum_{i=1}^{M} w_i p[\boldsymbol{x}_t \mid \boldsymbol{\mu}_i, \Sigma_i] \tag{10}$$

其中，$\boldsymbol{x}_t$ 是一个 D 维语音特征矢量，本文中的 $\boldsymbol{x}_t$ 即为一个音素的 12 维 WPPLP 特征矢量；w_i (i=1, 2,…, M)为混合权值，相当于每个高斯成员出现的概率，且 $\sum_{i=1}^{M} w_i = 1$，$p[\boldsymbol{x}_t \mid \boldsymbol{\mu}_i, \Sigma_i]$ 为 D 维高斯函数，即

$$p[\boldsymbol{x}_t \mid \boldsymbol{\mu}_i, \Sigma_i]=\frac{1}{(2\pi)^{D/2} \mid \sum_i \mid^{1/2}} \times \exp\{-\frac{1}{2}(\boldsymbol{x}_t-\boldsymbol{\mu}_i)^{\mathrm{T}} \sum_i^{-1} (\boldsymbol{x}_t-\boldsymbol{\mu}_i)\} \tag{11}$$

其中，$\boldsymbol{\mu}_i$ 为均值矢量；$|\sum_i|$ 为协方差矩阵。这里共有 M 个高斯分布函数，其参数为 $\boldsymbol{\mu}_i$ 和 $|\sum_i|$。每个函数受 w_i 加权后，取和得到 $\boldsymbol{x}_t$ 的概率分布。GMM 参数的估计采用极大似然准则，通过 EM[1](Expectation Maximization)迭代算法来实现。

5　实验结果与分析

选择在安静的环境下进行实验数据的录制，对语音信号进行 16 KHz 采样，16 bit 量化。录音人数为 330 人，汉语普通话水平中等，男女各半。每人分别录制 3 组含声母 f 和 h、sh 和 s、ch 和 c 后接不同韵母的单字，每组中每类不同声母单字各 30 个。每个语音信号文件经过语音处理技术中强制对齐的方法切除韵母部分，保留声母部分作为实验数据。其中 290 人的数据用作训练，40 人的数据用作识别。

实验中对每个语音信号分别提取 PLP, WPPLP 这 2 种特征参数，混合高斯模型的阶数选择 16, 64, 256 分别进行实验。实验流程如图 3 所示。

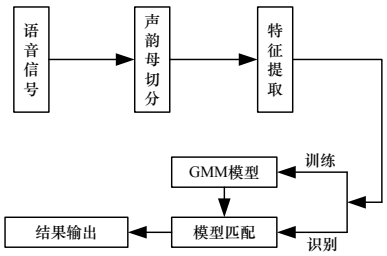

图 3　易混淆音素的识别流程

进行这 3 组易混淆音素实验后，所得识别结果分别如表 2~表 4 所示。

表 2　GMM 模型数为 16 时的识别结果　(%)

音素对	PLP 识别正确率	WPPLP 识别正确率	识别错误率相对下降	错误率平均相对下降
f	93.2	98.3	75.0	67.0
h	93.3	97.2	58.2	67.0
sh	93.1	95.2	30.0	48.0
s	91.5	97.1	65.9	48.0
ch	87.6	89.7	16.9	43.1
c	84.4	95.2	69.2	43.1

表 3　GMM 模型数为 64 时的识别结果　(%)

音素对	PLP 识别正确率	WPPLP 识别正确率	识别错误率相对下降	错误率平均相对下降
f	96.3	98.2	51.4	55.7
h	96.5	98.6	60.0	55.7
sh	94.1	95.6	25.4	50.0
s	93.7	98.4	74.6	50.0
ch	90.2	91.7	15.3	33.1
c	89.4	94.8	50.9	33.1

表 4　GMM 模型数为 256 时的识别结果　(%)

音素对	PLP 识别正确率	WPPLP 识别正确率	识别错误率相对下降	错误率平均相对下降
f	98.0	99.3	65.0	66.7
h	98.1	99.4	68.4	66.7
sh	95.6	97.4	43.2	33.6
s	97.2	98.7	24.0	33.6
ch	90.4	92.5	24.0	39.4
c	91.6	95.8	54.8	39.4

由这 3 组实验结果可见，对于汉语普通话易混淆音素 f 和 h、sh 和 s、ch 和 c 的识别，随着 GMM 阶数的增加，GMM 模型能够更好地逼近易混淆音素特征分布的概率密度函数，因此，采用 PLP 特征和 WPPLP 特征，每组音素的识别正确率均有提高；在相同的 GMM 阶数下，WPPLP 特征比传统 PLP 特征识别正确率均有明显提高，每组的识别错误率平均相对下降在 30%以上，最高相对下降 67%。验证了采用小波包分解理论可以提高易混淆音素识别性能的理论。

6　结束语

本文提出一种区分汉语普通话易混淆音素的新特征参数，借鉴传统 PLP 特征参数的提取原理，把小波包分解应用于汉语普通话易混淆音素特征参数的提取，提高了对语音信号的时频分辨率。采用 WPPLP 特征的识别错误率明显低于 PLP 特征，提高了识别性能。但是 WPPLP 特征参数对所有汉语普通话易混淆音素的区分性还有待进一步研究，继续寻找一种更为稳健的区分性特征，对汉语普通话易混淆发音的识别有很大的实用价值，这也是下一步的研究方向。

参考文献

[1] 王炳锡, 屈　丹, 彭　煊. 实用语音识别基础[M]. 北京: 国防工业出版社, 2005.

[2] 吴宗济, 林茂灿. 实验语音学概要[M]. 北京: 高等教育出版社, 1989.

[3] 刘　鸣, 戴蓓倩, 李　辉, 等. 基于离散小波变换和感知频域滤波的语音特征参数[J]. 电路系统学报, 2000, 5(1): 21-25.

[4] 崔锦泰. 小波分析导论[M]. 西安: 西安交通大学出版社, 1997.

编辑　顾姣健