2011 年 4 月 第 4 期
Vol. 32 No. 4 2011

VTS 特征补偿算法在语音识别中的实用性研究

杨 钊,杜 俊,胡 郁,刘庆峰,戴礼荣

(中国科学技术大学 讯飞语音实验室,安徽 合肥 230027)
E-mail:yangzhao@ mail. ustc. edu. cn

摘 要: 在语音识别实际应用中,由于噪声的多样性,会造成训练和测试的失配,导致系统性能下降. 特征补偿作为鲁棒性语音识别的一种重要方法,通过对训练和测试环境之间差异的研究,在特征空间中修正语音特征,使得修正后的测试语音特征能够更加接近训练语音特征. 本文介绍一种实用的基于环境模型矢量泰勒级数(VTS)近似的特征补偿算法. 首先验证传统的 VTS 离线算法在实际车载环境下的有效性;其次由于离线算法本身运算量很大,为了使其实用化,本文对算法进行改进,使其在提高效率的同时又能够保证与离线时相当的性能. 通过实验结果验证,本文提出的实用化 VTS 算法在识别性能上相当接近离线时最好的性能.

关 键 词: 失配;矢量泰勒级数;实用化;特征补偿
中图分类号: TP391 **文献标识码**:A **文 章 编 号**:1000-1220(2011)04-0782-05

Application of VTS Approximation Based Feature Compensation Approach to Speech Recognition

YANG Zhao, DU Jun, HU Yu, LIU Qing-feng, DAI Li-rong
(*iFly Speech Lab, University of Science and Technology of China, Hefei* 230027, *China*)

Abstract: In the environment with various noises, the mismatch between training and testing will result in a great reduction of speech recognition performance. As an effective method for robust speech recognition, feature compensation can refine testing noisy feature to be closer to training feature. In this paper, we proposed a practical feature compensation approach based on Vector Taylor Series (VTS) approximation using explicit model of environmental distortion. Firstly, we test and verify the effectiveness of traditional offline VTS algorithm in real car environment, however this offline algorithm has a large amount of calculation, in order to make it practical, the algorithm has been improved to increase efficiency and keep the performance comparable to the offline condition. From experimental results, performance of the practical VTS algorithm proposed in the paper is much close to the best performance of the offline condition.

Key words: mismatch; vector taylor series; practical; feature compensation

1 引 言

以混合高斯作为概率密度分布的隐马尔科夫模型(Hidden Markov Model, HMM)已经普遍应用于当今主流的自动语音识别系统中[1]. 然而,在大多数情况下,HMM 模型的训练数据都是在相对安静的环境下录制的. 用这些数据训练得到的系统,虽然在相同的环境下可以取得很好的识别性能,但在实际带噪环境下,由于噪声等因素的影响,测试数据与训练数据之间存在较大的失配,性能往往会变的很差. 因此,语音识别系统的噪声鲁棒性一直是研究的热点.

对于语音识别的噪声鲁棒性问题,有人提出了一种实用性比较强的特征增强/补偿算法,即一阶矢量泰勒级数(Vector Taylor Series VTS)[2,3]算法,它能够方便地利用噪声统计信息对特征进行补偿,进而提高识别性能. 在此基础上还有一些改进算法,如基于序列噪声估计的一阶 VTS 算法[4]、基于最大后验的一阶 VTS 算法[5]和高阶 VTS 算法[6],这些算法对噪声环境下的语音识别的性能都有一定程度的改善. 但无论是一阶 VTS 算法及其改进算法还是高阶 VTS 算法,均为离线算法,其性能的改善在很大程度上是以计算量为代价,而实际应用中,对语音识别的实时性要求很高,以上的各种离线算法均难以满足实用中这种实时性要求. 本文正是针对以上问题,提出了实用化的一阶 VTS 算法,在保证性能的基础上,大大提高了算法运行的效率和实时性. 结合实际录取的测试噪声数据,进行了一系列对比实验,证明了本文提出方法的有效性.

2 一阶 VTS 离线算法回顾[6]

假设带噪语音在时域上是符合下面这个显式的失真模型:

收稿日期:2010-01-08 收修改稿日期:2010-03-08 作者简介:**杨 钊**,男,1983 年生,硕士研究生,研究方向为语音识别、语音信号处理;**杜 俊**,男,1982 年生,博士,研究方向为语音识别;**胡 郁**,男,1978 年生,博士,研究方向为语音合成,语音识别以及语音评测;**刘庆峰**,男,1973 年生,博士,研究方向为语音合成;**戴礼荣**,男,1962 年生,教授,博士生导师,研究方向为语音信号处理、说话人与语种识别、多媒体通信、DSP.

$$y[t]=x[t]+n[t] \tag{1}$$

$x[t]$,$n[t]$和$y[t]$分别表示干净语音,加性噪声和带噪语音.

在忽略滤波器组间相关性的情况下,失真模型在 log-power 域可表示为:

$$y=\ln(e^x+e^n) \tag{2}$$

其中 y,x 和 n 分别表示带噪语音,干净语音和噪声的 log 功率谱.

为了得到干净语音信号的估计,我们首先需要得到准确的噪声模型参数.为此,在训练阶段,用未经过倒谱均值规整(Cepstral Mean Normalization CMN)的干净语音的 Mel 频率倒谱系数(Mel-Frequency Cepstral Coefficients ,MFCC)训练得到描述干净语音的 MFCC 特征分布的高斯混合模型(Gaussian Mixture Model GMM),$p(x_t^c)=\sum_{m=1}^{M}\omega_m N(x_t^c;\mu_{x,m}^c,\Sigma_{x,m}^c)$,其中 ω_m,$\mu_{x,m}^c$和$\Sigma_{x,m}^c$为第 m 个高斯的权重,均值向量和对角协方差矩阵.相应的模型参数可以通过以下公式从倒谱域转换到 log-power 域:

$$\mu_{x,m}^l=C^+\mu_{x,m}^c \tag{3}$$

$$\Sigma_{x,m}^l=C^+\Sigma_{x,m}^c(C^+)^T \tag{4}$$

其中 C^+为离散余弦变换(DCT)矩阵的 Moore - Penrose 逆,上标"l"和"c"表明在 log-power 域和倒谱域.

设噪声特征矢量 n^c 满足均值为 μ_n^c,对角斜方差为 Σ_n^c 的单高斯分布,则在识别阶段,对于任意一段语音,噪声模型参数(μ_n^c, Σ_n^c)按以下步骤估计:

步骤 1. 初始化:首先通过语音的初始静音段在倒谱域得到初始噪声模型参数的估计.

步骤 2. 将当前步骤得到的噪声模型参数通过以下公式从倒谱域转换到 log-power 域:

$$\mu_n^l=C^+\mu_n^c \tag{5}$$

$$\Sigma_n^l=C^+\Sigma_n^c(C^+)^T \tag{6}$$

步骤 3. 在 log-power 域上计算对于噪声估计和干净语音参数估计所需要的参数 $\mu_{y,m}^l$, $\Sigma_{y,m}^l$, $\Sigma_{xy,m}^l$, $\Sigma_{ny,m}^l$,这些参数的估计见后续(18)-(21)式.

步骤 4. 将以上计算所得的参数转换到倒谱域:

$$\mu_{y,m}^c=C\mu_{y,m}^l \tag{7}$$

$$\Sigma_{y,m}^c=C\Sigma_{y,m}^lC^T \tag{8}$$

$$\Sigma_{xy,m}^c=C\Sigma_{xy,m}^lC^T \tag{9}$$

$$\Sigma_{ny,m}^c=C\Sigma_{ny,m}^lC^T \tag{10}$$

步骤 5. 使用以下更新公式得到新的噪声模型参数:

$$\bar{\mu}_n=\frac{\sum_{t=1}^{T}\sum_{m=1}^{M}P(m|y_t)E[n_t|y_t,m]}{\sum_{t=1}^{T}\sum_{m=1}^{M}P(m|y_t)} \tag{11}$$

$$\bar{\Sigma}_n=\frac{\sum_{t=1}^{T}\sum_{m=1}^{M}P(m|y_t)E[n_tn_t^T|y_t,m]}{\sum_{t=1}^{T}\sum_{m=1}^{M}P(m|y_t)}-\bar{\mu}_n\bar{\mu}_n^T \tag{12}$$

其中

$$P(m|y_t)=\frac{\omega_m P_y(y_t|m)}{\sum_{l=1}^{M}\omega_m P_y(y_t|l)} \tag{13}$$

为了方便,在公式(11)-(12)中我们去掉了指示倒谱域的上标"c",$P(y_t|m)$为带噪语音 y_t 在通过矩匹配补偿得到的带噪语音模型的第 m 个高斯($N(y_t;\mu_{y,m},\Sigma_{y,m})$)上的概率密度函数,$E[n_t|y_t,m]$,$E[n_tn_t^T|y_t,m]$ 为相关条件期望,如下:

$$E[n_t|y_t,m]=\mu_n+\Sigma_{ny,m}\Sigma_{y,m}^{-1}(y_t-\mu_{y,m}) \tag{14}$$

$$E[n_tn_t^T|y_t,m]=E_n[n_t|y_t,m]E_n^T[n_t|y_t,m]+\Sigma_n-\Sigma_{ny,m}\Sigma_{y,m}^{-1}\Sigma_{yn,m} \tag{15}$$

步骤 6. 重复步骤 2-5n 次

由以上可以看出,噪声参数重估在倒谱域(步骤 5),而统计量的计算在 log-power 域(步骤 3),这是由于在 log-power 域容易得到统计量的显式表达式,并且可以考虑各维之间的相关性以得到统计量的准确估计.在 log-power 域中,对于(2)式的失真模型,对 y 一阶展开近似有:

$$y\approx\ln(\exp(\mu_x)+\exp(\mu_n))+\alpha(x-\mu_x)+(I-\alpha)(n-\mu_n) \tag{16}$$

其中 α 为:

$$\alpha=\frac{1}{1+\exp(\mu_n^l-\mu_x^l)} \tag{17}$$

因此我们可以得到步骤 3 中参数为(其中 A 为 α 构成的矩阵):

$$\mu_{y,m}^l=\log(\exp(\mu_{x,m}^l)+\exp(\mu_n^l)) \tag{18}$$

$$\Sigma_{y,m}^l=A\Sigma_{x,m}^lA^T+(I-A)\Sigma_n^l(I-A)^T \tag{19}$$

$$\Sigma_{yx,m}^l=(\Sigma_{xy,m}^l)^T=A\Sigma_{x,m}^l \tag{20}$$

$$\Sigma_{yn,m}^l=(\Sigma_{ny,m}^l)^T=(I-A)\Sigma_n^l \tag{21}$$

得到噪声模型参数估计以后,在倒谱域采用最小均方误差(MMSE)估计干净语音:

$$\hat{x}_t=E[x_t|y_t]=\sum P(m|y_t)E_x[x_t|y_t,m] \tag{22}$$

$E[x_t|y_t,m]$为对于第 m 个高斯在给定 y_t 的情况下 x_t 的条件期望:

$$E_x[x_t|y_t,m]=\mu_{x,m}+\Sigma_{xy,m}\Sigma_{y,m}^{-1}(y_t-\mu_{y,m}) \tag{23}$$

3　实用的 VTS 算法

从上节介绍的基于特征域的 VTS 离线算法我们可以看出,为了得到准确的噪声参数估计,需要用到整句语音信息,训练模型一般需使用高斯数为 128 或者更高的高斯混合模型,而且需要进行多次迭代才能得到较好的结果,这将大大增加运算量,降低系统的处理速度,这些都是在实用时难以接受的,为了使 VTS 算法实用化,必须在基本保持离线性能的基础上,尽量提高算法的效率,基于以上原因提出的实用化 VTS 算法其识别流程图如下页图 1 所示.

从离线 VTS 算法可以看出,噪声模型参数重估消耗了最多的运算量.为了提高算法效率,在图 1 所示的实用 VTS 算法的识别阶段,在特征补偿时去掉了传统的离线 VTS 算法所采用的参数重估过程,但是这样必然会带来噪声模型参数估计精度的下降,这时,噪声模型参数的初始化就显得尤其重要,并且从后面的实验可以看出,不论噪声参数是否重估,初始化的准确性对性能都有很大影响.为了得到更加准确的初始化参数以进行后续处理,从系统流程图可以看到,算法种采用了两种方式,即一遍解码和两遍解码方式,下面将分别对其

进行说明.

3.1 一遍解码方式

在传统的 VTS 离线算法中一般是固定选择前 n 帧进行噪声参数的初始化,但这种选择固定帧数的方法可能导致选取的段中包含语音帧或者包含的噪声帧太少以致对噪声信息的统计不充分,都会对性能有较大影响. 因此在基于一遍解码的实用算法中,为了提高噪声参数初始化的准确性,本文采用语音端点检测(VAD)算法找出语音中的静音段以获得良好的参数初始化.

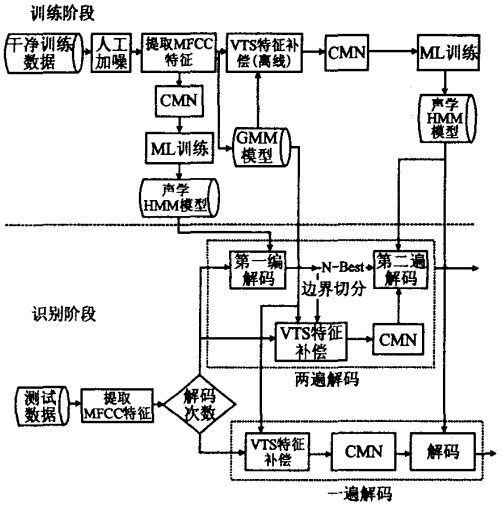

图 1 实用 VTS 特征补偿识别系统流程图

Fig. 1 Flowchart of our feature compensation approach

自动语音端点检测(Voice Activity Detection VAD)是语音信号处理系统的重要部分,作用是从输入信号中检测语音的起止点,良好的端点检测能够保证较好的噪声参数初始化,从而有效的保证识别的正确率. 本文的 VAD 是基于能量双门限算法,并且为了加快处理速度,只使用起始静音段.

3.2 两遍解码方式

两遍解码原理如图 1 中所示,输入语音首先通过识别器一遍解码,得到相应的 N-best 解码结果和边界切分. 第二遍解码则是在第一步得到的 N-best 解码结果的基础上,利用通过 VTS 特征补偿得到的新特征,再次解码,最终得到较好的识别结果.

在两遍解码方式中,第一遍解码采用的是一种快速解码方式,当一句话结束时解码也相应完成,得到边界切分,进而得到初始噪声参数,由于这时不需要进行参数重估,VTS 补偿算法运行量很小,而第二遍解码只是在第一遍解码基础上再次解码,只需找出 N-best 中概率最大的结果并输出,计算量也比较小,算法能达到实用的需求.

3.3 识别流程

在实用算法中我们对训练数据进行了人工加噪,并且采用加噪数据来训练 GMM,这样做的一个重要的原因就是在识别阶段,如果用干净数据训练得到的 GMM 进行特征补偿时会导致比较大的残差,使得识别性能下降. 同时我们也使用到了两个 HMM,分别为用加噪数据训练的 HMM 及用对加噪数据做离线 VTS 特征补偿后的数据训练的 HMM,这里做的离线 VTS 特征补偿是为了得到准确的噪声估计,分别使用两个 HMM 是为了在识别时尽量与特征匹配,以得到最好的识别结果. 实用算法的识别时处理流程为:

当一句话 x_t 来临时:

1. 提取 MFCC 特征

2. 判断解码方式,一遍解码方式跳到步骤 3,两遍解码方式跳到步骤 4

3. 使用固定的前 n 帧或使用 VAD 算法得到的初始静音段来初始化噪声模型参数,做实用 VTS 特征补偿,然后进行解码

4. 一遍解码,得到 N-Best 结果和边界切分,根据边界切分得到噪声模型参数,并做实用 VTS 特征补偿,在 N-Best 基础上进行第二遍解码,最终得到识别结果.

4 实验结果和分析

4.1 实验配置

为了验证本文提出方法的有效性,在以上给出的实际录取的真实数据库上进行了一系列实验. 对于一段语音,先去除直流分量,预加重(因子为 0.97),汉明窗加窗,帧长为 25ms,帧移为 10ms,提取 MFCC 特征参数,MFCC 特征是在功率谱上得到,总计为 13 维(包括 C0),计算一阶差分、二阶差分总计构成 39 维. Mel 频率滤波器组数为 23. 对于 HMM,它是基于音素的右相关声韵母的连续概率密度的隐马尔科夫模型,对于不同的声韵母分为 3、5 或 7 个状态,每个状态由 8 个高斯组成,其中静音的音素模型为 5 状态,每个状态 12 个高斯. 对用于实用的 VTS 特征补偿的 GMM,分别用了 8 高斯和 128 高斯,离线系统都是基于 GMM 为 128 高斯并进行噪声模型参数的重估. 同时还有一些配置在实用化 VTS 特征补偿算法中也起到了相当重要作用,下面将分别介绍:

4.1.1 训练数据加噪

在识别阶段时,做 VTS 特征补偿时需要用到 GMM,识别需要用到 HMM,为了达到更好的识别性能,对干净训练数据进行了人工加噪,加噪的信噪比共采用了 5 种,分别为 5dB、10 dB、15 dB、20 dB 和 clean,按以上信噪比将训练数据分为 5 份,比例为 1:1:1:1:1,再从抽取的噪声数据中随机取一段按已定信噪比加入到干净语音段,其中 clean 代表不加噪,即保持原语音不变.

4.1.2 HMM 训练

作为由干净数据加噪得到的训练数据或加噪数据进行离线 VTS 特征补偿后的数据,其 HMM 可以采用 SPR 方法[7]训练获得. 这么做主要基于以下两点考虑:一是使用 SPR 方法训练 HMM 速度快,二是得到的 HMM 模型在噪声环境下比传统的重训得到的模型更好. 在 SPR 方法中,首先根据干净的 HMM 在干净的数据集上得到帧或状态的切分,$L_{\text{jm}}^{t}(\tau)$. $O^{c}(\tau)$为加噪数据,则重估方程为:

$$\hat{\mu}_{jm}=\frac{\sum_{\tau=1}^{T}L_{\mathrm{jm}}^{s}(\tau)O^{c}(\tau)}{\sum_{\tau=1}^{T}L_{\mathrm{jm}}^{s}(\tau)} \tag{24}$$

$$\hat{\Sigma}_{jm}=\frac{\sum_{\tau=1}^{T}L_{\mathrm{jm}}^{s}(\tau)(O^{c}(\tau)-\hat{\mu}_{jm})(O^{c}(\tau)-\hat{\mu}_{jm})^{T}}{\sum_{\tau=1}^{T}L_{\mathrm{jm}}^{s}(\tau)} \tag{25}$$

其中

$$L_{\mathrm{jm}}^{s}(\tau)=p(q_{jm}(\tau)\mid S_{T}^{c},M) \tag{26}$$

$$S_{T}^{c}=S^{c}(1)S^{c}(2)\cdots S^{c}(T) \tag{27}$$

不过,以上都是建立在噪声平稳的情况下. 通过 SPR 训练模型时,帧/状态的切分是通过干净的模型和干净的数据得到,使帧/状态的切分并没有发生改变,是一种基于模型的补偿方式,训练得到的模型与干净模型具有一样的拓扑结构.

4.1.3　GMM 训练

VTS 离线特征补偿算法为了提高性能,一般使用高斯数比较大的 GMM. 但在实用时,为了加快处理速度,需要减少 GMM 高斯的数目,这时考虑到对实际性能的影响,有必要修改 GMM 涨高斯的策略. 当 GMM 高斯数比较小时,可以采用逐个涨高斯的训练方法,以得到比较精确的 GMM.

4.2　数据库介绍

实验中使用的干净数据大约 96 小时,共 14 万句,由 4 个数据库的子集组成,包括在录音环境下录制的标准普通话普通文本朗读数据(文本为预先选定的一些报刊杂志文章及剧本常用词汇等内容,拥有较广的覆盖率),还包括在办公室环境下在 5 种 PDA 设备上及在录音室环境下录制的四字短语数据. 加噪用的噪声数据是在不同车型、不同车况下录制的,该噪声数据的特点是比较平稳,并且低频占主要成分.

测试集共有 3 个,为在不同环境下录制的真实数据,分别为:

(1) 车载环境下录制的数据库,存在开窗户或开空调等背景背景噪声,共 1952 个条目,主要是包含人名和控制命令,信噪比较高,后面的总结中定义为测试集 1.

(2) 在食堂、道路和车载环境下录制的数据库,主要是包含人名和命令词,共 11993 个条目,信噪比较低,后面的总结中定义为测试集 2.

(3) 在办公室和卧室环境下录制的数据库,主要也是包含人名和命令词,共 7999 个条目,信噪比较高,后面的总结中定义为测试集 3.

需要对以上,实验中主要是针对测试集 1 来比较性能,测试集 2、3 并没有做任何优化,在此列出只是为了验证本方法的鲁棒性.

在实验测试实用算法时是比较了 3 种方法,如下定义:

(1) 方法 1:一遍解码,初始化固定选择前 n 帧(本文中为前 10 帧)

(2) 方法 2:一遍解码,使用语音端点检测算法

(3) 方法 3:使用两遍解码方式

传统离线一般采用的是前 10 帧初始化的方式,方法 1 是为了与传统离线方法性能进行比较. 方法 2 是在方法一基础上的改进,它们均采用一编解码方式,只不过方法 2 中使用语音端点检测算法使初始化更加准确,对噪声参数估计更好,性能也会带来一定提升. 方法 3 是采用两遍解码方式,它能够截出首尾静音段,使参数估计更加准确,带来性能进一步提升.

4.3　实验结果

为了验证人工加噪方法和 VTS 特征补偿算法的有效性,表 1 比较了几种离线算法的性能(性能评价指标为词正确率). 这里的离线算法均是用最大似然进行重估(4 步 EM 算法迭代),以下为其性能:

表 1　未加噪、加噪和离线 VTS 算法性能比较

Table 1　Compare of results of different methods

	测试集 1	测试集 2	测试集 3
方法 1	87.76	89.50	95.39
方法 2	89.65	90.14	95.36
方法 3	91.03	88.35	95.66
方法 4	91.29	88.76	95.60
方法 5	92.32	92.14	95.87

方法 1:干净训练集,不做 VTS

方法 2:加噪训练集,不做 VTS

方法 3:干净训练集,128 高斯 GMM, VTS 离线,前 10 帧初始化

方法 4:加噪训练集,128 高斯 GMM,VTS 离线,前 10 帧初始化

方法 5: 加噪训练集,128 高斯 GMM,VTS 离线,使用 Force -alignment 得到的边界进行初始化(离线最好性能)

从结果可以看出,对训练数据进行人工加噪相对于不加噪的情况性能有不少提升,做 VTS 比不做 VTS 的性能也有很大改善,同时,我们也看到,采用干净数据训练得到的 GMM 模型做 VTS 相对于干净训练集、不做 VTS 时,性能在测试集 1、3 有一定提升,但是在测试集 2 上反而有不小的下降,原因正如在第 2 节中提到的噪声参数初始化的影响,这在采用加噪数据训练的 GMM 做 VTS 上也得到体现,相对于加噪训练集、不做 VTS 的性能,在测试集 1、3 上有一定的提升,但是在测试集 2 上也有不小的下降. 但是当我们采用 Force-alignment得到的边界进行初始化,在进行参数重估时,

表 2　实用算法性能比较

Table 2　Compare of practical results

	测试集 1	测试集 2	测试集 3
GMM 为 128 高斯,实用算法方法 3	91.75	92.02	95.67
GMM 为 8 高斯,实用算法方法 1	91.39	90.44	95.60
GMM 为 8 高斯,实用算法方法 2	91.65	90.84	95.65
GMM 为 8 高斯,实用算法方法 3	91.96	92.04	95.80

可以看到,在测试集 1、2 上有了大幅度提升,测试集 3 也有一定的改善,这说明了噪声参数初始化的重要性. 并且从上我们还可以看出,由人工加噪数据得到模型与实际更匹配,进而带

来性能的改善.

表2(见上页)比较了几种实用算法的性能.其首先列出了GMM为128高斯实用算法方法3的性能.性能已经接近了离线最好的性能,但是128高斯仍然是一个比较复杂的模型,为了进一步加快处理速度,我们采用逐个涨高斯方法训练得到的8高斯的GMM,训练集做VTS并重训HMM,得到的性能如表2所示.

可以看出,对于GMM为8高斯,在实用算法方法2中,在做VAD时只需缓冲较短的时间以得到前静音段.在方法3中,第一遍解码得到10-Best结果,第二遍解码将从第一遍得到的结果进一步挑选.从以上结果可以看出,方法2相对于方法1识别性能有一定提升,但比起方法3性能要差一些.但是Two Pass需要进行两次解码,自然会带来更多的额外开销,而VAD只需用到一次解码,做端点检测速度较快,在速度上有一定优势,如果能设计出更好的VAD算法,必将进一步提高识别的性能.

表3 方法三中第一遍解码的覆盖率

Table 3 Coverage of the first pass decoding of method 3

	测试集1	测试集2	测试集3
覆盖率	98.92	98.57	99.45

对于实用算法方法3,第一遍解码的覆盖率如表3所示,可以看出,这时覆盖率已经比较大,再增加N的值虽然会增大覆盖率,但是也会带来计算量的增加,并且性能改善不多,得不偿失.

最后我们来考察一下该算法的在减少计算量、提高系统运行效率方面的有效性,在一台单独的服务器及由以上3个测试集组成的数据库上,我们分别测试了GMM为128高斯的VTS离线算法和GMM为8高斯的VTS实用算法,从结果来看,实用算法方法1速度比离线速度快18倍,考虑到方法2、3中会分别用到VAD算法和两遍解码,速度也要快10倍左右,可见VTS实用算法大大提高了运行速度,证明了本文方法提出方法的有效性.

5 总结和展望

在本文中,主要讨论了VTS实用特征补偿算法.通过以上实验我们已经验证了该方法的有效性,它的优点主要在于:比起传统的VTS离线算法,它大大减少了计算量,提高了系统运行效率,并且已接近了离线的识别性能,这将推动该算法的实用化.但本方法中对于噪声建模时,并没有考虑先验知识,并且对噪声模型的建模只采用了单高斯建模.如果将以上两方面扩展,必然还会提高识别系统的性能,这将是未来进一步研究的问题.

References:

[1] Baum L, Eagon J. An inequality with applications to statistical estimation for probabilistic functions of markov processes and to a model for ecology[J]. Bulletin of American Mathematical Society, 1967,73:360-363.

[2] Kim D Y,Un C K,Kim N S. Speech recognition in noisy environments using first-order vector taylor series[J]. Speech Communication,1998,24:39-49,.

[3] Moreno P J,Raj B,Stern R M. A vector taylor series approach for environment-independent speech recognition[C]. Proc. ICASSP, 1996,733-736.

[4] Ding G H,Wang X,Cao Y, et al. Sequential noise estimation for noise-robust speech recognition based on 1st-order VTS approximation[C]. In Proc. ASRU Workshop, 2005.

[5] Ding G H. Maximum a posteriori noise log-spectral estimation based on first-order vector taylor series expansion[J]. IEEE Signal Processing Letters, 2008,15:158-161.

[6] Du J,Huo Q. Feature compensation using high-order vector taylor series for noisy speech recognition[C]. Technical Memo, MSRA, January,2008.

[7] Gales M F. Model-based techniques for noise robust speech recognition[D]. Cambridge University,1995.