第四届全国人机语音通讯学术会议(NCMMSC-96)论文集，pp.32-35，1996 年 10 月，北京 (本文获本届学术会议优秀论文二等奖)

汉语语音听写机中的语音识别基元

郑　方　　吴文虎　　方棣棠
(清华大学计算机科学与技术系 100084)
(fzheng@sp.cs.tsinghua.edu.cn, (010)62594141)

摘要：语音听写机中语音模型的好坏直接影响到语言模型和听写机的性能，而语音识别基元的选取又会直接影响到语音模型的灵活性和鲁棒性。本文在一个大型数据库上对基于中心距离连续概率模型(CDCPM)的语音模型进行了大量关于语音识别基元的比较实验。实验表明，音节应是识别基元的首选。本文同时还对基于最近邻(NN)原则的输出观察向量计分方法进行了实验，取得了很好的效果。

关键词：识别基元，中心距离连续概率模型(CDCPM)，最近邻(NN)原则

1　引言

在语音听写机中，语音模型和语言模型是两个非常重要的部分。而语音模型的好坏会直接影响到语言模型性能的发挥，决定着听写机的质量。语音模型的一个基本问题是选取什么样的识别基元，本文分别以音素、声韵母、音节为识别基元进行比较实验。本文还对描述特征参数空间的类混合 Gauss 密度(MGD)[1] 和特征向量的最近邻(NN)原则计分方法进行了比较。

2　特征参数

在实验中，我们使用的特征参数是 D=16 阶倒谱系数$\{c_d\}_{d=1}^{D}$[2][3]及其线性回归系数$\{r_d\}_{d=1}^{D}$[4]，其中线性回归分析宽度为 5 帧。为了研究方便，我们把 D 阶倒谱系数$\{c_d\}_{d=1}^{D}$和 D 阶回归系数$\{r_d\}_{d=1}^{D}$都看成是 D 维欧氏空间中的向量，同时定义该空间上的两个向量 $\mathbf{c}_1$ 和 $\mathbf{c}_2$ 之间、 $\mathbf{r}_1$ 和 $\mathbf{r}_2$ 之间的距离度量为加权欧氏距离[5]：

$$y(\mathbf{c}_1,\mathbf{c}_2)=\sqrt{\sum_{d=1}^{D}w_d(c_{1d}-c_{2d})^2},\quad y(\mathbf{r}_1,\mathbf{r}_2)=\sqrt{\sum_{d=1}^{D}w_d(r_{1d}-r_{2d})^2} \tag{1}$$

其中 $\mathbf{w}=(w_1,w_2,\ldots,w_D)$ 为权向量。在实验中，权向量的分量等于倒谱特征向量相应维统计方差的倒数，这样使得在距离度量中各维对距离的贡献在统计意义上相等。

在进行了倒谱分析和回归分析以后，每一帧发音对应两个特征向量，即倒谱特征向量 $\mathbf{c}$ 及回归特征向量 $\mathbf{r}$ 。这两种特征有两种使用方式。一是把它们组合为一个 *D*2* 维大向量:

$$\mathbf{v}=(\alpha^{(c)}\mathbf{c},\alpha^{(r)}\mathbf{r}) \tag{2}$$

其中 $\alpha^{(c)}$ 和 $\alpha^{(r)}$ 分别是倒谱特征向量和回归特征向量在大向量中的平衡系数。在这样的情况下，*D*2* 维空间中向量 $\mathbf{v}_1$ 和 $\mathbf{v}_2$ 的距离度量为:

$$y(\mathbf{v}_1,\mathbf{v}_2)=\sqrt{\left[\alpha^{(c)}y(\mathbf{c}_1,\mathbf{c}_2)\right]^2+\left[\alpha^{(r)}y(\mathbf{r}_1,\mathbf{r}_2)\right]^2} \tag{3}$$

二是把两种特征看做是两个独立的部分，它们可以分别聚类和分别用不同的概率密度函数(PDF)进行描述。若记第 n 状态的倒谱特征的 PDF 为 $b_n^{(c)}(\mathbf{c})$，回归特征的 PDF 为 $b_n^{(r)}(\mathbf{r})$，

那么第 n 状态的输出向量 $\mathbf{v}=(\alpha^{(c)}\mathbf{c},\alpha^{(r)}\mathbf{r})$ 的匹配得分为：

$$b_n(\mathbf{v})=b_n^{(c)}(\mathbf{c})*b_n^{(r)}(\mathbf{r}) \tag{4}$$

3 语音模型

实验中使用的是基于中心距离正态 CDN(Center Distance Normal)分布的中心距离连续概率模型 CDCPM [5][6]。CDN 分布指服从正态分布的随机向量 $\boldsymbol{\xi}$ 离开其均值向量 μ_x 的距离所服从的分布，$\boldsymbol{\xi}$ 的伪概率密度函数(pseudo-PDF)形式为：

$$p(\boldsymbol{x};\mu_{\mathbf{x}},\mu_y)=\frac{2}{\pi\mu_y}\exp(-y^2(\boldsymbol{x},\mu_{\mathbf{x}})/\pi\mu_y^2)\overset{def}{=}\mathrm{N}_{CD}(\boldsymbol{x};\mu_{\mathbf{x}},\mu_y) \tag{5}$$

其中 μ_y 为 $\boldsymbol{\xi}$ 离开其均值向量 μ_x 的距离的均值。注意 $\mathrm{N}_{CD}(\boldsymbol{x};\mu_{\mathbf{x}},\mu_y)$ 不是 $\boldsymbol{\xi}$ 的 PDF，而是 $\boldsymbol{\xi}$ 与 $\mu_{\mathbf{x}}$ 之间的加权欧氏距离的 PDF，这里只是一种简单的记法。CDN 分布中的参数 $\mu_{\mathbf{x}}$ 及 μ_y 很容易从训练样本向量中估计出来。

关于特征参数分布的描述人们已经有很多的研究。为了更好地刻画特征空间的分布，最常用的描述方法是混合 Gauss 密度 MGD[1]和 Tied 混合 Gauss 密度[7]两种。对 CDN 分布，我们可采用类 MGD(即混合 CDN 分布)的描述方法：

$$\begin{aligned} b_n^{(c)}(\mathbf{c})&=\sum_{m=1}^{M^{(c)}}g_{nm}^{(c)}\mathrm{N}_{CD}(\mathbf{c};\mu_{xnm}^{(c)},\mu_{ynm}^{(c)})\\ b_n^{(r)}(\mathbf{r})&=\sum_{m=1}^{M^{(r)}}g_{nm}^{(r)}\mathrm{N}_{CD}(\mathbf{r};\mu_{xnm}^{(r)},\mu_{ynm}^{(r)}) \end{aligned} \tag{6}$$

其中 $1\le n\le N$(状态数), $1\le m\le M$(混合数), $1\le d\le D$(维数)，$b_n^{(c)}(\mathbf{c}),b_n^{(r)}(\mathbf{r})$ 分别表示第 n 状态的倒谱系数及其自回归系数的 PDF 函数。

Bayes 学习算法[8]同样可以用于 CDCPM，但在这样的描述中，需要通过训练得出 $g_{nm}^{(c)}$ 和 $g_{nm}^{(r)}$。我们在用聚类的方法得到(6)式中的参数后，在实验中尝试了基于最近邻(NN)准则的另外一种输出观察向量计分形式:

$$\begin{aligned} b_n^{(c)}(\mathbf{c})&=\max_{1\le m\le M^{(c)}}\mathrm{N}_{CD}(\mathbf{c};\mu_{xnm}^{(c)},\mu_{ynm}^{(c)})\\ b_n^{(r)}(\mathbf{r})&=\max_{1\le m\le M^{(r)}}\mathrm{N}_{CD}(\mathbf{r};\mu_{xnm}^{(r)},\mu_{ynm}^{(r)}) \end{aligned} \tag{7}$$

这样的实验取得了更好的效果。

4 识别单元的选择

汉语语音有 400 多个无调音节和四声[9]，共有 1300 多个有调音节。每个音节由声母和韵母组成，其中声母(包括零声母)22 个，韵母 38 个。汉语的韵母还可以更细地分为 19 个音素。连续语音识别中识别单元的选择就基于这样的语音学基础。

实验中对识别单元的选取进行了三种尝试：①按音素：19 个元音音素和 22 个辅音音素；②按声韵：22 个声母和 38 个韵母；③按音节：419 个无调音节。

实验表明，识别单元选取得越小，模型时空开销越小，灵活性越大，但训练数据库的标定就越困难，对语速和音变的适应也越差；而识别单元选得越大，灵活性就越小，但对语速和音变的适应越好。两者之间存在一个折中的选择。

5 实验

实验所使用的数据库是由年龄在 18 到 25 岁之间、来自几乎全国各个省份的 80 个人(40 个男声和 40 个女声)录制的，其中每位发音人的发音有：单音节(11 组×100 个)；二字词(63 组×100 个)；三字词(11 组×100 个)；四字词(10 组×100 个)；五字词(1 组×76 个)；六字词(1 组×23 个)；七字词(1 组×10 个) 。

该数据库中每五位的词表组成一个大词表，因此 80 人的发音实际上共有 80/5 =16 遍。每一遍的大词表中，音节的安排并不是均匀的，而是决定于它们在汉语词典列出的所有词中出现的次数。除此之外，每人再念 1 组×10 个各不相同的句子。

说话人要求按普通话发音，但可以带一点口音，而且背景有一定的噪音。这样做的目的是为了制作一个现实世界的数据库，使模型的研究更面向实用系统。

数据库的量很大，共有 25GB 左右，约合 230 小时的语音数据。

我们在这个大型数据库上男性发音人的单音节数据进行了 46 组比较实验，限于篇幅，本文只列出几组，但所有组的实验结果都支持下面所给出的结论。下表中，如无特殊说明，状态数 N=6，混合数 M=4，维数 D=16。前 20 人的数据构成训练集，后 20 人的数据构成识别集，所列结果为两集测试的平均。

1、输出观察向量计分方法的比较：识别基元为韵母；特征参数仅有倒谱系数。结论是基于 NN 原则的计分结果远好于基于类 MGD 的计分结果。

表 1　观察向量计分方法的比较

前 n 名	1	2	3	4	5	6	7	8	9	10	12	18
NN	78.22	91.48	95.52	97.54	98.53	99.12	99.40	99.59	99.70	99.77	99.9	n/a
MGD'	70.22	86.18	91.61	94.48	96.21	97.45	98.24	98.79	99.16	99.35	n/a	99.9

2、倒谱系数和回归系数的组合方式比较：识别基元为韵母；计分方法为 NN。结论：倒谱系数 CEP 与其回归系数 ARCEP 单独聚类、单独计分效果好得多。

表 2　倒谱系数和回归系数的组合方式比较

前 n 名		1	2	3	4	5	6	7	8	9	10	11	12
分	1	80.23	92.83	96.56	98.11	98.82	99.28	99.48	99.61	99.73	99.80	99.9	n/a
别	2	79.19	92.49	96.17	97.90	98.77	99.22	99.46	99.62	99.74	99.81	99.9	n/a
整	1	78.47	91.79	95.80	97.77	98.71	99.18	99.44	99.62	99.74	99.81	99.9	n/a
体	2	78.06	91.47	95.56	97.48	98.51	99.13	99.43	99.61	99.71	99.77	n/a	99.9

分别 1：CEP 和 ARCEP 分别聚类，用(6)式分别计分；

分别 2：ARCEP 的聚类类别服从 CEP 的聚类，用(6)式分别计分；

整体 1：按大向量聚类，$\alpha^{(c)}=\alpha^{(r)}=1$，规整化因子 G=0.316 [5]；

整体 2：按大向量聚类，$\alpha^{(c)}=\alpha^{(r)}=1$，规整化因子 G=0.1 [5]；

3、识别基元的比较。下表给出分别以音素、声韵母和音节为基本单元识别出音节的结果。其中识别所用为帧同步算法。结论：以音节为基本识别单元，性能好，但开销大。

表 3　识别基元的比较

前 n 名	1	2	3	4	5	6	7	8	9	10	100	132	162
音素	79.20	89.50	93.53	95.79	97.77	98.00	99.01	99.13	99.24	99.35	n/a	n/a	99.9
声韵	82.30	91.93	95.60	96.67	98.03	98.52	99.07	99.23	99.36	99.45	n/a	99.9	n/a
音节	85.08	94.01	96.60	97.79	98.50	98.92	99.12	99.31	99.45	99.50	99.9	n/a	n/a

【参考文献】

[1] **Wilpon, Jay G., Rabiner, Lawrence R., Lee, Chin-Hui, Goldman, E.R.**, "Automatic Recognition of Keywords in Unconstrained Speech Using Hidden Markov Models," *IEEE Trans. on ASSP*, vol.38, No.11, Nov. 1990, pp.1870-1878

[2] **Makhoul, J.**, "Linear Prediction: A Tutorial Review," Proc. *IEEE*, vol. 63, Apr. 1975, pp.562-580

[3] **Gold, B., Rader, C.M.**, "*Digital Processing of Signals*, " New York: McGraw-Hill, 1969, p.246

[4] **Furui, S.**, "Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum," *IEEE Trans.on ASSP*, vol. ASSP-34, No. 1. Feb., 1986, pp.52-59

[5] **郑方、吴文虎、方棣棠，** "CDCPM 及其在语音识别中的应用，"已经被*《软件学报》*录用，1996

[6] **郑方、吴文虎、方棣棠，** "连续距离密度的分段概率模型 CDD-SPM，" *第三届全国人机语音通讯学术会议论文(NCMMSC-94)*，1994 年 10 月，重庆，pp.238-241

[7] **Bellegarda, J.R., Nahamoo, D.**, "Tied Mixture Continuous Parameter Modeling for Speech Recognition," *IEEE Trans. on ASSP*, vol.ASSP-38, No.12, Nov. 1990, pp.2033-2045

[8] **Gauvain, J.-L., Lee, C.-H.,** "Bayesian Learning for Hidden Markov Model with Gaussian Mixture State Observation Densities," *Speech Comm.*, Vol. 11, 1992, pp.205-213

[9] **陈永彬、王仁华，** *《语言信号处理》*，中国科学技术大学出版社，1990 年

Speech Recognition Units in the Chinese Dictation Machines

Fang Zheng, Wenhu Wu, and Ditang Fang

ABSTRACT: In the Chinese Dictation Machine, the performance of the acoustic model will directly affect the use of the language model and the quality of the machine, and flexibility of the acoustic model is, to some extent, depend on the choosing of the speech recognition unit. A great deal of comparison experiments have been done across a giant corpus using CDCPM-based acoustic models. The results show that SYLLABLE should be a best choice. Meanwhile, experiments on NN-based output observation scoring scheme have been done with good performance.

Keywords: Speech Recognition Unit, Center-Distance Continuous Probabilistic Model, Nearest Neighbor Rule