# 一种基于与噪声特征距离最大的语音识别方法*

张 军 英

（西安电子科技大学 计算机系）

1991 年 6 月 14 日收到

传统的语音模板建立方法使得语音识别系统中各语音的特征模板长度不一，而对于大词汇表的语音识别来说，这将给模板库的查找、添加、修改（学习）、乃至语音的识别等带来较大的困难．本文提出了一种基于与噪声特征的 DTW 距离最大的语音帧数归整方法，及基于此的模板建立和语音识别方法，提出并定义了判断语音识别系统识别性能的性能评价函数．实验结果表明，这里所提出的方法能以少得多的计算量获得满意的识别性能．

## 一、引 言

传统的语音模板的建立方法，是在多个训练语音中经过两两的语音特征动态时间弯折，找出与各次语音特征的匹配距离为最小的那个语音特征作为该语音的特征模板，这使得在语音识别系统中各个语音的特征模板长度不一，而对于大词汇表的语音识别来说，语音特征模板库中各语音模板长度的不一致性，将给库的查找、添加、修改（学习）、乃至语音的识别等都带来较大的困难．本文为实现语音模板库的长度的一致性，提出了一种基于与噪声特征距离最大的语音特征帧数归整方法，并提出了基于此的语音模板建立和语音识别方法，实验结果表明了这种方法的有效性和可行性．

## 二、基于与噪声特征距离最大的语音识别方案

### 1. 语音特征的帧数规整

为了将各次发音的语音特征规整成等帧数的语音特征（其帧数应选为大量语音的帧数平均值，约为 25 帧），我们首先建立标准的 25 帧噪声模板，称之为无语音模板或空模板．对于用 LPC（线性预测）系数描述语音特征的情形来说，标准的噪声信号是白噪声信号，其功率谱是平的，故空模板第 $i$ 帧（$i=1—25$）的 LPC 系数都是 $a(i,0)=1$，$a(i,k)=0(k=1—p$，$p$ 为 LPC 阶数），由此建立起空模板 A．

语音之所以称之为语音，是因为它是人所发生的一种不同于噪声的特殊声音．正因为如此，为了将语音特征 $b(i,k)(i=1—J,k=0—p)$ 帧数规整为 25 帧的语音特征 $c(i,k)(i=1—25,k=0—p)$，以语音特征 $B$ 与空模板 $A$ 距离最大为原则进行动态时间弯折，即找到 $A$、$B$ 的合适的匹配路径，使 $A$ 与 $B$ 的总体平均距离为最大，如图 1($a$) 所示．若动态时间弯折的基本路径和权重选择如图 1($b$) 所示，则求匹配路径的递推表达式为：

$$g(1,1)=2d(1,1) \tag{1}$$

$$g(i,j)=\max\begin{cases} g(i-1,j-2)+2d(i,j-1)+d(i,j) \\ g(i-1,j-1)+2d(i,j) \\ g(i-2,j-1)+2d(i-1,j)+d(i,j) \end{cases} \tag{2}$$

$$D(A,B)=\frac{g(I,J)}{I+J} \tag{3}$$

* 国家自然科学基金资助课题



?1994-2017 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

(1)、(2)、(3) 式中，$d(i,j)$ 为语音的第 $j$ 帧特征向空模板第 $i$ 帧的弯折距离，$g(i,j)$ 为使语音特征的前 $j$ 帧向空模板的前 $i$ 帧弯折的最大弯折距离，$D(A,B)$ 为 $B$ 以最大距离弯折至 $A$ 的总体平均距离，$I$ 取为 25. 为了减小计算量及与实际可能性相符，在匹配时加上了限制线 $j=i\pm r$. 运用上述方法得到了最优路径后，语

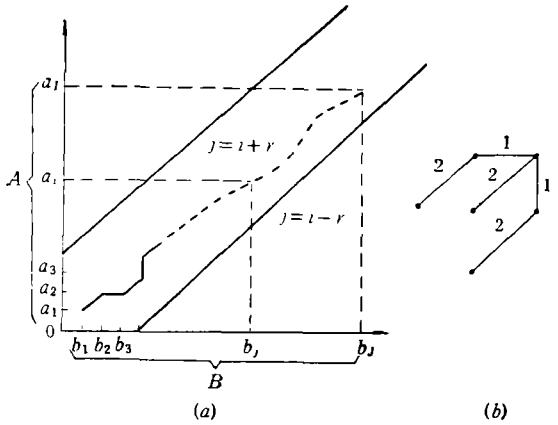


图 1 基于与噪声特征距离最大的语音特征帧数规整图
(a) 使总体平均距离最大的匹配路径
(b) 基本路径和权重选择

音特征 $B$ 在与噪声特征距离最大意义下在 25 帧特征空间的投影 $C$ 的第 $i$ 帧 $c(i,k)(i=1—25,\ k=0—p)$ 可通过语音特征 $B$ 匹配上空模板的第 $i$ 帧的所有语音帧特征的算术平均得到，即

$$c(i,k)=\frac{\sum_j b(j,k)}{l(i)} \qquad (4)$$

其中，求和是对匹配上空模板第 $i$ 帧的所有语音帧特征求和，$l(i)$ 是匹配上空模板第 $i$ 帧的所有语音帧的帧数. 这样，就通过找出语音特征 $B$ 在与噪声特征 $A$ 距离最大意义下在25帧特征空间上的投影 $C$，实现了语音特征的时间规整.

基于正交投影原理，我们还可以得出另一种语音特征的帧数规整方案. 根据正交投影的基本原理，将语音特征 $B$ 与 25 帧特征空间的原点 $a(i,k)=0(i=1—25,k=0—p)$ 进行以距离最小为原则的动态时间弯折，即式 (2) 中将 max 换为 min，根据所得到的最优路径按 (4) 式计算，从而得到语音特征 $B$ 的基于正交投影原理的在 25 帧特征空间的投影 $c(i,\ k)$ $(i=1—25,k=0—p)$，作为语音特征 $B$ 的帧数规整结果，如图 2 所示.

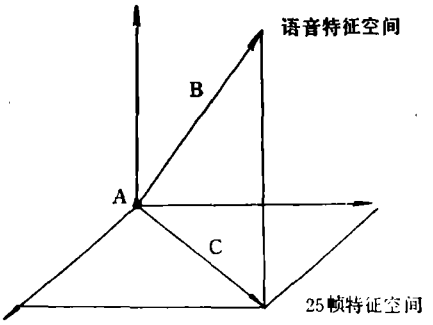


图 2 基于正交投影的语音特征帧数规整图

**2. 语音模板建立及匹配距离计算**

当有多个训练语音作为训练和形成该语音的模板时，这个模板可通过各训练语音在 25 帧特征空间上的投影的算术平均得到. 用这种方法建立语音模板库，保证了无论训练语音的帧数是多少，各语音模板的帧数都是 25 帧，给库的添加、查找、修改(学习)、乃至语音的识别都带来了许多的方便. 为了实现语音识别，样本语音特征与模板的匹配距离的计算有以下三种方法:

(1) 样本语音特征与语音模板以距离最小为原则直接进行动态时间弯折，这与传统的匹配方法一致，只是模板的形成方法不同.

另外两种方法均依据以下的思路. 既然模板的建立是在进行了语音特征的帧数规整的预处理之后得到的，那么，出于处理的一致性的考虑，也应对样本语音的特征进行同样的预处理，即先求样本语音特征在 25 帧特征空间上的投影，然后用这个投影与模板进行匹配，这时的匹配距离计算有以下两种方法:

(2) 样本语音特征在 25 帧特征空间的投影与语音模板的距离作为匹配距离.

(3) 样本语音特征在 25 帧特征空间的投影向语音模板的以最小距离意义下的动态时间弯折距离.

**3. 语音识别的性能评价方法**



?1994-2017 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

语音识别方法的性能最终应该用语音识别率来描述，但运用合理的性能评价函数也不失是一种有效的性能评价方法。

若用某种语音识别方法进行语音匹配，所得到的匹配距离为：发与模板所表示的音相同的音所得到的匹配距离在 $d_1 < D < d_2$ 的区域上。发与模板所表示音不相同的音所得到的匹配距离在 $d_3 < D < d_4$ 的区域上，如图 3($a$) 所示，通常有 $d_3 > d_1, d_4 > d_2$，定义两区域的绝对距离 $abs$ 和相对距离 $rel_1$ 分别为

$$abs = d_3 - d_2 \tag{5}$$

$$rel_1 = \frac{d_3 - d_2}{d_4 - d_1} \tag{6}$$

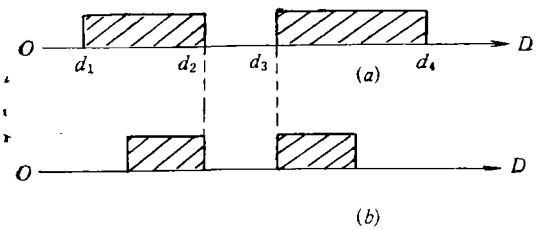


图 3　语音识别的性能评价，$abs$ 相同但 ($a$)$rel_1$ 较小和 ($b$)$rel_1$ 较大的情况

若用 $per1$ 来表征这个语音识别方法的性能，明显地，$abs$ 越大，表明为正确识别语音所应给出的距离门限的选择范围越大，因此性能值 $per1$ 也应越大，在 $abs$ 相等的情况下，如图 3($a$)、($b$) 所示，明显地图 3($a$) 的性能比图 3($b$) 的性能好，而图 3($a$) 的 $rel_1$ 比图 3($b$) 的 $rel_1$ 小，所以 $rel_1$ 越小则性能值 $per1$ 应越大。因此定义语言识别方法的性能为两区域的绝对距离与由相对距离所引起的绝对距离的增量之和，即

$$per1 = abs + abs(1 - rel_1) \tag{7}$$

用类似的方法还可定义另一种性能度量

$$per2 = abs + abs(2 - rel_2) \tag{8}$$

其中

$$rel_2 = \frac{d_3 - d_2}{d_3 - d_1} + \frac{d_3 - d_2}{d_4 - d_2} \tag{9}$$

实验中对各方法的性能的观察与用 $per1$、$per2$ 计算的结果完全一致，

## 三、实验研究及结果

我们用 Turbo C 在 Super PC 机上对上述方法进行了实验研究，基本的实验参数如下：

抗混迭滤波器：60—4500Hz 带通滤波器；采样频率：10kHz/s；预加重系数：0.9357；帧长：256ms；帧移：128ms；窗函数：Hamming 窗；LPC 阶数：12；距离测度：板仓测度：在实验中，每音均发音 12 次，其中前 5 次用于训练，后 7 次用于匹配，所选择的音为由二男二女发的 '$LA$' 音和 '$JIN$' 音，不同于模板的音为他们发的数字 '0'、'1'、'2'…'10' 的音。用本文给出的各种方法所得到的匹配距离所计算出的性能评价值见表 1，其中性能评价计算所需的 $d_1$、$d_2$、$d_3$、$d_4$ 均为由相应的匹配距离的均值和方差计算而得到的统计数值。在表 1 中，DTW1、DTW2、DTW22 是指用基于与噪声特征距离最大时间规整方法建模并分别用相应的"二、2"中 (1),(2),(3) 方法进行匹配，DTW3、DTW4、DTW44 是指用基于正交投影原理的时间规整方法建模并分别用相应的"二、2"中 (1),(2),(3) 方法进行匹配。明显地，DTW1、DTW2、DTW22 的性能优于 DTW3、DTW4、DTW44 的性能。这就提出了一个问题，为什么基于与噪声距离最大的时间弯折方法比基于正交投影的方法要优越得多？基于正交投影原理的时间规整方法是使语音特征与 25 帧特征空间的原点的总体平均距离最小，从局部看，第 $j$ 帧语音特征向第 $i$ 帧弯折的局部距离为

表 1　基于与噪声特征距离最大和基于正交投影的时间规整方法的比较

| DTW | 性能评价值 | |
|---|---|---|
| | $per1$ | $per2$ |
| DTW1 | 2.171197 | 2.226915 |
| DTW2 | 4.022538 | 4.211206 |
| DTW22 | 3.262091 | 3.375461 |
| DTW3 | 1.523006 | 1.608420 |
| DTW4 | 0.817455 | 0.943284 |
| DTW44 | 0.921800 | 0.990246 |



?1994-2017 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

$$d(i,j)=\sum_{k=0}^{p}[b(j,k)-0]^2 \qquad (10)$$

$$=1+\sum_{k=1}^{p}b^2(j,k)$$

因为 $\sum_{k=1}^{p}b^2(j,k)$ 正好是语音特征 $B$ 的第 $j$ 帧与空模板第 $i$ 帧的匹配距离，所以上述的局部距离与语音特征 $B$ 与空模板第 $i$ 帧的局部距离只差常数 1，因此使总体平均距离最小的正交投影的结果与使语音特征与噪声的空模板的总体平均距离最小的结果是完全一致的。实验研究也完全证明了这一点，即基于正交投影原理的时间弯折方法正好与基于与噪声特征距离最大的时间规整方法相反，本质上是一种基于与噪声距离最小的时间弯折方法，这种弯折方法尽量减小语音与噪声的差异，明显地是不适用于语音识别的，因此其用于语音识别的性能较差是理所当然的。

在基于与噪声特征距离最大的语音识别方法中，匹配的方法不同也导致了性能的不同，由表 1 可见，DTW1、DTW2、DTW22 的性能顺序是：DTW2 > DTW1 > DTW22，即“二、(2)”的方法最好。从计算量看，若语音模板数为 $n$，用“二、(1)”的方法需样本语音与每一模板作弯折，共需 $n$ 次弯折，“二、(2)”的方法首先将样本语音与空模板进行一次弯折，然后与每一模板求欧氏距离，共需作的弯折次数仅为 1 次，而“二、(3)”的方法样本语音首先与空模板弯折，弯折结果再向模板弯折，共需 $n+1$ 次弯折。所以从计算量上看，“二、(2)”的计算量明显减少，使得匹配速度提高了约 $n$ 倍，这对于大词汇表的语音识别是至关重要的。综上，DTW2 在匹配性能上和运算量上都是最好的。

当选择的模板语音为 ‘$LA$’ 音，而不同于模板的音选为 ‘$A$’、‘$O$’，‘$E$’ 等音时，各种方法的性能评价结果见表 2($a$) 所示，而若将 ‘$A$’ 音也视为是 ‘$LA$’ 音所得到的性能评价见表 2($b$) 所示。很明显，($b$) 的性能较 ($a$) 的性能有显著提高，这说明上述方法对声母不同韵母相同的音仍容易发生误识，这也是传统方法同样存在的问题，这需要通过提高声母距离在总体匹配距离中的比重的方法加以解决[2]。

表 2　声韵分割对识别性能的影响

($a$)

| DTW | 性能评价值 | |
|---|---|---|
| | *per1* | *per2* |
| DTW1 | 3.420805 | 3.636990 |
| DTW2 | 3.597657 | 3.952658 |
| DTW22 | 3.542139 | 3.770438 |

($b$)

| DTW | 性能评价值 | |
|---|---|---|
| | *per1* | *per2* |
| DTW1 | 4.851592 | 5.193031 |
| DTW2 | 6.067523 | 6.511900 |
| DTW22 | 5.633274 | 6.056029 |

本文提出的方法的识别性能已用性能评价方法进行了性能的评价，但为了与传统的DTW方法进行比较，下面我们给出这里方法的粗略的识别率。由于本文的实验还不很充分，这时用识别率来描述识别精度是不太合适的，因为识别率是一个只有在进行了非常大量的实验基础上所得到的统计数值，在目前没有进行充分实验的情况下，只可能给出一个非常粗略的识别率数值供参考。在对应于表 2($a$) 的实验中，实验的 4 × 7 次中有 4 次误识到，正确识别率为 85.7%，这个数值就整个语音的识别而言已算是较高的了，而在对应于表 2($b$) 的将 ‘$A$’ 音也视为 ‘$LA$’ 音的识别实验中，在实验的 4×7 次中没有一次误识别，即正确识别率为 100%，这说明如果用本文的方法进行音节的识别可望达到很高的正确识别率。而传统的 DTW 方法的音节识别精度，就文献 [3] 上记录，在全部 2251 个音节中，误识别就达 223 个，其正确识别率只有 90%。

由此可见，基于与噪声特征距离最大的语音识别方法，在模板库的建立与管理、识别速度以及识别精度上都将是一种非常有效的方法



?1994-2017 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

## 四、结　　语

综上，我们提出了一种基于与噪声特征距离最大的语音特征帧数规整方法及在此基础上的模板建立和语音识别方案，所建立的模板库中各模板的帧数相等，给库的建立、添加、修改(学习)等带来许多方便，并提出了对语音匹配方案的一种性能评价方法，实验表明了对样本语音特征进行帧数归整后直接与模板匹配能够获得较高的匹配性能，且识别速度比传统方法提高约 $n$ 倍，对于大词汇表语音识别来说，由于 $n$ 很大，所以这里的方法是非常有意义的.

### 参 考 文 献


[1] 杨道淳，声学技术，6-3(1987)，24—27.

[2] Haltsonen S., *IEEE Trans.* on *ASSP.*, **33**-2(1985), 449—453.

[3] 俞一彪，袁保宗，信号处理，3-4(1987)，206—213.


# SAW 功率倒谱分析系统


李吉孚　　严晓兰

(合肥电子工程学院)



1991 年 8 月 15 日收到



倒谱 (Cepstrum) 技术已广泛应用在信号处理中，例如目标分类，提取多径效应环境中的信号[1]，和语音处理[2]等. 本文介绍用 SAW 器件实现功率倒谱系统的原理、设计和实验结果，同时也指出它的用途.


## 一、引　　言

倒谱有三种：功率倒谱(Power Cepstrum)、复倒谱 (Complex Cepstrum) 和相位倒谱 (Phase Cepstrum)[3]. 本文仅讨论功率倒谱.

设时域信号为 $S(t)$，其功率倒谱 $C(\tau)$ 定义为信号 $S(t)$ 的功率谱的对数的功率谱：

$$C(\tau)=|FT\{\log|FT[S(t)]|^2\}|^2 \quad (1)$$

这里 $FT[\cdot]$ 表示 Fourier 变换. 可见倒谱分析要经过两次 Fourier 变换. 输入信号经第一个频谱分析器将信号从时域变到频域，求得其频谱. 通过真对数放大器之后，由第二个频谱分析器将信号由频域变至伪时域 (Pseudo-Time Domain)，也称为倒频域 (Quefrency Domain)，即可求得输入信号的倒谱. 两个频谱分析器之间插入一个真对数放大器，它使两个函数频域相乘变成相加运算. 因此，倒谱分析系统有解卷积功能.

对于某些具体的应用，有时省去中间的对数放大. 简化后的功率倒谱定义为时域信号幅频谱的 Fourier 逆变换的幅值平方.

早期倒谱分析系统，主要采用数字 FFT 技术通过计算机软件来实现，一般不是实时处理. 应用近代快速数字技术，也只能实时处理语音一类的慢变信号，对于宽带快变信号，目前还无能为力. 然而，应用 SAW 器件构成的功率倒谱分析系统能克服此缺陷，实时处理快变宽带信号. 这给倒谱技术开辟了广阔的应用前景. 特别是在电子对抗领域，用它来侦察敌方通信、雷达信号的特征，以便实施干扰.

本文所述系统的实现及实验结果，迄今尚未见国内资料报导. 有关 SAW 倒谱系统的国外文献，也仅从理论上简单地阐述了倒谱系统有高的处理增益，而本系统通过具体实验，有力地证实了系统确实具有提高信噪比的优点，从而扩大了系统在信号处理中的应用范围.




?1994-2017 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net