一种引入帧间相关信息的 HMM 语音识别方法 [1]

赵 力 邹采荣 吴镇扬

(东南大学无线电工程系 南京 210096)

摘 要 该文提出了一种基于复数帧段输入 HMM 的语音识别方法，它采用相继的复数帧组成的特征参数向量作为语音识别 HMM 的输入，能有效地在语音识别 HMM 中引入帧间相关信息。为了进一步改善复数帧段输入 HMM 的输出概率分布函数，作者还提出了用 MGDF 和 RBF 函数作为复数帧段输入 HMM 的输出概率分布函数的方法。通过对非特定人汉语孤立数字和连续数字语音识别试验，证实了该文提出的引入帧间相关信息方法的有效性。

关键词 语音识别，隐马尔可夫模型，帧间相关信息，复数帧段输入

中图号 TP391.42

1 引 言

虽然隐马尔可夫模型 (HMM) 是现在最流行的语音识别模型，然而基本型的 HMM 有一个固有的缺陷，就是它采用状态输出独立假设，影响了 HMM 描述语音信号时间上帧间相关动态特性的能力。为了弥补这一缺陷，已经有许多改进方法被提出。例如：增加 HMM 状态数和在时间轴方向利用回归系数法 [1]；使用线性或非线性预测器法 [2]；利用多项式回归函数法 [3]；利用条件概率 HMM 的方法 [4]；利用模拟人的听觉顺向时频特性的动态倒谱系数法 [5] 等。这些提案对于改善传统输出独立 HMM 的缺陷都是有效的方法，但是它们实现起来较为复杂。

利用语音帧间相关信息最直接最简便的方法，是采用相继的复数帧组成的特征参数矢量作为输入特征量的方法。这种方法最初是由井手等人提出 [6]，Ostendorf 等人把这一方法推广到了连续语音识别系统 [7]。本文利用这一设想，提出了一种基于复数帧段输入 HMM 的语音识别方法，首先从理论上推导了复数帧段输入 HMM 的可行性，并且为了改善复数帧段输入 HMM 的输出概率分布函数，提出了用修正高斯分布函数 (MGDF) 和径向基函数 (RBF) 代替传统高斯分布函数的方法。最后通过对非特定人汉语孤立数字和连续数字语音识别试验，证明了提出的复数帧段输入 HMM 可以较好地改善输出独立 HMM 的缺陷，是一种有效而简便的利用帧间相关信息的方法。

2 复数帧段特征量 HMM

假设输入特征参数序列 $Y = y_1, y_2, \cdots, y_T$，HMM 的状态序列 $X = x_1, x_2, \cdots, x_T$，则对于 Y 的 HMM 的输出概率可以由下式导出。

$$\begin{aligned} P(y_1, y_2, \cdots, y_T) &= \sum_x P(y_1 y_2 \cdots y_T, x_1 x_2 \cdots x_T) \\ &= \sum_x P(y_1 y_2 \cdots y_T | x_1 x_2 \cdots x_T) P(x_1 x_2 \cdots x_T) \\ &= \sum_x \prod_i P(y_i | y_1 y_2 \cdots y_{i-2} y_{i-1}, x_1 x_2 \cdots x_{i-1} x_i) P(x_i | x_1 x_2 \cdots x_{i-1}) \end{aligned} \quad (1)$$

[1] 1999-06-11 收到，2000-04-23 定稿

$$\approx \sum_x \prod_i P(y_i|y_{i-3}y_{i-2}y_{i-1}, x_{i-1}x_i)P(x_i|x_{i-1}) \tag{2}$$

$$= \sum_x \prod_i \frac{P(y_{i-3}y_{i-2}y_{i-1}y_i|x_{i-1}x_i)}{P(y_{i-3}y_{i-2}y_{i-1}|x_{i-1}x_i)} P(x_i|x_{i-1}) \tag{3}$$

$$\approx \sum_x \prod_i \frac{P(y_{i-1}y_i|x_{i-1}x_i)}{P(y_{i-1}|x_{i-1}x_i)} P(x_i|x_{i-1}) \tag{4}$$

$$= \sum_x \prod_i P(y_i|y_{i-1}, x_{i-1}x_i)p(x_i|x_{i-1}) \tag{5}$$

$$\approx \sum_x \prod_i P(y_i|x_{i-1}x_i)p(x_i|x_{i-1}) \tag{6}$$

这里 (2) 式和 (3) 式是 4 帧宽度的条件概率 HMM, (4) 式和 (5) 式是 2 帧宽度的条件概率 HMM, (6) 式是标准形式 HMM 的输出概率算式。对 (3) 式和 (4) 式，认为在 $x_{i-1}x_i$ 发生的前提下，$y_{i-3}y_{i-2}y_{i-1}$ 以及 y_{i-1} 几乎必然发生，所以作为近似计算，只取它们的分子部分作为近似输出概率的计算式，得到如下 (7) 式和 (8) 式，并且把它们定义为 4 帧宽度和 2 帧宽度的复数帧段输入 HMM 。

$$\sum_x \prod_i \frac{P[y_{i-3}y_{i-2}y_{i-1}y_i|x_{i-1}x_i]}{P(y_{i-3}y_{i-2}y_{i-1}|x_{i-1}x_i)} p(x_i|x_{i-1}) \approx \sum_x \prod_i P[y_{i-3}y_{i-2}y_{i-1}y_i|x_{i-1}x_i]p(x_i|x_{i-1}) \tag{7}$$

$$\sum_x \prod_i \frac{P[y_{i-1}y_i|x_{i-1}x_i]}{P(y_{i-1}|x_{i-1}x_i)} p(x_i|x_{i-1}) \approx \sum_x \prod_i P[y_{i-1}y_i|x_{i-1}x_i]p(x_i|x_{i-1}) \tag{8}$$

复数帧段输入 HMM 的输入是由相继的复数帧特征参数矢量按顺序组合成的一个复合特征参数矢量，每个复数帧段特征参数的段移为一帧。这些复数帧段特征参数作为语音输入特征数据在模型训练和语音识别时使用。

3 对复数帧段输入 HMM 输出概率分布函数的修正

标准连续 HMM 的输出概率分布函数是如下面 (9) 式所示的高斯分布函数。

$$P(\boldsymbol{y}|i,j) = \sum_{k=1}^{L} w_{ijk} \frac{1}{(2\pi)^{p/2}|\Sigma_{ijk}|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu}_{ijk})\Sigma_{ijk}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}_{ijk})^t\right\} \tag{9}$$

在复数帧段输入 HMM 中，连续几帧信号作为一个特征参数矢量，输入特征参数矢量的维数 p 将增加。由于 (9) 式中计算协方差矩阵 Σ_{ij} 所需的乘法次数是 p^2+p, 当 p 增加时，计算量和内存所用空间都将按平方指数增长。另外，更重要的是随着 p 的增加，Σ_{ij} 的推定误差将增大，同时由于 HMM 参数增多，当训练数据不足时，训练的模型精度将下降，从而降低识别性能。为了避免直接计算 Σ_{ij} 所引起的计算量和计算误差增大的问题，在复数帧段输入 HMM 中，必须对 (9) 式进行修正。本文提出了两种修正方法，并通过实验比较了这两种方法。

3.1 利用 MGDF 的修正方法 [8]

把 (10) 式所示的协方差矩阵和它的特征值与特征向量的关系。

$$\Sigma_{ijk} = \sum_{d=1}^{p} \lambda_{dijk}\Phi_{dijk}\Phi_{dijk}^t \tag{10}$$

代入 (9) 式，对 (9) 式进行修正可得

$$p(\boldsymbol{y}|i,j)=\sum_{k=1}^{L}w_{ijk}\Bigg/\left[((2\pi)^{p/2}\left(\prod_{d=1}^{p}\lambda_{dijk}\right)^{\frac{1}{2}}\right]^{-1}$$
$$\times\exp\left\{-\frac{1}{2}\sum_{d=1}^{p}\Bigg/\left[\lambda_{dijk}\left(\Phi_{dijk}^{t}(\boldsymbol{y}-\boldsymbol{\mu}_{dijk})\right)^{2}\right]^{-1}\right\} \tag{11}$$

由于高次的特征值和特征向量的推定误差比低次的大，所以令高于某一阈值 m 的特征值 $\lambda_{m+1},\lambda_{m+2},\cdots,\lambda_p$ 为一固定的等值 $N_0\sigma^2/(N+N_0)$。这里 N 是学习样本的总数，N_0 是信赖度常数，σ^2 是和特征值有关的量 (本文中通过实验选择 σ^2 为 $\lambda_{m+1},\lambda_{m+2},\cdots,\lambda_p$ 的平均值)。我们令 $N_0/(N+N_0)=\alpha(0<\alpha<1)$ 以及 $\alpha\sigma^2=h^2$，并根据 (12) 式的关系，可推出 (13) 式的修正高斯分布函数 MGDF。以 (13) 式代替 (9) 式作为 HMM 的输出概率分布函数。(13) 式中的参数 α 以及分割点 m 由实验确定。

$$\sum_{d=1}^{p}[\Phi_{dijk}(\boldsymbol{y}-\boldsymbol{u}_{ijk})^{t}]^{2}=||\boldsymbol{y}-\boldsymbol{u}_{ijk}||^{2} \tag{12}$$

$$p(\boldsymbol{y}|i,j)=\sum_{k=1}^{L}w_{ijk}\Big/\left[(2\pi)^{p/2}\left(h^{2(p-m)}\prod_{d=1}^{m}\lambda_{dijk}\right)^{1/2}\right]$$
$$\times\exp\left\{-\frac{1}{2h^{2}}\left[||\boldsymbol{y}-\boldsymbol{u}_{ijk}||^{2}+\sum_{d=1}^{m}\frac{h^{2}-\lambda_{dijkd}}{\lambda_{dijkd}}(\Phi_{dijk}^{t}(\boldsymbol{y}-\boldsymbol{u}_{ijk}))^{2}\right]\right\} \tag{13}$$

3.2 利用 RBF 的修正方法 [9]

采用 RBF 函数，代替 (9) 式作为 HMM 的输出概率分布函数。本文中我们采用的 RBF 函数是 VQ-PNN(Probabilistic Neural Network) 函数中的一种 [10]，如 (14) 式所示。

$$P(\boldsymbol{y}|i,j)=\sum_{k=1}^{L}w_{ijk}\frac{1}{(2\pi)^{p/2}\sigma_{ijk}^{p}}\exp\left\{-\frac{(\boldsymbol{y}-\boldsymbol{\mu}_{ijk})(\boldsymbol{y}-\boldsymbol{\mu}_{ijk})^{t}}{2\sigma_{ijk}^{2}}\right\} \tag{14}$$

式中 $\boldsymbol{\mu}_{ijk}$ 表示 HMM 从第 i 状态到第 j 状态转移时的第 k 个均值向量，σ_{ijk} 是和 $\boldsymbol{\mu}_{ijk}$ 对应的标准偏差，我们假定各维数之间相互独立，而且具有一个共同的标准偏差。

4 实验和结果

复数帧段输入 MGDF-HMM 和 RBF-HMM 的训练是采用 Segmental k-mean 法进行的。首先采用非复数帧段特征参数训练初始 HMM 模型，然后用 Viterbi 方法，针对初始 HMM 的每一个状态，分割和收集训练用数据，最后利用各状态分割和收集的训练用数据，采用复数帧段特征参数，求取相应的均值和方差，组成 MGDF 和 RBF 函数。

实验是针对非特定人的汉语孤立数字和连续数字的语音识别。在孤立数字识别实验中，邀请 75 名男性话者每个人对汉语 10 个数字各发音 4 遍 (共 3000 个数据)，其中 50 人的发音作为训练用数据，另 25 个人的发音作为评价识别用数据。在连续数字识别实验中，选用 35 类 4 位数汉语连续数字语音，邀请 30 名男性话者每个人对 35 个 4 位数字各发音 2 遍，共 2100 个发音作为评价识别用数据。实验中模拟语音信号经 12kHz 采样，$1-0.98Z^{-1}$ 的预加重，窗长 21.33ms(256 点), 窗移 10ms 的汉明窗后，进行 14 阶 LPC 分析，然后从 14 阶 LPC 系数中求出 10 阶的 Mel 倒谱系数 (临界带倒谱系数) 和 Mel 倒谱的 10 阶的线性回归

系数，并计算归一化能量和一阶差分能量。实验用 HMM 是汉语数字音节单位 5 状态混合连续分布 HMM， HMM 所取的 Mixture 数分别是 2， 4 和 8 。在汉语连续数字语音识别实验中，采用 One Pass DTW(Viterbi scoring) 方法，求取最佳的孤立数字 HMM 的连接序列。为了进行比较，我们分别用传统的独立输出连续 HMM 、复数帧段输入 MGDF-HMM 和 RBF-HMM 进行了孤立数字和连续数字的语音识别比较实验。实验结果如表 1 所示。

表 1 识别实验结果 [%]

(a) 测试数据集 (孤立数字)

	2 帧宽度			4 帧宽度			6 帧宽度		
Mixture	2	4	8	2	4	8	2	4	8
独立输出 HMM	95.1	95.5	96.0	95.1	95.5	96.0	95.1	95.5	96.0
帧段 MGDF-HMM	977	97.6	98.2	98.2	99.2	98.9	98.5	99.4	99.0
帧段 RBF-HMM	98.5	99.4	99.5	99.2	99.7	99.5	99.5	99.7	99.8

(b) 训练数据集 (孤立数字)

	2 帧宽度			4 帧宽度			6 帧宽度		
Mixture	2	4	8	2	4	8	2	4	8
独立输出 HMM	97.2	98.1	98.4	97.2	98.1	98.4	97.2	98.1	98.4
帧段 MGDF-HMM	98.9	99.2	99.2	99.3	99.5	99.4	99.3	99.7	99.3
帧段 RBF-HMM	98.8	99.7	99.8	99.7	99.8	99.7	99.8	99.8	99.9

(c) 测试数据集 (连续数字)

	2 帧宽度			4 帧宽度			6 帧宽度		
Mixture	2	4	8	2	4	8	2	4	8
独立输出 HMM	82.8	83.2	82.7	82.8	83.7	82.7	82.8	83.8	82.9
帧段 MGDF-HMM	84.0	84.8	84.6	85.6	86.1	86.0	85.8	86.8	86.6
帧段 RBF-HMM	85.4	86.0	85.0	86.0	86.9	86.4	86.7	87.8	87.7

从表中给出的 2 帧， 4 帧和 6 帧宽度复数帧段输入 HMM 的实验结果可知，这些宽度的复数帧段输入 HMM 的识别性能都优于独立输出 HMM 。同时，我们也进行了大于 6 帧宽度的复数帧段输入 HMM 的识别性能比较实验。结果表明，宽度过宽不利于识别性能的改善。另外，我们过去的比较实验已经证明，在独立输出 HMM 中，单独利用帧间动态线性回归系数 (Δ 参数) 进行识别，识别性能比只利用倒谱或倒谱 $+\Delta$ 倒谱要差 [9] 。其次从不同输出概率分布函数的角度来看， RBF-HMM 的识别性能优于 MGDF-HMM ，所以选择 RBF 作为复数帧段输入 HMM 的输出概率分布函数更为合适。

5 结 论

本文提出了一种在语音识别 HMM 中引入帧间相关信息的方法。它采用相继的复数帧组成的特征参数矢量作为 HMM 的输入，这种方法是利用语音帧间相关信息最直接最简便的方法。同时我们还提出了两种适合于复数帧段输入 HMM 的输出概率分布函数，并通过实验证明了 RBF 函数形式比 MGDF 函数形式更为有效。通过对非特定人汉语数字语音识别实验，证明了本文提出的方法能够较好地改善 HMM 表述语音信号时间相关等动态特性的能力，提高 HMM 的语音识别性能，是一种有效而简便的在语音识别 HMM 中引入语音帧间相关信息的方法。虽然本文中我们仅仅将该算法应用在汉语数码语音识别中，该算法也可以用于其它语音识别领域。作为今后的研究课题，我们将进一步扩大复数帧段输入 HMM 的实验范围，并寻找更合适有效的复数帧段输入 HMM 的输出概率分布函数。

参 考 文 献

[1] V. N. Gupta, M. Lennig, P. Mermelstein, Integration of acoustic information in a large vocabulary word recognizer, ICASSP-87, Dallas, USA, 1987.2, 697–700.

[2] 坪香英一，ニュ - ラルネット驱动型 HMM, IEICE, Technical Report, 1989, SP89-83, 33–41.
[3] L. Deng, M. Aksmanoric, X. Sun, C. F. J. Wu, Speech recognition using hidden Markov models with polynomial regression functions as stationary states, IEEE Trans. on Speech & Audio Processing, 1994, (4), 507–520.
[4] C. J. Wellekens, Explicit correlation in hidden Maarkov model with optimized inter-frame dependence, ICASSP-95, Detroit, USA, 1995.1, 209–212.
[5] 相川清明，河原英纪，顺向マスキングの时间周波数特性を模拟した动的ケプストラムを用いた音韵识别，日本通信电子学会论文志 (A), 1993, J76-A(11), 1514–1521.
[6] 井手和之，牧野正三，时间－周波数を用いた无声破裂音の识别，日本音响学会论文志，1982, 39(5), 321–329.
[7] M. Ostendorf, S. Roukos, A stochastic segment model for phoneme-based continuous speech recognition, IEEE Trans. on Acoust., Speech & Signal Processing, 1989, ASSP-37(12), 1857–1869.
[8] T. Wakabayashi, S. Tsuruokaet,*ed al.*, On the size and variable transformation of feature vector for handwritten character, IEICE, J76-D- II (12), 2495–2503.
[9] L. Zhao, H. Suzuki, S. Nakagawa, A comparison study of probability functions in HMMs through spoken digit recognition, IEICE, TRANS.INF and SYST., 1995, E78-D(6), 669–675.
[10] S. Nakagawa, Estimation of probability density function and a posteriori probability and evaluation by vowel recognition, IEICE, Technical Report, 1992, SP92-24, 61–72.

A METHOD OF HMM SPEECH RECOGNITION INTRODUCED INTER-FRAME CORRELATION

Zhao Li　Zou Cairong　Wu Zhenyang

(*Department of Radio Engineering, Southeast University, Nanjing* 210096, China)

Abstract This paper applies segmental unit into HMM for speech recognition. In this model, several successive frames are combined and treated as an input vector. It expects that segmental unit input HMM would be effective to describe the inter-frame correlation information and has also proposed the MGDF and RBF to further improve output probability function. By comparing them with the traditional HMMs based on their speech recognition performance rates through the experiments of speaker-independent spoken digit (isolated/connected) recognition, the validity of the proposed appraoch could be verified.

Key words Speech recognition, Hidden Markov model, Inter-frame correlation information, Segmental unit input

赵　力：　男，1958 年生，博士后，研究方向为语音信号处理.
邹采荣：　男，1950 年生，教授，博士生导师，研究方向为信号处理.
吴镇扬：　男，1949 年生，教授，博士生导师，研究方向为信号处理.