ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.17, No.12, December 2006, pp.2547−2553
DOI: 10.1360/jos172547

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

语音识别确认中的置信特征和判定算法*

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)

Confidence Measures and Integrating Algorithm in Utterance Verification

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, Fax: +86-10-66504252, E-mail: yanbf@chinaunicom.com.cn

Yan BF, Zhu XY, Zhang ZJ, Zhang F. Confidence measures and integrating algorithm in utterance verification. *Journal of Software*, 2006,17(12):2547−2553. http://www.jos.org.cn/1000-9825/17/2547.htm

Abstract: In this paper, an approach is presented for integrating several confidence measures to verify utterance based on support vector machine (SVM). Segmental filler-based posterior probability parameters and linear predictive coding (LPC) recognition difference measure are derived from the verified utterance. An SVM classifier is trained to integrate several confidence measures to make final decision. Experimental results show that confidence measures and the SVM classifier are effective for utterance verification.

Key words: confidence measure; utterance verification; speech recognition; support vector machine

摘 要: 提出了一种基于支持向量机的联合多种置信特征进行语音识别确认的判定方法.从待确认语音中提取出分段的后验概率和线性预测编码识别结果置信特征,其中后验概率根据垃圾模型近似计算得到;设计支持向量机分类器联合多种置信特征给出最终确认结果.实验结果表明,所提出的置信特征和支持向量机分类器取得了很好的确认效果.

关键词: 置信特征;语音识别确认;语音识别;支持向量机

中图法分类号: TP301 文献标识码: A

随着自动语音识别技术的发展,语音识别系统的应用越来越广泛.但是,语音识别系统的性能仍然不能满足实际应用的需要,识别结果中存在很多错误识别的词,语音识别系统在得到初步的识别结果以后,必须通过后处理过程,正确评价识别结果的可靠性,检出并拒识非语音的噪声、词表外词和识别错误.

语音识别算法一般利用最大后验概率决策规则进行识别,得到的识别结果满足

$$\hat{W} = \arg\max_{w} p(W \mid X) = \arg\max_{w} \frac{p(X \mid W)p(W)}{p(X)} = \arg\max_{w} p(X \mid W)p(W) \quad (1)$$

其中 X 是输入语音特征向量;W 是词表中的某一个词;$p(W)$是根据语言模型统计出的词 W 出现的先验概率;$p(X)$

* Supported by the National Natural Science Foundation of China under Grant No.60272019 (国家自然科学基金)
Received 2004-02-09; Accepted 2005-08-24

是特征向量 X 出现的先验概率.语音识别的任务是比较词表中所有词模型的后验概率 $p(W|X)$的相对大小进行决策.给定 X,对于所有词模型,$p(X)$是常量.在比较词模型后验概率的相对大小时,$p(X)$忽略不计,即识别器给出的识别结果是词表中相对最匹配的词,而不是置信度足够大的词.在实际的语音识别系统中,后处理的语音确认过程给出识别候选结果的置信度水平,并根据置信度大小接受或拒绝候选结果.

在理论上,当给定 X、识别结果为 $\hat{W}$ 时,根据后验概率 $p(\hat{W}|X)$ 就可以评价识别结果的可靠性,后验概率本身就是语音确认中很好的置信特征.然而在计算后验概率时,较准确地估算 $p(X)$是非常困难的.all-phone 方法[1]根据所有模型累计计算得到 $p(X)$,这种方法在计算和时间上的开销非常大.catch-all 方法[2]去掉搜索空间的语法限制,任何一个词可以连接其他所有的词,任意一个词序列都可以被识别出来,搜索出的最佳路径的概率似然值近似为 $p(X)$;lattice-based 方法[3]在词图中通过前后向算法近似计算 $p(X)$;N-best 方法[4]仅根据识别结果的前 N 个词候选近似计算 $p(X)$.但是,这些方法计算得都不够精确.

本文提出基于垃圾模型的快速近似计算 $p(X)$的方法,给出代表置信度水平的后验概率值:借鉴模型参数共享的思想,以基于分散度的距离标准,采用改进的合并分级聚类算法[5]对所有音节进行聚类,每一类音节对应的样本训练一个垃圾模型,根据较少的垃圾模型快速地近似计算 $p(X)$.基于分散度距离标准的选择,可以更充分地利用有限的训练样本,对合并分级聚类算法的改进可以避免各分类结果中样本数过于不均的问题.考虑到汉语发音的特点,即大部分音节分为声母和韵母两部分,根据能量将待确认音节分为两个子段,对每个子段分别计算后验概率的统计值作为置信特征,以强调各个局部的匹配.

另外,本文选择有别于识别阶段所用语音特征的其他特征进行二次识别,根据两次识别的结果是否相同给出置信度水平.识别阶段使用反映人类听觉特性的 Mel 频率倒谱系数(Mel frequency cepstral coefficient,简称 MFCC)特征[6].然后从待确认语音段的原始录入数据中提取出 LPC(linear predictive coding)特征[7],在 LPC 特征空间对待确认语音段进行二次识别,以与 MFCC 特征空间识别结果是否一致作为 LPC 识别置信特征.

在得到多种置信特征以后,必须联合多种置信特征给出最后的确认结果.语音确认要解决的实际上是一个分类问题,在第 1 步识别阶段给出待确认的识别结果以后,第 2 步中的确认过程将待确认的识别结果分成两类(“对”或“错”).线性 Fisher 判决方法[8]计算得到最优的分类线的权向量后,利用先验知识在分类线上选择分界阈值点进行分类.前馈神经网络方法[9]联合多种置信特征进行确认,以每一种置信特征作为网络的每一维输入,网络的输出为[0,1]区间的一个实数,表示待确认结果正确的概率.

传统统计学研究的内容是样本无穷大时的渐进理论,即当样本数据趋于无穷多时的统计性质.而实际问题中样本数据往往是有限的.在有限训练样本集条件下,基于统计学习结构风险最小理论的 SVM(support vector machine),应用在语音识别中,获得了较好的结果[10].本文在语音确认过程中采用 SVM 方法联合多种置信特征进行判决,并与 Fisher 线性判决和神经网络方法进行了比较.实验结果表明,SVM 判定算法获得了更好的确认性能.

1 置信特征

理论上,给定的语音特征序列 X,词模型 W 所对应的后验概率 $p(W|X)$本身就是词 W 的置信特征 $C_{MAP}(W|X)$.根据贝叶斯公式

$$C_{MAP}(W|X)=p(W|X)=\frac{p(W)p(X|W)}{p(X)} \tag{2}$$

在实际的语音识别系统中,比较词模型后验概率彼此之间的相对大小时,$p(X)$一般都忽略不计.但是,在计算代表置信度水平的后验概率时,必须计算 $p(X)$.

all-phone 方法对词表中所有模型累计计算 $p(X)$:

$$p(X)=\sum_{W}p(W)p(X|W) \tag{3}$$

这种方法的计算量非常大,尤其是在大词表的连续语音识别中,必须计算每一个词模型 W 对应的概率值 $p(X|W)$,不能进行有效的剪枝以加快搜索速度.

为了减少计算量,本文提出基于垃圾模型的快速近似计算 $p(X)$的方法,给出代表置信度水平的后验概率值.首先,借鉴模型参数共享的思想,以基于分散度的距离标准,采用改进的合并分级聚类算法[5]对所有音节进行聚类,每一类音节对应的样本训练一个垃圾模型 F,从而大规模地减小声学空间中的模型数量;然后,根据较小的垃圾模型集快速计算 $p(X)$:

$$p(X)=\sum_{F}p(F)p(X\mid F) \tag{4}$$

根据系统的应用领域和聚类结果,可以统计出先验概率 $p(W)$和 $p(F)$,近似认为所有词的 $p(W)$相同,式(2)可以改写为

$$C_{MAP}(W\mid X)=\frac{p(W)p(X\mid W)}{\sum_{F}p(F)p(X\mid F)}=\frac{p(X\mid W)}{\sum_{F}p(X\mid F)N_F} \tag{5}$$

其中,N_F 是垃圾模型 F 对应的类别中含有的音节数目.

在本文中,词模型和垃圾模型都采用连续的隐马尔可夫模型(constrained hidden Markov model,简称 CHMM).对每一帧特征向量分别计算帧级别的对数后验概率置信特征,即正规化对数似然得分(normalize log-likelihood,简称 NLL)[2]:

$$C_{NLL}(W\mid x_t)=\log p(x_t\mid\lambda_W)-\log\sum_{F}p(x_t\mid\lambda_F)N_F \tag{6}$$

其中λ_W,λ_F 分别为词 W 和垃圾模型 F 的 HMM 模型.

根据待确认语音每一帧的置信特征,分别统计出下列置信分数:

- 算术平均置信特征

$$C_{AM}(W\mid X)=\frac{1}{N}\sum_{t=1}^{N}C_{NLL}(W\mid x_t) \tag{7}$$

- 几何平均置信特征

$$C_{GM}(W\mid X)=\left(\prod_{t=1}^{N}\left|C_{NLL}(W\mid x_t)\right|\right)^{1/N} \tag{8}$$

- 标准差置信特征

$$C_{AD}(W\mid X)=\frac{1}{N}\sum_{t=1}^{N}\left|C_{NLL}(W\mid x_t)-C_{AM}(W\mid X)\right| \tag{9}$$

- 最小值置信特征

$$C_{\mathrm{MIN}}(W\mid X)=\min_{t}C_{NLL}(W\mid x_t) \tag{10}$$

其中,N 为输入语音特征 X 的帧数,C_{AM} 反映模型对整个待确认序列匹配的平均情况,C_{GM} 强调待确认序列中置信度水平较低部分的影响,标准差 C_{AD} 体现置信分数的变化,C_{MIN} 反映待确认序列中置信度水平最低的置信分数.为了进一步考察模型与待确认语音段局部的匹配情况,考虑到汉语发音的特点,大部分音节分为声母和韵母两部分,声母部分的能量较低,因此,根据能量的变化将待确认语音分为两段,对声母和韵母部分统计出 C_{AM},C_{GM},C_{AD},C_{MIN} 置信特征,分别进行确认.

另外,本文从其他途径给出置信度水平的评价方法——LPC 识别置信特征.在识别阶段,使用的是 MFCC 特征,MFCC 在一定程度上反映了人类的听觉特性.在确认阶段,使用不同于 MFCC 的特征进行识别,根据两种特征下识别结果是否匹配来确定是否接受待确认的语音.在两种完全不同的特征空间上对待识别的模式进行划分,如果两者的分类结果一致,则认为对模式的划分是正确的.LPC 模拟了人类发声声道的特点,选择 LPC 特征对待确认语音进行二次识别,如果两种特征下识别结果一致,则 LPC 置信特征为 1,否则为 0.

2 判定算法

在得到多种置信特征以后,所面临的问题就是如何利用这些置信度信息,联合多种置信特征给出最后的确认结果.

2.1 Fisher线性判决

Fisher 线性判决方法把高维空间的样本投影到一条最优分类线上,在一维空间上对高维空间中的样本进行分类判决.首先,从训练语音样本中提取出置信特征正例矢量和反例矢量,然后根据这些矢量计算得到最优分类线的方向,权向量 $\vec{p}$.在识别时,对待确认词的高维置信特征矢量 $\vec{C}$ 进行如下线性投影,得到一维的置信分数 r,

$$r = \vec{p}^T \cdot \vec{C} \tag{11}$$

根据置信分数 r,计算决策后验概率值 c

$$c = \frac{p(r\,|\,\text{correct})\,p(\text{correct})}{p(r\,|\,\text{incorrect})\,p(\text{incorrect})} \cong \frac{p(r\,|\,\text{correct})}{p(r\,|\,\text{incorrect})} \tag{12}$$

在训练样本中,正例矢量和反例矢量的置信分数 r 的分布符合高斯分布.分别对正例矢量和反例矢量建立高斯概率密度函数模型,根据训练样本统计出两个高斯函数的均值和方差,以计算 $p(r|\text{correct})$和 $p(r|\text{incorrect})$.

2.2 神经网络

本文采用的是一个 3 层的前馈神经网络,如图 1 所示.

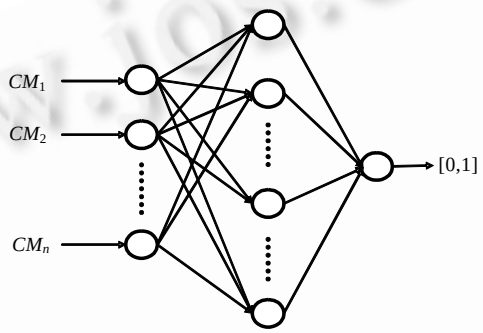

Fig.1 Architecture of neural network

图 1 神经网络结构

在训练时,正例置信特征矢量的输出为 1,反例置信特征矢量的输出为 0.在识别时,网络的输出值为[0,1]间的实数,此分值可以看作待确认语音正确的概率值.在训练学习过程中,使用均方误差和标准反向传播算法.输出函数采用 Sigmoid 函数:

$$f(x) = \frac{2}{1+\mathrm{e}^{-2x}} - 1 \tag{13}$$

2.3 支持向量机

SVM 是建立在统计学习理论的 VC 维理论和结构风险最小原理基础上的学习算法,根据有限的样本信息,在模型的复杂性和学习能力之间寻求最佳折衷,以期获得最好的推广能力.SVM 使用结构风险最小化来寻找最优分类面,在数学上证明了这等价于寻找最小真实风险.所以,支持向量机在有限样本条件下的推广能力很好.

SVM 是从线性可分情况下的最优分类面发展而来的.所谓最优分类线就是要求分类线不但能将两类正确分开(训练错误率为 0),而且使分类间隔最大.分类线方程为 $x\cdot w+b=0$,我们可以对其进行归一化,使得对线性可分的样本集$(x_i,y_i)(i=1,\ldots,n,x\in R^d,y\in\{+1,-1\})$,满足 $y_i[(w\cdot x_i)+b]-1\geq 0, i=1,\ldots,n$.此时,分类间隔等于 $2/||w||$,使间隔最大等价于使$||w||^2$ 最小.满足上述条件且使分类间隔最小的分类面叫做最优分类面,分类线上的训练样本点称作支持向量.

利用 Lagrange 优化方法可以把上述最优分类面问题转化为其对偶问题,即在约束条件 $\sum_{i=1}^{n} y_i\alpha_i = 0$ 且$\alpha_i\geq 0$, $i=1,\ldots,n$ 下,对α_i 求解下列函数的最大值:

$$Q(\alpha)=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j(x_i\cdot x_j) \tag{14}$$

容易证明,解中将只有少部分α_i不为 0,对应的样本就是支持向量.解上述问题后得到的最优分类函数是

$$f(x)=\mathrm{sgn}\{(w\cdot x)+b\}=\mathrm{sgn}\left\{\sum_{i=1}^{n}\alpha_i^* y_i(x_i\cdot x)+b^*\right\} \tag{15}$$

上式中的求和符号只对支持向量进行,b^*是分类阈值.

对于非线性问题,可以通过非线性变换将非线性问题转化为某个高维空间中的线性问题,在高维空间中求得最优分类面.采用适当的核函数就可以实现非线性问题某一变换后的线性分类,其计算复杂度没有增加.本文选择多项式内积函数为核函数

$$K(x_i,x_j)=\left(\frac{1}{256}(x_i\cdot x_j)+1\right)^3 \tag{16}$$

3 实 验

本文通过等错误率(equal error rate,简称 EER)来评价系统的确认性能.语音确认阶段存在两种错误:错误拒绝(false rejection)和错误警报(false alarm).错误拒绝率和错误警报率分别为

$$p(\text{错误拒绝})=\frac{\text{错误拒绝的词数}}{\text{识别器正确识别的词数}} \tag{17}$$

$$p(\text{错误警报})=\frac{\text{错误接受的词数}}{\text{识别器错误识别的词数}} \tag{18}$$

通过调整确认阈值,可以改变错误拒绝率和错误警报率,当两者相同时,得到等错误率 *EER*,

$$EER=p(\text{错误拒绝})=p(\text{错误警报}) \tag{19}$$

在垃圾模型的训练中,411 个无调拼音音节共分为 9 类,每一类音节对应的样本训练出一个垃圾模型,以 9 个垃圾模型来描述声学空间,计算后验概率置信特征的计算量小于 all-phone 方法.在对数据进行分类测试时,如果某个拼音在聚类时属于某类,且识别结果也为该类对应的垃圾模型,则认为是一个正确分类,统计此正确率作为垃圾模型好坏的评判标准.表 1 是音节聚类结果,表 2 是垃圾模型的识别结果.

Table 1 Result of syllable cluster

表 **1** 音节分类结果

Cluster	1	2	3	4	5	6	7	8	9
Samples' quantity	66	59	43	28	68	41	62	20	24

Table 2 Recognition result of filler model

表 **2** 垃圾模型的识别结果

Candidates quantity	1	2	3	4	5
Accuracy (%)	90.1	97.6	99.0	99.6	99.8

从音节的分类识别的测试结果看,前两选的结果达到 97.6%,这个聚类结果已经完全满足构建垃圾模型的需求.

这里比较 all-phone,*N*-best 方法与基于垃圾模型(filler-based)的后验概率计算方法的确认性能.其中,*N*-best 后验概率计算方法见式(20),*N*=9.

$$C_{MAP-Nbest}=\frac{p(X\mid W)}{\sum_{n=1}^{N}p(X\mid W_n)} \tag{20}$$

表 3 是上述 3 种方法统计的各种后验概率置信特征和本文提出的 LPC 识别结果置信特征的确认性能比较.Filler-based 方法的确认性能要优于常用的 *N*-best 方法.虽然 all-phone 方法的确认性能最好,但其计算量太大,且不能用于搜索过程中需要剪枝的连续语音识别.另外,本文提出的 LPC 识别结果置信特征也获得了较好的确

认性能,具体分析 MFCC 和 LPC 两种特征的识别结果可以发现:对于 MFCC 特征识别错误的大多数音节,LPC 特征识别结果也是错误的.但是,两者的识别结果并不一致.此时,LPC 置信特征可以正确地拒识 MFCC 特征的错误识别结果,降低错误警报率;对于那些不易混淆的音节,两种特征的识别结果大致是正确的,从而防止大量错误拒绝情况的发生.

Table 3 Utterance verification EER by several confidence measures (%)

表 3 各种置信特征确认性能 EER(%)

Confidence measure	CAM	CGM	CAD	CMIN
all-phone	20.3	23.4	29.5	31.6
N-best	27.1	29.7	36.1	39.8
Filler-Based	22.7	23.9	31.6	34.1
LPC measure		19.7		

我们分别采用 Fisher 线性判决方法、神经网络方法和 SVM 方法联合多种置信特征进行确认,实验结果见表 4.

Table 4 Verification result of combination decision method

表 4 联合判定方法的确认性能

Classifier	Fisher	NN	SVM
EER (%)	14.9	13.6	12.2

表 4 中,SVM 联合置信分数的确认性能要优于 Fisher 线性判决和神经网络方法.作出操作点特征曲线图(receiver operating characteristic,简称 ROC),进一步地评价 3 种联合判定算法的确认性能.如图 2 所示,在相同的错误警报率条件下,SVM 方法的错误拒绝率最低.

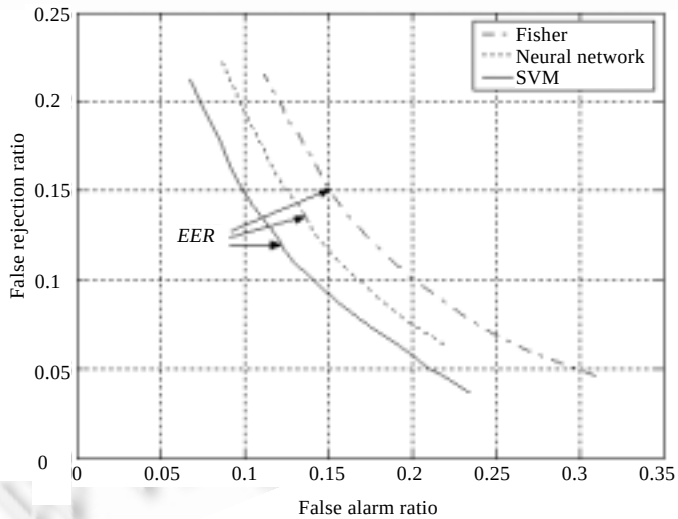

Fig.2 The ROC curve of three classifiers

图 2 联合判定方法的 ROC 曲线图

4 结 论

本文提出了基于垃圾模型的后验概率置信度的快速近似计算方法,基于分散度距离标准的改进合并分级聚类算法训练出了较好的垃圾模型.另外,我们还提出了 LPC 识别置信特征.得到多种置信特征以后,采用 SVM 方法联合多种置信特征进行语音确认.实验结果表明,本文提出的置信特征获得了较好的确认性能;并且,SVM 联合判定算法要优于 Fisher 线性判决和神经网络方法.

References:

[1] Young YS. Detecting misrecognitions and out-of-vocabulary words. In: Proc. of the IEEE Int'l Conf. on Acoustic, Speech and Signal Processing, Vol 2. Adelaide: IEEE, 1994. 21−24.

[2] Kamppari SO, Hazen TJ. Word and phone level acoustic confidence scoring. In: Proc. of the IEEE Int'l Conf. on Acoustic, Speech and Signal Processing, Vol 3. 2000. 1799−1802. http://groups.csail.mit.edu/sls/publications/2000/Simo-ICASSP.pdf

[3] Wessel F, Schluter R, Macherey K, Ney H. Confidence measures for large vocabulary continuous speech recognition. IEEE Trans. on Speech and Audio Processing, 2001,9(3):288−298.

[4] Wessel F, Schluter R, Ney H. Using posterior word probabilities for improved speech recognition. In: Proc. of the IEEE Int'l Conf. on Acoustic, Speech and Signal Processing, Vol 3. 2000. 1587−1590.

[5] Yan BF, Zhu XY. Continuous speech recognition and verification based on a combination score. Journal of Software, 2003,14(12): 2014−2020 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/14/2014.htm

[6] Schroeder M. Direct (nonrecursive) relations between cepstrum and predictor coefficients. IEEE Trans. on Acoustics, Speech and Signal Processing, 1981,29(2):297−301.

[7] Markel J, Gray A. Linear prediction of speech. Proc. of the IEEE, 1978,66(2):266−267.

[8] Hazen T, Seneff S, Polifroni J. Recognition confidence scoring and its use in speech understanding systems. Computer Speech and Language, 2002,16:49−67.

[9] Xiong ZY, Wu WH, Xu MX. Comparison and integration of confidence measure calculation method. In: Proc. of the 6th National Conf. on Man-Machine Speech Communication. 2001. 447−449 (in Chinese with English abstract).

[10] Cortes C, Haffner P, Mohri M. Weighted automata kernels-general framework and algorithms. In: Proc. of the 9th European Conf. on Speech Communication and Technology. 2003. 989−992.

附中文参考文献:

[5] 严斌峰,朱小燕.基于联合得分的连续语音识别确认方法.软件学报,2003,14(12):2014−2020. http://www.jos.org.cn/1000-9825/14/2014.htm

[9] 熊振宇,吴文虎,徐明星.置信度计算方法的比较和结合.见:徐明星编.第 6 届全国人机语音通讯学术会议.2001.447−449.

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,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.