文章编号:1001 - 9081(2005)06 - 1345 - 02

基于小波子带分解的特征参数对语音自动切分的改进

秦 欢,柴佩琪,陈 锴
(同济大学 电子与信息工程学院,上海 200092)
(sqinhuan@21cn.com)

摘 要:采用了基于小波子带分解的特征提取方法,根据 DCT 和 DWT 两种去相关方法的不同,得到语音信号的特征参数分别为 Subband Based Cepstral(SBC)和 Wavelet Packet Parameters(WPP)。实验切分结果表明,基于小波子带分解的特征参数比 MFCC 取得更好的切分效果。

关键词:隐马尔可夫模型;语音自动切分;Mel 频率倒谱系数;小波子带分解
中图分类号: TP18 **文献标识码**:A

Improvement on automatic speech segmentation using wavelet packet transform features

QIN Huan, CHAI Pei-qi, CHEN Kai
(*Department of Computer Science and Engineering, Tongji University, Shanghai,* 200092)

Abstract: Two new feature parameters based on wavelet packet transform were proposed intend of MFCC. According to the difference of decorrelation method, the two feature parameters were named as subband based cepstral parameters (SBC) and wavelet packet parameters (WPP). The tests indicate that SBC and WPP achieve better performance than MFCC.

Key words: HMM; automatic speech segmentation; MFCC(Mel-Frequency Cestrum Coefficient); wavelet packet transform

目前,广泛应用的自动切分的方法是使用隐马尔可夫模型(HMM)在声学数据的基础上对大语料库进行自动切分,将语料分割成以单音(或单词)为单位,并准确给出每个单音(或单词)在原始语料中的时间起始点。目前的系统大多采用 Mel 频率倒谱系数(Mel-Frequency Cestrum Coefficient, MFCC)作为语音信号特征参数。文献[1]提出了一种基于种子的 HMM 自动切分方法,采用 MFCC 作为特征参数,取得了比较好的切分效果。

本文采用了两种新的基于小波子带分解来提取特征参数的方法,取代 MFCC。实验结果表明,这两种特征参数相比较于 MFCC 取得了更好的切分效果。

1 基于小波子带分解的特征参数

1.1 MFCC 参数简介

MFCC 定义为语音信号经过快速傅里叶变换后所得的加窗短时信号的实倒谱。一般的 MFCC 的计算过程大致如下:

1) 对每一帧语音信号进行短时傅里叶变换得到其频谱。

2) 求它的频谱幅度的平方,即能量谱,并用一组三角形滤波器在频域对能量谱进行带通滤波。

3) 对滤波器组的输出取对数,然后再进行傅里叶逆变换,即得到 MFCC 参数。

在上述基本语音参数基础上加入时域衍生选项,包括一阶回退系数(Delta 系数)和二阶回退系数(加速系数),通常可以取得更佳的效果。

1.2 基于小波子带分解的特征参数

基于小波子带分解的特征参数提取在方法上与 MFCC 求取过程相似,它首先对时域语音信号分帧加窗,然后采用小波包变换把加窗信号分解成 24 个子带的系数,然后计算每个子带的能量谱,并对能量谱系数统一取对数,最后对所得到的特征参数进行去相关得到最终的特征参数。根据去相关方法的不同,最终的特征参数可分为 SBC 和 WPP 两种。SBC 采用 DCT 变换去相关,而 WPP 采用 DWT 变换去相关。

对于语音信号 s(n),基于小波子带分解的特征参数提取的总体框架如图 1。

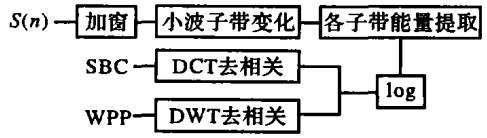


图 1 基于小波子带变换特征参数的提取

1.3 小波子带分解及能量谱

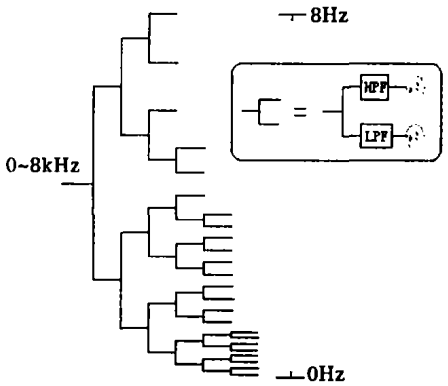


图 2 24 子带小波分解树

小波包变换将每一帧的语音信号分解成 24 个小波子带系数,本文所用的 24 - subband 小波包分解树的结构如图 2 所示。图中,HPF 和 LPF 分别指在用小波包分解时用到的高通和低通滤波器。

收稿日期:2004 - 12 - 03;修订日期:2005 - 01 - 26

作者简介:秦欢(1980 -),男,硕士研究生,主要研究方向:模式识别与语音信号处理; 柴佩琪(1935 -),上海人,教授,博士生导师,主要研究方向:模式识别、语音信号处理、人工智能; 陈锴(1977 -),安徽人,博士研究生,主要研究方向:模式识别与语音信号处理.

按照这样的子带划分，每个子带的频率范围如表 1 种所示。从表中可以看出这样的划分着重突出了位于 0 ~ 1000Hz 频率范围内的子带（主要为对应于第 6 层分解的 8 个低频子带），实际语音信号的绝大部分能量也集中于这些频带范围之内。第 9 到第 18 个子带的频率位于 1 000Hz ~ 3 500Hz，每个子带的频率宽度相等，都是 250Hz。第 19 到第 21 个子带占据 3 500Hz ~ 5 000Hz 的频带，每个子带的带宽为 500Hz。最后第 22 到第 24 个子带占据 5 000Hz ~ 8 000Hz 的频带，每个子带的带宽都为 1 000Hz。实验结果表明，这样的划分取得比较好的切分效果。

表 1 24 个子带的频率范围（Hz）

子带	频率范围	子带	频率范围	子带	频率范围
1	0 ~ 125	9	1 000 ~ 1 250	17	3 000 ~ 3 250
2	125 ~ 250	10	1 250 ~ 1 500	18	3 250 ~ 3 500
3	250 ~ 375	11	1 500 ~ 1 750	19	3 500 ~ 4 000
4	375 ~ 500	12	1 750 ~ 2 000	20	4 000 ~ 4 500
5	500 ~ 625	13	2 000 ~ 2 250	21	4 500 ~ 5 000
6	625 ~ 750	14	2 250 ~ 2 500	22	5 000 ~ 6 000
7	750 ~ 875	15	2 500 ~ 2 750	23	6 000 ~ 7 000
8	875 ~ 1 000	16	2 750 ~ 3 000	24	7 000 ~ 8 000

每一帧语音信号按照以上小波树结构进行小波包分解后，对于每个子带，其能量谱通过用该子带所有小波系数的平方和取平均而得到，具体计算方法如下：

$$S_i = \frac{\sum_{m \in i} [wp(i,m)]^2}{N_i}$$

其中，S_i 为第 i 个子带能量谱的值，$i = 1,2,\cdots,24$。

wp：为每一帧语音时域信号所有的小波系数的集合。

$wp(i,m)$：为第 i 个子带的第 m 个小波系数。

N_i：为第 i 个子带小波系数的总数。

1.4 去相关

由于在用 HMM 模型对参数进行训练和识别时，输出概率模型采用高斯模型模拟，并且协方差矩阵为对角阵，而从语音信号提取的特征参数将作为 HMM 的训练参数。这样就要求各维特征参数之间的相关性比较低，以便与对角协方差矩阵相适应。因此在对以上步骤取得的特征参数有必要进行去相关处理。对于 24 维的能量谱参数，本实验中采取了两种去相关的方法，分别为 DCT 和 DWT，最终求得的特征参数对应为 SBC 和 WPP。

1）SBC：

$$SBC(n) = \sum_{i=1}^{24} \log S_i \cos\left(\frac{n(i-0.5)}{24}\pi\right), n = 1,2,\cdots,13$$

2）WPP：

通常对于 24 维的特征参数，采用 Daubechies4 小波进行 3 层小波变换后，特征维数仍为 24。我们对每一维的参数值通过方差统计分析发现，只有前 13 维的参数变换较大，而后面各维变化较小，并且参数的能量主要集中在前 13 维。因此，在进行小波变换后，保留前 13 维参数作为最终去相关取得的结果。

因此在两种去相关方法中，最终提取出来的特征参数都是 13 维。对于去相关后的 13 维特征参数，分别求取一阶和两阶 Delta 系数，最后的特征参数由三部分组成，分别是：去相关后的数据（13），一阶 Delta 系数（13），两阶 Delta 系数（13），一共 39 维，与本文中所用 MFCC 参数的维数一样。

2 基于 HMM 的自动切分概述

基于 HMM 的自动切分的一般做法是先对所有的语音进行特征提取，得到对应的特征参数；用部分语料的特征参数作为 HMM 模型的输入参数训练得到每个单音的初始化模型；然后用所有语料的特征参数重估训练初始化模型，得到每个单音最终的 HMM 模型；最后，根据这些模型用识别的方法对所有语料进行时间调整，得到每个单音的时间边界。

3 实验结果

本实验语料全部来自同一名女性专业播音员的发声，选取一个小时左右的英文语料，平均长度 15s 左右，采样率为 16kHz。其中，取出部分作为训练各个单音初始 HMM 的语音，它们都配有准确的时间标记，即所有单音在语音中的时间起始点已经准确给出。窗长取 16ms，帧偏移为 3ms。小波包分解采用 Daubechies4 小波的滤波器。

3.1 相关性统计

本实验中，对去相关后的 13 维特征参数，首先进行归一化，归一化后的特征向量用 X 表示；然后通过计算 X 自相关矩阵所有非对角线元素的和来估计相关性。具体计算方法如下：

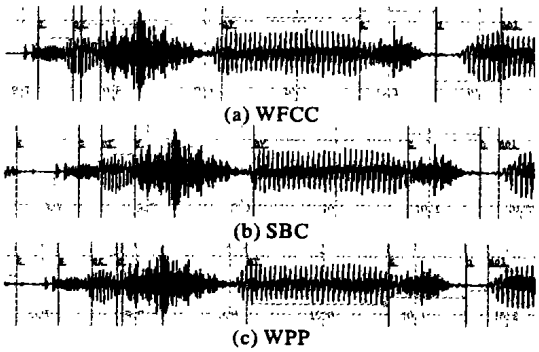
$$R = E[XX^T], \eta = \sum_{i=1}^{T} \sum_{m=1}^{13} \sum_{\substack{n=1 \\ n \neq m}}^{13} R(n,m)$$

其中，X 为归一化的去相关数据，共 13 维。η 为相关性统计量。T 为统计时总共使用的统计样本数（以帧为单位）。根据该式，如果参数各维相关性比较低的话，那么其统计值应该相应的较小。

统计结果表明：对于本文的特征参数提取方法，DCT 的统计值略小于 DWT 的统计值，也就是说 DCT 的去相关效果略好于 DWT，这一点从最后的切分结果中也有所反映。

3.2 MFCC、SBC 和 WPP 切分结果对比

从切分结果可以看出：SBC 和 WPP 的切分结果相似，SBC 略优于 WPP，这体现在：在单词之间 s，z，zh，sh 等一类辅音边界点上以及单词内的辅音和元音分界点上，SBC 稍好于 WPP，不过误差一般都在 10ms 之内。这一点与相关性统计的结果一致。SBC 和 WPP 相对于 MFCC，都有很大改进。



(a) WFCC

(b) SBC

(c) WPP

图 3 结果对比

参考文献：

[1] 祝瑶卿，柴佩琪. 基于 HMM 连接语音自动切分中的初始化模型研究[J]. 微型电脑应用，2003，19(7).

[2] SARIKAYA R，PELLOM B，HANSEN JHL. Wavelet Packet Transform Features with Application to Speaker Identification[A]. NORSIG - 98，IEEE Nordic Signal Processing Symposium[C]，Vigso，Denmark，1998.81 - 84.

[3] The HTK Book (for HTK Version 3.1) © COPYRIGHT 1995 - 1999 Microsoft Corporation[Z]，version 3.1.

[4] 易克初，田斌，付强. 语音信号处理[M]. 北京：国防工业出版社，2000.