ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.18, No.4, April 2007, pp.878−883
DOI: 10.1360/jos180878

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于邻接空间的鲁棒语音识别方法*

严斌峰[1,2+], 朱小燕[1], 张智江[2], 张 范[2]

[1](清华大学 计算机科学与技术系,北京 100084)

[2](中国联合通信有限公司,北京 100032)

Robust Speech Recognition Based on Neighborhood Space

YAN Bin-Feng[1,2+], ZHU Xiao-Yan[1], ZHANG Zhi-Jiang[2], ZHANG Fan[2]

[1](Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

[2](China United Telecommunications Corporation, Beijing 100032, China)

+ Corresponding author: Phn: +86-10-66505147, E-mail: yanbf@chinaunicom.com.cn, http://www.chinaunicom. com.cn

Yan BF, Zhu XY, Zhang ZJ, Zhang F. Robust speech recognition based on neighborhood space. *Journal of Software*, 2007,18(4):878−883. http://www.jos.org.cn/1000-9825/18/878.htm

Abstract: This paper presents an approach to robust speech recognition based on neighborhood space, which can achieve performance robustness under mismatch between training and testing conditions. This approach uses neighborhood space of each underlying model to produce Bayesian predictive density as observation probability density. Experimental results show that the proposed method improves the performance robustness.

Key words: model space; acoustic model; speech recognition; Bayesian predictive density; pattern recognition

摘 要: 提出了一种基于邻接空间模型的鲁棒语音识别方法,解决测试集和训练集差别导致的识别正确率过低的问题.在以声学模型为中心的邻接空间中计算贝叶斯预测概率密度值,作为观察概率输出分值进行识别.实验表明,相对于传统语音识别方法,鲁棒识别方法在保证干净测试集的识别率没有很大下降的前提下,对含噪测试集的识别率获得了较大的提高.

关键词: 模型空间;声学模型;语音识别;贝叶斯预测密度;模式识别

中图法分类号: TP391 文献标识码: A

随着自动语音识别(automatic speech recognition,简称 ASR)技术的发展,语音识别在对话系统、语音控制系统等领域的应用越来越广泛.然而,在含噪真实环境的应用中,语音识别的正确率会有大幅度的下降,其原因主要在于存在着训练集和测试集的差别:训练语音一般在实验室环境下录制,环境噪声强度相对于在线语音识别时很小.因此,提高语音识别系统的鲁棒性是语音识别技术面临实用所必须解决的重要问题.

前端(front-end)的鲁棒语音识别主要研究鲁棒语音特征提取的信号处理方法以减少噪声的影响.在提取特征之前,先估计噪声的强度[1],采用谱减(spectral subtraction)方法降低噪声强度,然后提取出某种鲁棒的语音特征进行识别[2−5].后端(back-end)方法主要研究声学模型(一般是隐马尔可夫模型,简称 HMM)的模型补偿(model compensation)技术[6−10],减小测试集和训练集的差别带来的影响.这些方法都需要某些先验的知识给出衡量两

* Supported by the National Natural Science Foundation of China under Grant No.60272019 (国家自然科学基金)
Received 2004-02-02; Accepted 2005-08-24

者差别的评价机制——映射函数(mapping function),其损失参数(nuisance parameter)可以根据少量的训练和测试数据,以最大似然(maximum likelihood,简称 ML)或最大后验(maximum a posterior,简称 MAP)准则估计得到.然而在实际应用中,识别时只有测试语音和预先训练出的声学模型,并没有可用的先验知识,无法给出衡量测试集和训练集差别的评价机制的具体形式.

本文提出了一种基于邻接空间的后端鲁棒语音识别方法.首先,在模型空间中建立声学模型的邻接空间——以声学模型为中心的邻近区域,并给出邻接空间模型的参数表示;然后,采用贝叶斯预测算法[11−13],根据测试集语音噪声的强度决定邻接空间的大小,在邻接空间中计算出贝叶斯预测概率密度值(Bayesian predictive probability density),作为观察输出分值进行鲁棒的语音识别.实验表明,相对于传统的语音识别方法,该方法在保证干净语音测试集的识别率没有很大下降的前提下,对叠加高斯白噪声和有性别差异的测试集的识别率都获得了较大的提高.

本文第 1 节介绍邻接空间.第 2 节介绍基于邻接空间模型的鲁棒语音识别方法.第 3 节是实验结果和分析.第 4 节是结论.

1 邻接空间

假设语音识别器共有 N 个初始的声学模型,为$\{\lambda_i|1\le i\le N\}$,每一个初始声学模型可以看作模型空间$I$中的一个点.对于某个初始声学模型$\lambda$,定义模型空间中紧密围绕在点$\lambda$周围的区域为$\lambda$的邻接空间$\Lambda$,如图 1 所示.邻接空间是初始声学模型的鲁棒表达形式,由于训练时声学模型参数估计的误差、测试集和训练集的差别,相对测试集,最优(即“真实”)的模型参数与预先训练出的初始模型的参数之间存在一定的偏差;当邻接空间的大小设置适当时,在模型空间中“真实”的模型参数分布于初始声学模型对应的邻接空间中.

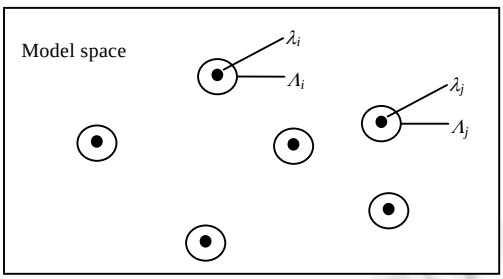

Fig.1 Initial acoustic model and its neighbourhood space in model space

图 1 模型空间中的初始声学模型及其邻接空间

基于连续 HMM 的语音识别系统中,L 个状态的初始声学模型λ的模型参数为$\{a,A,\theta\}$:$a=\{a_1,a_2,\ldots,a_L\}$是初始状态分布;$A=\{a_{ij}|1\le i,j\le L\}$是转移矩阵;$\theta=\{\theta_1,\theta_2,\ldots,\theta_L\}$是状态的观察输出概率密度函数,一般选择混合高斯密度函数的形式,$\theta_i=\{\zeta_{ik},M_{ik},\Sigma_{ik}\}$, $k=1,...,K$; $i=1,...,L$; $\sum_{k=1}^{K}\zeta_{ik}=1$,$K$ 为高斯混合的个数;ξ_{ik} 是第 k 个高斯函数的权重;M_{ik},Σ_{ik} 分别为第 k 个高斯函数的均值向量和协方差矩阵.假设输入的语音特征向量为 $O=\{O_1,O_2,\ldots,O_T\}$,对于语音向量 O_t,观察输出概率密度由公式(1)计算:

$$p(O_t\mid\theta_i)=\sum_{k=1}^{K}\zeta_{ik}\,p(O_t\mid\theta_{ik}) \tag{1}$$

$p(O_t|\theta_{ik})$是第 k 个高斯函数的输出概率密度值:

$$p(O_t\mid\theta_{ik})=\frac{1}{\sqrt{2\pi}\left|\Sigma_{ik}\right|^{1/2}}\exp\left\{-\frac{1}{2}(O_t-M_{ik})^T\sum\nolimits_{ik}^{-1}(O_t-M_{ik})\right\} \tag{2}$$

模型参数中最重要的是高斯函数的均值向量,因此,本文只考虑均值向量的变化,给出模型λ对应的邻接空间模型Λ的参数表示$\{a,A,\theta^*\}$:初始状态分布和转移矩阵与λ的相同;$\theta^*=\{\theta_1^*,\theta_2^*,...,\theta_L^*\}$, $\theta_i^*=\{\zeta_{ik},M_{ik}^*,\Sigma_{ik}\}$,

$k=1,...,K;\ i=1,...,L;\ \sum_{k=1}^{K}\zeta_{ik}=1$,混合权重和高斯函数的协方差矩阵不变,高斯函数的均值向量 M_{ik}^{*} 的第 d 维 M_{ikd}^{*} 在区间[M_{ikd}–C,M_{ikd}+C]内均匀分布,其中 C 为控制邻接空间大小的常量.

2 鲁棒语音识别

本文提出了一种基于邻接空间模型的后端鲁棒的语音识别方法:采用贝叶斯预测算法,在模型空间Γ中,以每一个声学模型λ为中心建立其对应的邻接空间,假设"真实"的λ参数均匀分布在邻接空间中,计算出贝叶斯预测概率密度值,作为观察输出分值进行鲁棒的识别.

2.1 贝叶斯预测决策

语音识别算法一般采用最大后验概率决策规则进行识别,识别结果 $\hat{\lambda}$ 满足:

$$\hat{\lambda}=\arg\max_{\lambda\in\Gamma}p(\lambda\mid O)=\arg\max_{\lambda\in\Gamma}\frac{p(O\mid\lambda)p(\lambda)}{p(O)}\cong\arg\max_{\lambda\in\Gamma}p(O\mid\lambda)\tag{3}$$

上式中,计算 $p(O|\lambda)$的λ真实分布是未知的,通常只能给出近似的分布,根据训练数据,采用参数估计的最优化准则(例如 ML,MAP 判别式训练等)估计出近似分布函数的参数.

贝叶斯预测决策方法对λ的近似分布与真实分布之间的误差进行预测,采用先验概率密度函数 $p(\lambda|\varphi)$预测λ的未知真实分布,其中:φ为超参数(hyper-parameter),以贝叶斯预测概率密度值 $\tilde{p}(O\mid\lambda)$ 近似为 $p(O|\lambda)$的真实值:

$$\tilde{p}(O\mid\lambda)=\int p(O\mid\lambda)p(\lambda\mid\varphi)\mathrm{d}\lambda\tag{4}$$

先验函数 $p(\lambda|\varphi)$的选择是贝叶斯预测决策方法的关键,一种 $p(\lambda|\varphi)$的定义[11−13]如下:

$$p(\lambda\mid\varphi)=\prod_{i=1}^{L}p(\theta_i\mid\varphi)=\prod_{i=1}^{L}\left(p(\zeta_{i1},...,\zeta_{iK}\mid\varphi)\prod_{k=1}^{K}p(M_{ik},\Sigma_{ik}\mid\varphi)\right)\tag{5}$$

其中,$p(\zeta_{i1},\ldots,\zeta_{iK}|\varphi)$和 $p(M_{ik},\ldots,\Sigma_{iK}|\varphi)$分别符合 Dirichlet 分布和 Normal-Wishart 分布[14].

本文只考虑 HMM 中状态混合高斯函数的均值向量分布的不确定性,为简便起见,假设真实模型参数在初始声学模型λ的邻接空间Λ内均匀分布,则

$$p(\lambda\mid\varphi)=\frac{1}{2C}\tag{6}$$

2.2 鲁棒识别

给定 HMM 参数λ,对于输入语音 $O=\{O_1,O_2,\ldots,O_T\}$,计算贝叶斯概率预测值:

$$\tilde{p}(O\mid\lambda)=\sum_{S}p(S)\left\{\prod_{t=1}^{T}\tilde{p}_{S_t=i}(O_t\mid\theta_i)\right\}\tag{7}$$

上式中的求和符号是对所有可能的状态序列 S 求和,其中状态上的贝叶斯概率预测值为

$$\tilde{p}(O_t\mid\theta_i)=\sum_{k=1}^{K}\zeta_{ik}\tilde{p}(O_t\mid\theta_{ik})\tag{8}$$

语音特征向量的各维之间彼此是独立的,因此,

$$\tilde{p}(O_t\mid\theta_{ik})=\prod_{d=1}^{D}\tilde{p}(O_{td}\mid\theta_{ikd})\tag{9}$$

其中,D 是语音特征向量的维数,根据式(4)、式(6),有

$$\begin{aligned}\tilde{p}(O_{td}\mid\theta_{ikd})&=\frac{1}{2C}\int_{M_{ikd}-C}^{M_{ikd}+C}\frac{1}{\sqrt{2\pi}\left|\Sigma_{ikd}\right|^{1/2}}\exp\left\{-\frac{(O_{td}-M_{ikd}^{*})^2}{2\Sigma_{ikd}}\right\}\mathrm{d}M_{ikd}^{*}\\&=\frac{1}{2C}\left\{\Phi\left(\frac{M_{ikd}-O_{td}+C}{|\Sigma_{ikd}|^{1/2}}\right)-\Phi\left(\frac{M_{ikd}-O_{td}-C}{|\Sigma_{ikd}|^{1/2}}\right)\right\}\end{aligned}\tag{10}$$

$\Phi(\cdot)$为误差函数:

$$\Phi(z)=\int_{-\infty}^{z}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-x^2/2}\mathrm{d}x \tag{11}$$

误差函数的近似计算公式如下:

$$\Phi(z)=\frac{1+erf\left(z/\sqrt{2}\right)}{2} \tag{12}$$

当 $z\geq 0$ 时,

$$\begin{aligned} erf(z)&=\int_{0}^{z}\frac{2}{\sqrt{\pi}}\mathrm{e}^{-x^2}\mathrm{d}x\\ &\approx 1-t\exp(-z^2-1.26551223+1.26551223t+0.37409196t^2+\\ &\quad 0.09678418t^3-0.18628806t^4+0.27886807t^5-1.13520398t^6+\\ &\quad 1.48851587t^7-0.82215223t^8+0.17087277t^9) \end{aligned} \tag{13}$$

$$t=\frac{1}{1+z/2} \tag{14}$$

否则,

$$erf(z)=-erf(-z) \tag{15}$$

该近似计算方法的最大误差为 1.2×10^{-7}[14].

对每一个声学模型分别计算 $\tilde{p}(O\mid\lambda)$,最后的识别结果 $\hat{\lambda}$ 满足:

$$\hat{\lambda}=\arg\max_{\lambda\in\Gamma}\tilde{p}(O\mid\lambda) \tag{16}$$

在线识别时,识别系统定期地计算录入语音的信躁比大小,根据经验自适应地调整邻接空间大小相关的 C 值,流程如图 2 所示.

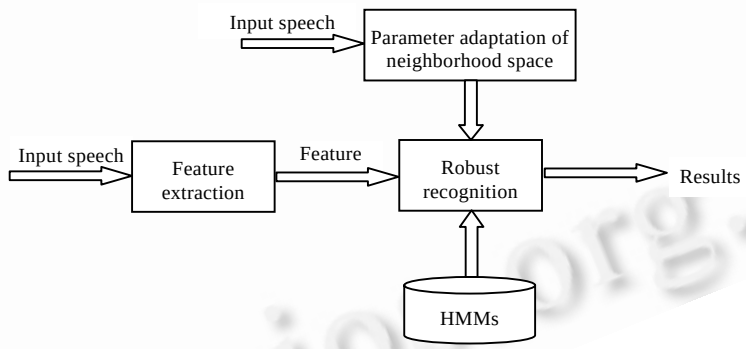

Fig.2 Flow of online robust speech recognition

图 2 在线鲁棒语音识别流程

3 实 验

3.1 实验数据

本文所采用的语音数据库是清华大学智能技术与系统国家重点实验室录制的 CIDS 语音数据库,其中的样本都是 11025Hz 采样率、16 位和单声道的语音信号.实验中使用 CIDS 数据库中的 60 个男声和 40 个女声的 13 22 个有调拼音语音,其中:前 40 个男声训练 411 个无调拼音的连续 HMM 模型;后 20 个男声作为干净语音测试集(与训练集的录制条件完全相同),40 个女声作为有性别差异的测试集,后 20 个男声的原始数据分别叠加不同信噪比的高斯白噪声作为有噪声的测试集.

识别时生成的 MFCC(Mel frequency cepstral coefficient)特征向量是 39 维的向量,分别是 12 维的 Mel 特征系数及其一二阶差分;一维的能量特征系数及其一二阶差分.

3.2 实验结果

本文分别设定不同的决定邻接空间大小的常量 C,对有性别差异的测试集和不同信噪比的含噪测试集进行识别,并与传统的语音识别(Baseline)方法进行了比较,实验结果见表 1.

Table 1 Experimental results

表 1 实验结果

C	Testing speech data (%)							
	Clean speech	Gender different speech	Contaminated speech at several SNR levels					
			−5	0	5	10	15	20
0.00 1	71.3	44.5	9.5	12.4	16.1	29.8	52.2	65.3
0.01	68.5	46.7	11.1	13.6	17.8	31.2	55.3	64.7
0.1	64.1	50.8	12.9	14.3	23.2	35.5	52.9	62.5
0.2	62.6	52.9	13.7	16.9	26.4	38.4	49.6	59.4
0.3	61.4	53.6	14.0	17.1	28.5	37.9	46.2	57.1
0.4	57.3	51.4	16.6	23.3	27.1	35.4	43.4	52.9
0.5	55.9	50.3	19.4	24.7	26.8	31.1	41.3	50.7
Baseline	73.7	44.1	8.6	11.5	15.7	29.5	48.2	59.8
Improved	−2.4	9.5	10.8	13.2	12.8	8.9	7.1	5.5

由于识别时模型参数相对于最大似然的训练模型参数作了细微的抖动,因此对于干净语音测试集来说,鲁棒语音识别方法相对于传统方法的识别率必然会下降,表 1 所示最小下降了 2.4%;而对于有性别差异的和不同信噪比的含噪语音的测试集,识别率有不同程度的提高,其中,对于不同信噪比的含噪语音的测试集,信噪比小(噪声大)的测试集识别率的提高相对更大,信噪比为 0 时,识别率的提高达到 13.2%.同时,随着信噪比的减小,获得最高识别率的鲁棒识别方法的参数 C 值越高,说明当测试集与训练集的差异越大时,真实声学模型参数均匀分布的邻接空间的范围也越大.对于干净的语音测试集,随着 C 的减小,鲁棒识别方法更接近传统方法的识别率;当 C 足够小时,邻接空间缩小为与其对应的初始声学模型,忽略计算误差,鲁棒识别方法与传统方法的识别率近似相同.

4 结 论

本文提出了基于邻接空间的鲁棒语音算法,采用贝叶斯预测算法,在模型空间中,以每一个声学模型为中心建立其对应的邻接空间,在此邻接空间中计算出贝叶斯预测概率密度值作为观察输出分值进行鲁棒的语音识别.实验表明,本文提出的鲁棒语音识别方法获得了较好的结果,相对于传统语音识别方法,对有性别差异和叠加高斯白噪声的测试集的识别率有较大的提高.

本文对克服稳定噪声环境下语音识别错误率增大问题,提出了有益的尝试.在下一步的工作中,应该针对不同类型的噪音环境,采用统计方法更加准确地探究相应高斯均值向量的实际分布;另外,在突变噪音环境下,高斯方差向量的实际分布也会较训练参数有一定的误差.因此,研究如何在模型层面上对方差进行一定的修正,也是本文后续工作中的课题.

References:

[1] Hirsch H, Ehrlicher C. Noise estimation technique for robust speech recognition. In: Proc. of the IEEE Int'l Conf. on Acoustic, Speech and Signal Processing. 1995. 153−156.

[2] Hermansky H, Tibrewala S, Pavel M. Towards ASR on partially corrupted speech. In: Proc. of the Int'l Conf. on Spoken Language Processing. 1996. 458−461. http://ieeexplore.ieee.org/xpls/abs_all.jsp?tp=&arnumber=607154

[3] Zhang XY, Wang F, Zheng F, Xu MX, Wu WH. Integrating sub-band information into feature extraction for robust speech recognition. Journal of Chinese Information Processing, 2002,16(1):19−24 (in Chinese with English abstract).

[4] Zhu QF, Alwan A. Non-Linear feature extraction for robust speech recognition in stationary and non-stationary noise. Computer Speech and Language, 2003,17(4):381−402.

[5] Yuo KH, Wang HC. Robust features for noisy speech recognition based on temporal trajectory filtering of short-time autocorrelation sequences. Speech Communication, 1999,28(1):13−24.

[6] Gauvain, JL, Lee CH. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. IEEE Trans. on Speech and Audio Processing, 1994,2(2):291−298.

[7] Gales, MJF, Young SJ. Robust continuous speech recognition using parallel model combination. IEEE Trans. on Speech and Audio Processing, 1996,4(5):352−359.

[8] Renevey P, Drygajlo A. Statistical estimation of unreliable features for robust speech recognition. In: Proc. of the IEEE Int'l Conf. on Acoustic, Speech and Signal Processing. 2000. 1731−1734.

[9] Leggetter CJ, Woodland PC. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. Computer Speech and Language, 1995,9(2):171−185.

[10] Sankar A, Lee CH. A maximum likelihood approach to stochastic matching for robust speech recognition. IEEE Trans. on Speech and Audio Processing, 1996,4(3):190−202.

[11] Huo Q, Lee CH. On-Line adaptive learning of the continuous density hidden Markov model based on approximate recursive Bayes estimate. IEEE Trans. on Speech and Audio Processing, 1997,5(2):161−172.

[12] Huo Q, Jiang H, Lee CH. A Bayesian predictive classification approach to robust speech recognition. In: Proc. of the IEEE Int'l Conf. on Acoustic, Speech and Signal Processing. 1997. 1547−1550.

[13] Jiang H, Hirose K, Huo Q. Robust speech recognition based on viterbi Bayesian predictive classification. In: Proc. of the IEEE Int'l Conf. on Acoustic, Speech and Signal Processing. 1997. 1551−1554.

[14] Vapnik V. The Nature of Statistical Learning Theory. 2nd ed., New York: Springer-Verlag, 2000.

附中文参考文献:

[3] 张欣研,王帆,郑方,徐明星,吴文虎.基于子带信息的鲁棒语音特征提取框架.中文信息学报,2002,16(1):19−24.

严斌峰(1977－),男,江西鹰潭人,博士,主要研究领域为人工智能,信号处理,通信技术.

朱小燕(1957－),女,教授,博士生导师,CCF高级会员,主要研究领域为人工智能,人工神经网络,模式识别,人机交互.

张智江(1963－),男,博士,教授级高工,主要研究领域为通信技术.

张范(1961－),男,教授级高工,主要研究领域为通信技术.