基于多路差别子空间的语速变化语音的识别*

吕成国　韩纪庆　王承发

哈尔滨工业大学计算机科学与技术学院 哈尔滨 150001

E-mail: cgl258@sina.com

摘　要：目前，对正常情况下的语音识别技术的研究已经做了很多，但是针对变异语音识别的研究做得还很少。语速变化是发音变异的一种，本文建立了快、慢和正常语速的语音库，运用差别子空间方法对语速变化的语音进行了训练和识别，并对其进行了改进，提出了多路差别子空间方法。实验结果表明，这种方法对语速变化的语音有良好的识别效果。

关键词：语速变化，　多路差别子空间，　语音识别

Variational Speed Speech Recognition Based on Multi-Difference Subspace Method

LU Chengguo　HAN Jiqing　WANG Chengfa

School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001

E-mail: cgl258@sina.com

Abstract: At present, numerous studies have been made for speech recognition under normal environment, but researches for variation speech recognition have been not made so much relatively. Speed variation is a kind of speech variation, a speed variation corpus include of fast, slow and normal speech is constructed. We implement and improve training and recognition algorithm based on difference subspace methods, propose multi-difference subspace method. The experimental results show that the system has good recognition performance for variational speed speech.

Keyword: Speed variation, Multi-Difference subspace, Speech recognition

1 引言

目前语音识别技术已经取得很大成就，孤立词识别系统和连续语音识别系统都已经达到了很高的识别率。但是，当训练与识别的环境不匹配时，绝大多数识别系统的识别率会明显下降。在环境异常情况如恐惧、愤怒、环境噪声影响下话者由于心理紧张和情绪变化语音会

*本文承国家自然科学基金项目（项目号 60085001）的资助。

发生变异，话者在身体不适（如感冒）时和生理情况受到影响（如加速度变化）时语音也会发生变异。发音变异可以引起各语音参数的不同变化，以致常规语音识别系统的识别率大大下降。从七十年代开始，国外就有人分析研究发音变异对语音识别性能产生的影响，而直到八十年代末，才开始有人研究顽健的变异语音识别问题。近年来，变异条件下顽健语音识别的研究已经逐步得到重视，在 ICASSP、EUROSPEECH 等重要国际会议论文集中经常有这方面的研究论文。针对有噪变异语音研究，国外已建立了由实际和模拟环境采集的包扩多种变异情况的语音数据库，实验统计了各种变异条件下多种参数的变化规律。Hansen[1-3]、Lippman[4]和 Chen[5]等都对变异语音的识别进行了研究，提出了对不同变异语音进行补偿和提高识别率的方法，但至今仍处于探索比较阶段。

语速变化是发音变异的一个方面，当人们遇到紧急情况或心理紧张时，语速会变快，思考时或犹豫不决时，语速会变慢。快、慢情况的语音与正常语音相比，主要是元音持续时间的长短不同[2]，同时基音频率也会发生变化，但是快的情况更复杂一些，有声母和韵母之间过渡音段丢失的现象，而中间字的清音部分有时也会丢失。

2 语速变化的语音数据库

我们在实验室环境下，录制了快、慢和正常语速的语音数据，建立了语音库，词表大小 30 词，包括 10 个单字词、10 个双字词和 10 个三字词。

为了检验所录制的多重话语风格语音数据的有效性，我们做了一个主观听觉判断实验，随机播放所录制的三种话者风格语音数据，让 4 名实验者分别通过听觉来主观判断语音的话语风格类型，每种风格抽取 2 组各 30 个语音样本，共有 180 个语音样本。最终 3 种话语风格的主观听觉判断实验的结果如表 1 所示。

表 1 多重话语风格的主观听觉判断结果

判断结果 \ 语音类型		快	慢	正常
实验者 1	快	168	0	8
	慢	0	180	2
	正常	12	0	170
实验者 2	快	164	0	8
	慢	0	180	0
	正常	16	0	172
实验者 3	快	172	0	8
	慢	0	180	2
	正常	8	0	170
实验者 4	快	162	0	14
	慢	0	180	0
	正常	18	0	166
平均正确率（%）		92.5	100	94.2

从以上结果可以看出，所建立的语音数据库中，慢的情况正确率达 100%，满足研究实验的需要。对于判断错误的结果，我们进行了分析，如果四名实验者中有两名以上判断错误，我们认为数据不可靠，需重新录制在数据库中进行替换，如果只有一名实验者判断错误，我们认为是实验者的原因，数据基本可靠，在数据库中保留。

3 差别子空间法语音识别的基本原理

由于语音产生的时序性，不仅不同的人对同一个词的发音有差别，而且同一个人对同一个词的两遍发音之间也是有差别的。由此，同一个词的多遍发音之间，有相同的共性部分，也有不同的差别部分，差别子空间法就是利用语音之间的这一性质来进行语音的识别。M. Bilginer[6]等用差别子空间法进行非特定人的孤立词语音识别，取得了很好的识别效果，我们采用差别子空间法进行变异的语音识别。

设 R^n 为 n 维特征矢量空间，$x=(x_1,x_2,\cdots,x_n)$, $y=(y_1,y_1,\cdots,y_n)$, $x,y\in R^n$。定义 $<x,y>=x_1y_1+x_2y_2+\cdots x_ny_n$; $\|x\|=<x,x>^{1/2}$.设某一词有 m 个训练样本 $a_1,a_2,\cdots a_m$。$a_i\in R^n, i=1,\cdots,m$，表示 m 个不同的话者或语音样本。定义：$b_1=a_2-a_1, b_2=a_3-a_1,\cdots,b_{m-1}=a_m-a_1$；式中 $b_i(i=1,\cdots,m-1)$ 表示了不同话者或语音样本之间的差别。一般 $a_1,a_2,\cdots a_m$ 是线性无关的，所以 $b_1,b_2,\cdots b_{m-1}$ 也是线性无关的。则 $B=span\{b_1,b_2,\cdots,b_{m-1}\}$ 称为 $a_1,a_2,\cdots a_m$ 的差别子空间。对 $b_1,b_2,\cdots,b_{m-1}$ 进行正交化得到 B 的一组标准正交基 $z_1,z_2,\cdots,z_{m-1}$，即 $\|z_i\|=1,<z_i,z_j>=\delta_{ij}$。设 a_i 在 B 上的投影为 $\overline{a_i}$，则

$$\overline{a_i}=<a_i,z_1>z_1+<a_i,z_2>z_2+\cdots+<a_i,z_{m-1}>z_{m-1}, i=1,\cdots,m$$

$\overline{a_i}$ 反映了 a_i 的个性特征，或者说是与其它特征矢量之间的差别，而 $a_i-\overline{a_i}$ 反映了 a_i 的共性特征。可以证明 $a_i-\overline{a_i}$ 与 i 无关，而且与选择哪个特征矢量作参考矢量无关。定义 $a_{com}=a_i-\overline{a_i}(i=1,\cdots,m)$ 为某个词的共性特征矢量，它表征了这个词的共性特征。假定

$c_i \in R^n, i = 1,\cdots,s$ 为测试样本的特征矢量，s 为测试集中测试样本特征矢量的总数。

设 $\overline{c_i^k}$ 为 c_i 在第 k 个差别子空间中的投影，即

$$\overline{c_i^k} = < c_i, z_1^k > z_1^k + < c_i, z_2^k > z_2^k + \cdots + < c_i, z_{m-1}^k > z_{m-1}^k, i = 1,\cdots,m$$

则 $c_{rem}(k) = c_i - \overline{c_i^k}$ 定义为 c_i 关于第 k 个词的剩余矢量。

训练时对词表中每个词构造差别子空间，每个词一个差别子空间。通过训练确定每个子空间的共性矢量与标准正交基。识别时，对于测试集中的每个特征矢量，都分别在所有差别子空间的标准正交基上求投影，然后这个特征矢量减去其投影得到剩余矢量，测试集中第 i 个特征矢量在第 k 个差别子空间上的剩余矢量表示为：

$$c_{\text{remaining}}^k = c_i - \overline{c_i}^k = c_i - \sum_{j=1}^{m-1} \left\langle c_i, z_j^k \right\rangle z_j^k .$$

每个剩余矢量与相应共性矢量相比较，失真最小者为结果。采用距离测度判别函数时表示为：

$$result = \arg\{ \min_{1 \le k \le \text{wordnum}} (\left\| c_{\text{remaining}}^k - a_{\text{common}}^k \right\|) \} .$$

其中 wordnum 为词表大小。

4 多路差别子空间识别方法

我们对差别子空间方法作了改进，设计了多路差别子空间的识别算法，对话语风格变化的语音进行训练和识别。其原理就是对词表中每个词都设计三个差别子空间，分别对应快、慢和正常三种情况，识别时对测试集中的每个特征矢量，都在每个词的三个差别子空间上求剩余矢量，采用距离测度求失真，最小的失真为这个词的失真，词表中所有词的最小失真对应的那个词，为识别结果。多路差别子空间方法的识别简图如图 1 所示：

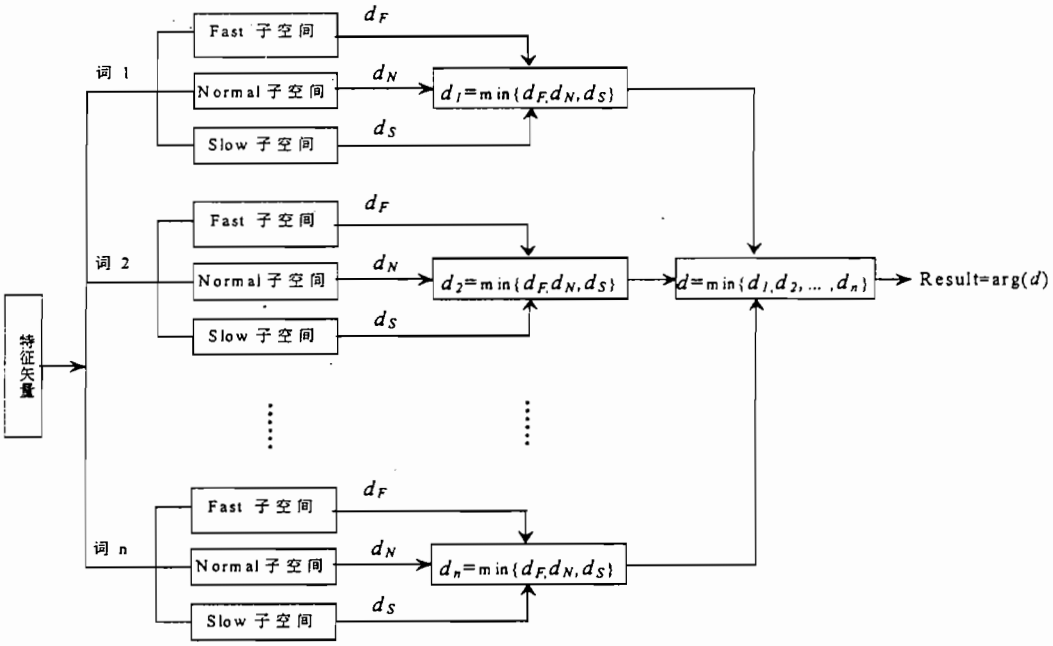

图 1 多路差别子空间方法简图

图中 Fast、Normal 和 Slow 子空间分别表示对应于语速快、正常和慢的词的子空间，d 表示失真测度。

5 实验结果与分析

我们实现了基于差别子空间法和多路差别子空间方法的语音训练与识别系统，并作了关于 HMM 多重风格训练方法、差别子空间的多重风格训练方法与多路子空间方法的对比实验。差别子空间的多重风格混合训练法是指用包含三种风格的每个词的所有训练语音构造一个差别子空间进行训练，词表中每个词对应一个差别子空间；而多路差别子空间法是用每个词的每种风格的训练语音分别构造一个差别子空间，词表中每个词对应三个差别子空间。

用语音库中三种情况各 12 遍共 36 遍语音数据组成训练集，测试集包括三种情况各 4 遍，特征提取采用 39 维特征向量，包括能量、12 阶 Mel 倒谱系数及其一阶、二阶差分，分别采用三种方法进行训练、识别，实验结果如表 2 所示：

表 2 采用三种方法的对比实验结果

训练方法 / 识别结果(%)	HMM 方法	差别子空间方法	多路差别子空间方法
快	81.3	64.0	90.0
慢	86.7	66.3	98.3
正常	90.7	71.3	91.7
总体平均	86.2	67.2	93.3

从结果可以看到，对多重话语风格的语音识别，多路差别子空间法的识别要明显好于差别子空间的多重风格混合训练法和 HMM 的多重风格训练方法，平均识别率达到了 93.3%。我们对采用多路差别子空间方法训练识别的错误识别结果进行了分析，发现有两种情况：一种是词本身的识别错误，另一种是类间识别错误，这种错误只有一种，就是“正常”的识别成“快”的情况，而这可能就是导致慢速情况语音比正常语音识别率高的原因。

6 结束语

本文提出多路子空间法的语音识别方法，对语速变化的语音识别取得了很好的识别效果，此方法对于多语种的语音识别、方言以及话语风格变化的语音识别都有良好的借鉴意义。

参 考 文 献

[1] J.H.L.Hansen, M.A.Clements. Stress Compensation and Noise Reduction Algorithms for Robust Speech Recognition. ICASSP'89, 1989, p266-269

[2] J.H.L. Hansen and S. Bou-Ghazale. Robust Speech Recognition Training via Duration and Spectral-based Stress Token Generation. IEEE Trans. On Speech and Audio Processing, 1995, 3(5): 415-421.

[3] John H. L. Hansen, Sahar E. Bou-Ghazale. Getting Started with SUSAS: A Speech under Simulated and Actual Stress Database. ESCA. Eurospeech97:1743~1746

[4] R.P.Lippmann, E.A.Martin, D.B.Paul. Multi-Style Training for Robust Isolated-Word Speech Recognition.ICASSP'87, 1987:705-708

[5] Y.Chen. Cepstral Domain Talker Stress Compensation for Robust Speech Recognition. IEEE Trans. On acoustics, Speech and Signal Processing, 1988, 36(4):433-439

[6] M. Bilginer Gulmezoglu, V. Dzhafarov, M. Keskin and A. Barkana. A novel approach to isolated word recognition. IEEE Transactions on Speech and Audio Processing, 1999, 7:620~628