语音识别中空间相关性信息的利用

余 鹏, 王作英
(清华大学电子工程系, 北京 100084)

摘 要: 在语音识别中不同语音单元的语音信号特征之间不是独立的, 描述不同声学状态信号特征之间的相互关系的信息称为"空间相关性". 空间相关信息在语音识别参数估计算法(训练和自适应)中有非常重要的作用. 本文对这种相关性作了探讨, 提出了一种在语音识别中应用空间相关信息的新方法. 我们用线性方程来描述空间相关性所体现出来的不同语音单元特征之间的依赖性, 通过分组 K-L 变换的方法来估计这组线性约束的相关系数, 并给出一种结合空间相关信息的训练方法. 实验结果表明, 空间相关的先验知识对语音识别训练模块的稳健性有明显的提高.

关键词: 语音识别; 声学层码本; 参数估计; 空间相关性; 线性约束
中图分类号: TP391 文献标识码: A 文章编号: 0372-2112 (2002) 07-0964-03

Using Spatial Correlation Information in Speech Recognition

YU Peng, WANG Zuo-ying
(*Dept. of Electronics Engineering, Tsinghua, Beijing 100084, China*)

Abstract: In Speech Recognition, features from different acoustic units are not independent. The correlation between different acoustic units is called "Spatial Correlation". Spatial Correlation is important for acoustic model estimation. In the paper, we proposed a new method of using spatial information in speech recognition. We use linear equation to subscribe spatial correlation, calculate equation coefficients by K-L transformation, and develop a new training algorithm with the linear constraints. Experimental results show the new method brings significant improvement in error reduction.

Key words: speech recognition; acoustic codebook; parameter estimation; spatial correlation; linear constraints

1 引言

如果将说话人的语音识别码本作为概率空间的样本点的话, 那么声学模型参数本身是一组随机变量. 这组随机变量的概率分布描述的是声学模型的不同状态在特征空间的分布情况, 我们称这种模型参数的分布为一种"空间结构".

空间结构信息已经作为先验知识被应用于语音识别自适应算法中. 如在 MAP 算法[3]和 MLLR 算法[4]中, 利用 Speaker Independent 模型参数作为先验知识, 这实际上是空间结构信息中的"均值"信息; 另一个例子是 MLMI 自适应算法[2], 利用一组说话人的 Speaker Dependent 模型参数作为先验知识, 这可以看作空间结构的一个粗略的样本集.

但在以上几例中, 都没有直接考虑不同声学状态之间的相关信息. 事实上, 声学模型不同状态不是相互独立的, 它们之间的相关信息对于跨状态的参数估计非常重要. MLLR 算法采用状态聚类的方法将状态分组, 并对同一组状态采用相同的线性变换, 这可以看作是空间相关信息的一种应用, 但这种应用显然是粗略的.

对于空间相关信息的研究主要问题有三个, 即如何描述空间相关性、如何得到空间相关性和如何应用这种相关性. 本文针对这三个方面的问题给出了一个解决方案, 实验结果表明采用空间相关信息的新的训练方法对于不充分数据量训练有明显的效果.

2 空间相关规律的数学描述形式

正态声学模型参数包含特征均值矢量和它的协方差矩阵, 称为码本. 在下面的讨论中, 我们将只考虑码本均值矢量, 此时一个声学模型可以看成是包含各个状态码本均值矢量的一个大矢量. 设声学码本状态数为 NS, 特征维数为 NF, 则声学模型相当于一个 $NS \times NF$ 维的大矢量, 我们称这个矢量为"码本矢量", 所在的空间称为"码本空间", 记作 $\boldsymbol{R}_{NS\times NF}$.

在信息论中, 两个随机变量 A、B 的相关性可以用他们的互信息 $I(A, B)$ 来衡量, 互信息越大, 相关性越强, 反之则弱. 当 $I(A, B) = 0$ 时, 表示 A、B 相互独立; 而当 $I(A, B) = H(A) = H(B)$ 时, 就表明 A、B 间存在确定性的相互关系. 直接描述空间的概率相关性是困难的, 但是确定性的相互约束则易于利用. 我们的研究正是针对这种确定性的空间相关约束. 记一个声学码本矢量为 $\boldsymbol{C}$, 则这种确定性的空间相关规律就可以表示为矢量 $\boldsymbol{C}$ 所满足的一组方程, $SC_n(\boldsymbol{C}) = 0$, $n = 0, 1, \cdots, NT$, 其中 NT 表示方程的数目, 这里每个方程我们称之为一个"约束". 我们假设这种约束是线性的, 这样, 空间相关性就可以表示成为这样的一组线性方程:

$$\begin{cases} SC_1 = (\boldsymbol{C}, A_1) - b_1 = \sum_{s=1}^{NS\cdot NF} c_s \cdot a_{1s} - b_1 = 0 \\ SC_2 = (\boldsymbol{C}, A_2) - b_2 = \sum_{s=1}^{NS\cdot NF} c_s \cdot a_{2s} - b_2 = 0 \\ \cdots\cdots\cdots \\ SC_{NT} = (\boldsymbol{C}, A_{NT}) - b_{NT} = \sum_{s=1}^{NS\cdot NF} c_s \cdot a_{NT,s} - b_{NT} = 0 \end{cases}$$

这里 $A_{ij} \in \boldsymbol{R}_{NS\times NF}$, $b_i \in \boldsymbol{R}$, $i = 1, 2, \cdots, NT$, $j = 1, 2, \cdots, NS \cdot NF$ 为约束系数, $(\cdot, \cdot)$ 表示 $\boldsymbol{R}_{NS\times NF}$ 空间的内积.

收稿日期: 2000-11-22; 修回日期: 2001-09-20

3　采用 K-L 方法来得到约束

选出一组说话人来组成统计集, 采用常规的码本估计方法得到每个人的码本矢量, 记整个矢量集为 $\boldsymbol{K}$. 我们的目标是找到一组约束, 使得 $\boldsymbol{K}$ 中的每个码本矢量都符合这组约束.

从前面我们知道, 码本空间 $\boldsymbol{R}_{NS\times NF}$的维数是$NS\times NF$, 因此上一个线性约束有 $NS\times NF+1$ 个参数, 这可以看作一个约束矢量. 估计这么一个大矢量对数据量的要求是很高的, 为此我们先对码本空间进行分割, 将 $\boldsymbol{R}_{NS\times NF}$分割为较小的乘积子空间, 在每个子空间内, 我们只需估计一个小得多的矢量.

首先, 我们假设空间相关性仅存在于语音学中发音相近的状态之间, 依据语音学知识将声学状态进行分组, 将发音相近的状态分在一组中. 将特征空间表示为 $\boldsymbol{R}_{NS\times NF}=\boldsymbol{R}_{NS^1\times NF}\times\boldsymbol{R}_{NS^2\times NF}\times\cdots\times\boldsymbol{R}_{NS^{NG}\times NF}$, 其中 NG 为分组的数目, NS^i, $i=1,2,\cdots,NG$ 为各组中的状态数. 分组的个数应该根据码本矢量集的大小来定, 以使在每个子空间内, 矢量集的大小都足够稳健地估计约束矢量.

进一步, 如果只考虑同一组状态之间同一维特征的相互关系, 子空间 $\boldsymbol{R}_{NS^i\times NF}$将再一次分割为$NF$ 个更小的乘积子空间$\boldsymbol{R}_{NS^i}$. 我们将最后生成的小空间称为“相关子空间”. 图 1 解释了码本空间的分割.

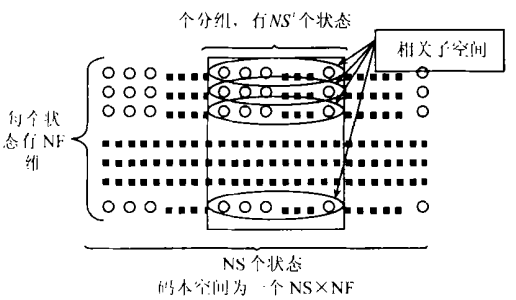

图 1　码本空间、状态分组和相关子空间的关系

线性空间中一组线性约束对应于全空间的一个线性流形, 因此求取线性约束相当于寻找与统计集 $\boldsymbol{K}$ 正交的线性流形, 这可以借助于 K-L 变换. K-L 变换将全空间分成相互正交的子空间, 如果某个子空间的特征值为零, 相当于 $\boldsymbol{K}$ 在这个子空间上的投影的平均能量为零, 即 $\boldsymbol{K}$ 正交于这个子空间.

在每个相关子空间中, 首先作 K-L 变换. 记

$$\boldsymbol{R}=\sum_{k=1}^{NC}(\boldsymbol{C}'_k-\overline{\boldsymbol{C}}')\cdot(\boldsymbol{C}'_k-\overline{\boldsymbol{C}}')^T/NC$$

为训练矢量的协方差矩阵, 其中 NC 代表训练矢量的个数, $\boldsymbol{C}'_k$ 代表训练矢量 $\boldsymbol{C}_k$ 在当前相关子空间上的截断矢量(即从一个 $NS\times NF$ 的大矢量中取出 NS^i 个分量组成一个小矢量), $\overline{\boldsymbol{C}}'=\sum_{k=1}^{NC}\boldsymbol{C}'_k/NC$ 代表截断矢量的均值.

对 $\boldsymbol{R}$ 作特征值分解

$$\boldsymbol{R}=\boldsymbol{E}\cdot\Sigma\cdot\boldsymbol{E}^T=(\boldsymbol{e}_1,\boldsymbol{e}_2,\cdots,\boldsymbol{e}_{NS'})\cdot diag(\sigma_1,\sigma_2,\cdots,\sigma_{NS^i})\cdot(\boldsymbol{e}_1,\boldsymbol{e}_2,\cdots,\boldsymbol{e}_{NS'})^T$$

其中 σ_j 是$\boldsymbol{R}$ 的特征值, $\boldsymbol{e}_j$ 是对应的特征矢量. 取特征值中最小的几个值, 不妨设为 $\sigma_{NS^i-NM^i+1}$, $\sigma_{NS^i-NM^i+2}$, $\cdots$, σ_{NS^i}使满足 $\sum_{j=NS^i-NM^i+1}^{NS^i}\sigma_j<(1-\lambda)\cdot\sum_{j=1}^{NS^i}\sigma_j$, 其中 λ 是我们预设的门限值. 由 $(\boldsymbol{C}'-\overline{\boldsymbol{C}}')\perp span(\boldsymbol{e}_{NS^i-NM^i+1},\boldsymbol{e}_{NS^i-NM^i+2},\cdots,\boldsymbol{e}_{NS^i})$, 得到关于 $\boldsymbol{C}'$ 的NM^i 个线性约束.

$$\begin{cases}\sum_{j=1}^{NS^i}c'_j\cdot e_{NS^i-NM^i+1,j}=(\boldsymbol{C}',\boldsymbol{e}_{NS^i-NM^i+1})=(\overline{\boldsymbol{C}}',\boldsymbol{e}_{NS^i-NM^i+1})=\sum_{j=1}^{NG}\overline{c}_j\cdot e_{NS^i-NM^i+1,j}\\ \sum_{j=1}^{NS^i}c'_j\cdot e_{NS^i-NM^i+2,j}=(\boldsymbol{C}',\boldsymbol{e}_{NS^i-NM^i+2})=(\overline{\boldsymbol{C}}',\boldsymbol{e}_{NS^i-NM+2})=\sum_{j=1}^{NG}\overline{c}_j\cdot e_{NS^i-NM^i+2,j}\\ \cdots\cdots\cdots\\ \sum_{j=1}^{NS^i}c'_j\cdot e_{NS^i,j}=(\boldsymbol{C}',\boldsymbol{e}_{NS^i})=(\overline{\boldsymbol{C}}',\boldsymbol{e}_{NS^i})=\sum_{j=1}^{NS^i}\overline{c}_j\cdot e_{NS^i,j}\end{cases}$$

由于 $\boldsymbol{C}'$ 是 $\boldsymbol{C}$ 的一个截断, 因此上式也就是 $\boldsymbol{C}$ 的一组约束. 将所有相关子空间得到的约束组合起来, 就得到关于 $\boldsymbol{C}$ 的所有约束.

$$SC_n(\boldsymbol{C})=(\boldsymbol{C},A_n)-b_n=\sum_{s=1}^{NS\cdot NF}c_s\cdot a_{n,s}-b_n=0,\ n=1,2,\cdots,NT$$

其中$NT=\sum_{i=1}^{NG}NM^i$ 为所有约束的个数.

4　利用空间相关约束的码本估计算法

HMM 码本估计算法的原理是找到一组码本参数, 使得在这组参数下的 HMM 模型对训练数据能够给出最大的似然值, 这是一个最优化过程.

$$\boldsymbol{C}=\arg\max(P(S|\boldsymbol{C}))$$

其中 S 代表训练特征数据.

在这个算法中加入空间相关性约束, 只须对最优化算法中加入约束. 即

$$\begin{cases}\boldsymbol{C}=\arg\max(P(S|\boldsymbol{C}))\\ \text{s.t.}\quad SC_n(\boldsymbol{C})=0,\ n=1,2,\cdots,NT\end{cases}$$

我们系统采用单高斯、全协方差矩阵模型, 训练方法是一个基于欧氏距离的 K 均值算法. 这相当于最优化算法

$$\boldsymbol{C}=\arg\min\left(\sum_{i=1}^{NS}\sum_{k=1}^{N_i}\sum_{j=1}^{NF}(c_{i\times NF+j}-s^i_{k,j})^2\right)$$

其中 $c_{i\times NF+j}$表示码本矢量 $\boldsymbol{C}$ 中对应第 i 个状态、第 j 维的分量, $s^i_{k,j}$表示训练数据中对应第 i 个状态的第k 个矢量的第j维, N_i 表示训练数据中对应第 i 个状态的矢量数目.

当加入空间相关信息后, 训练算法变为

$$\begin{cases}\boldsymbol{C}=\arg\min\left(\sum_{i=1}^{NS}\sum_{k=1}^{N_i}\sum_{j=1}^{NF}(c_{i\times NF+j}-s^i_{k,j})^2\right)\\ \text{s.t.}\quad SC_n(\boldsymbol{C})=\sum_{s=1}^{NS\cdot NF}c_s\cdot a_{n,s}-b_n=0,\ n=1,2,\cdots,NT\end{cases}$$

这是个一个凸集上的凸泛函优化问题, 根据最优化理论, 这个问题有唯一的最优解。用 Lagrange 乘子法求得其解如下

记$\overline{s}_{i,j}=\sum_{k=1}^{N_i}s^i_{k,j}/N_i$, $i=1,2,\cdots,NS$, $j=1,2,\cdots,NF$

$$v_{n,m}=\sum_{s=1}^{NS\cdot NF}a_{m,s}\cdot a_{n,s}/N_{[s/NF]},\ n,m=1,2,\cdots,NT$$

$$w_n = b_n - \sum_{i=1}^{NS}\sum_{j=1}^{NF} a_{n,\,i\times NF+j}\cdot \bar{s}_{i,\,j},\ n = 1, 2, \cdots, NT$$

$$\boldsymbol{U} = \boldsymbol{V}^{-1}\cdot \boldsymbol{W},$$

$$c_{i\times NF+j} = \bar{s}_{i,\,j} + \sum_{n=1}^{NT} u_n \cdot a_{n,\,i\times NF+j}/N_i$$

在最后的解中, $\bar{s}_{i,\,j}$ 是不加入空间相关约束时的码本估计公式, 后面一项 $\sum_{n=1}^{NT} u_n \cdot a_{n,\,i\times NF+j}/N_i$ 是空间相关性约束带来的修正.

求解最后结果时, 矩阵 $\boldsymbol{V}$ 是一个分块对角阵, 它的求逆可以分解成为 $NS^i \times NS^i$ 维的小矩阵的求逆.

5 实验结果

以下实验采用的实验数据集是国家 863 高科技计划提供的数据, 共 77 个文件, 均为男声数据, 每个文件 600 句左右. 实验采用的语音识别模型是 HMM 模型的一个改进模型 DDBHMM(基于段长分布的隐含马尔可夫模型[1]) , 对每个状态采用全协方差的单高斯分布来描述, 特征提取 14 维 MFCC 系数加上能量维和一、二阶差分共 45 维. 实验中, 取前 70 个文件作为空间相关性训练集, 后 7 个文件作为测试集. 测试时, 将 400 句之前留出用于训练说话人相关码本, 400 句之后用于进行识别测试. 以下给出的实验数据均为声学的误识率.

利用语音学知识, 将汉语中的声母分成 20 组, 形成 900 个相关子空间; 韵母分成 41 组, 形成 1845 个相关子空间.

分别只对声母应用约束、只对韵母应用约束、对声母韵母同时应用约束进行实验, 实验时采用 100 句用于估计说话人相关码本, λ 取为 0.9. 结果如下.

表 1 对声母、韵母分别进行约束实验

	约束	m93	m94	m95	m96	m97	m98	m99	平均
基线	0	25.1	26.7	19.5	35.1	27.5	19.2	27.7	25.8
声母	4277	25.6	26.9	20.6	35.5	26.2	19.5	26.3	25.8
韵母	17452	23.4	25.1	18.6	35.0	27.1	18.5	26.1	24.8
联合	21729	23.3	25.1	19.4	35.2	25.5	19.6	24.6	24.7

在联合约束的情况下, 误识率下降了 4.6%. 同时可以看到, 对声母进行约束的效果没有对韵母进行约束的效果好, 这是因为汉语中声母状态一般较短, 稳定段短, 过渡段长; 而韵母状态则稳定段比较长, 因此韵母之间的相互约束关系比较明显.

改变 λ 值, 分别设定为 0.999、0.995、0.990、0.950、0.900、0.800、0.500 产生空间相关约束, 将约束用于进行测试文件的说话人相关码本的训练, 再将训练出的码本用于识别测试. 只对韵母约束进行测试, 采用 100 句进行码本估计.

表 2 门限 λ 对结果的影响

λ	约束	m93	m94	m95	m96	m97	m98	m99	平均
基线	0	25.1	26.7	19.5	35.1	27.5	19.2	27.7	25.8
0.999	732	25.1	26.6	19.5	35.0	27.2	19.0	27.3	25.7
0.995	3507	25.0	26.7	19.3	34.6	27.2	19.3	26.6	25.5
0.990	5595	25.3	26.9	19.3	34.5	27.2	19.6	27.3	25.7
0.950	13231	24.4	26.4	18.6	34.4	27.2	19.5	26.5	25.3
0.900	17452	23.4	25.1	18.6	35.0	27.1	18.5	26.1	24.8
0.800	21662	22.7	25.0	18.6	35.2	27.1	18.6	26.7	24.8
0.500	26010	22.3	23.7	18.8	36.0	29.8	18.9	27.2	25.2

从图中可以看出, 门限值 λ 取得比较大时, 约束数目较少, 对效果改善不明显. 而当 λ 取的过小时, 约束的精确度降低, 效果也会下降. 因此, λ 值取到 0.9 左右是比较合适的.

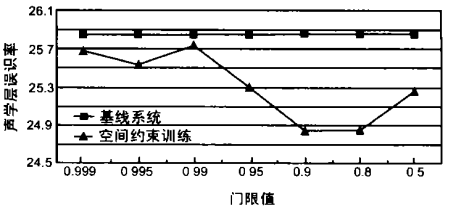

图 2 门限值 λ 对实验结果的影响

6 结论

本文对在语音识别码本估计中空间相关的先验知识的利用进行了探讨. 通过采用线性约束的形式描述声学状态的空间相关性和利用语音学知识将全码本空间分割为小的相关子空间的方法, 解决了对码本空间结构估计数据量不足的问题, 并给出了在语音识别训练算法中应用空间相关性的算法.

从实验结果可以看出, 汉语中韵母状态相互之间的约束性要强于声母状态, 这是因为汉语发音中韵母段往往比声母段发音更为充分且稳定. 对门限值参数分析实验表明, 在约束较少时, 增加约束数目可以进一步改善训练效果, 但当约束太多时, 由于约束的精确度下降, 效果也会下降.

文中采用线性约束的形式来描述空间相关性, 这使我们可以直接用 K-L 变换来求取约束参数, 也给训练算法中的应用带来了方便. 如果采用其他的约束形式, 如多项式约束、对数约束、指数约束等, 应能够更精确地描述空间相关性, 但会大大增加算法复杂度. 对空间相关性的研究而言, 如何找到最有效率的描述形式仍是值得讨论的. 这将是我们下一步的研究目标.

参考文献:

[1] 王作英. 基于段长分布的 HMM 语音识别模型 [A]. 第二届全国汉字、汉语识别会议论文集 [C]. 1989.

[2] 王作英, 刘丰. Speaker adaptation using maximum likelihood model interpolation [A], Proceeding of ICASSP [C]. 1999.

[3] Chirr-Hui Lee, Chirr-Heng Lin, Bing-Hwang Huang. A study on speaker adaptation of the parameters of continuous density hidden markov models [J]. IEEE Trans. On Signal Processing, 1999, 39: 806–814.

[4] C J Leggetter, P C Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models [J]. Computer Speech and Language, 1995, 9(2): 171–185.

作者简介:

余 鹏 男, 1976 年生于上海市, 1997 年毕业于上海交通大学电子工程系, 获学士学位, 现是清华大学电子工程系硕博连读研究生, 研究方向为语音信号处理.