第26卷 第3期
2000年5月

自 动 化 学 报
ACTA AUTOMATICA SINICA

Vol. 26, No. 3
May, 2000

短文

# 基于 *K*-MEANS 算法的语境相关矢量量化[1)]

许晓斌* 丁 丰 林碧琴 袁保宗

(北方交通大学信息科学研究所 北京 100044)

(* E-mail: xiaobinxu@263.net)

**摘 要** 研究用于连续语音识别的语境相关矢量量化技术. 提出采用 *k*-means(*k*-均值)算法逐一地调整决策树叶子所包含的各个语境,实现对音素模型的混合密度的优化. 实验结果表明,采用 *k*-means 算法的语境相关矢量量化得到的平均分布密度比简单合并决策树叶子所得到的平均分布密度提高4%～10%.

**关键词** 连续语音识别,语境相关矢量量化,*k*-means 算法,混合密度的优化.

# CONTEXT-DEPENDENT VECTOR QUANTIZATION BASED ON *K*-MEANS ALGORITHM

XU Xiaobin DING Feng LIN Biqin YUAN Baozong

(*Institute of Information Science, Northern Jiaotong University, Beijing* 100044)

**Abstract** An approach based on *k*-means algorithm to implement context-dependent vector quantization for continuous speech recognition is presented. With the approach, individual phonetic context of each leaf in a decision tree can be moved from one leaf to another, in order to optimize the mixture density given by that tree. Experimental results given in this paper demonstrate that the average likelihood given by this approach based on *k*-means algorithm is four to ten percent higher than that obtained with the previous method of optimizing mixture densities by merging leaves of decision trees.

**Key words** Continuous speech recognition, context-dependent vector quantization, *k*-means algorithm, optimization of mixture density

1)国家自然科学基金资助项目.

收稿日期 1997-12-23 收修改稿日期 1999-11-20

# 1 引言

在目前大多数语音识别系统中，为使音素模型能够准确描述该音素在各种语境中的发音，采用与发音语境相关的矢量量化方法[1,2]来确定音素模型中的混合密度. 每个混合密度的发音数据是用与发音语境无关的音素模型对训练数据做 Viterbi alignment 而获得，这种语境无关音素模型使用相应音素在各种发音语境中的训练数据训练，训练可靠性较高，能够准确地将每一帧训练数据分别划分到相应的状态(转移)中.

但是，现有机器的速度和存储空间大小限制了音素模型可容纳的混合密度分量的数目. 因此，大量不同语境的发音数据被迫要用同一混合分量来表示. 这样，就要求在矢量量化之后，决策树中只产生很少的几片(目前一般不超过10)终止节点(叶子)，作为混合密度的分量.

如何准确地产生这些叶子或混合分量，是语境相关矢量量化要研究的问题.

本文给出一种改进的语境相关矢量量化方法，其特点体现在：1)事先不给出二值语境问题，二值语境问题由矢量量化过程自动给出；2)决策树中每一个叶子的每一种语境的训练数据集合可以单独合并到其它叶子中，以获得对音素模型的混合密度的优化.

# 2 现有方法中存在的问题

现有两种语境相关矢量量化方法. 一种方法[2]是手工设定决策树节点的二值语言问题(或称语境问题)，在矢量量化过程中，这些二值语言问题不可改变. 这种矢量量化方法的缺点是设定的二值语言问题的主观性较大，会影响分类(混合密度)的精确程度. 另一种方法[1]则相反，不需要手工设定决策树节点的二值语言问题，并且在矢量量化过程中，各个节点的二值语言问题可以改变，从而实现对混合密度的优化.

然而，后一种方法的缺点是，它的二值语言问题的改变是通过合并两片发音数据分布密度相似的叶子(混合分量)而实现的. 由于每片叶子可能包含多个发音语境，从而使得每次合并是多个发音语境的数据的“融合”，而不是只将其中某一发音语境的训练数据合并到其它叶子中. 此时，可能会因为无法对每一个发音语境做逐一的调整，而不能获得最优的混合密度.

# 3 基于 $k$-means 方法的语境相关矢量量化

基于对上述问题的分析，给出一种基于 $k$-means 算法的语境相关矢量量化方法. 在这种方法中，首先利用文[1]中给出的语境相关矢量量化算法，为各音素的每一个状态(转移)生成一棵具有 $k$ 片叶子的决策树，然后利用 $k$-means 算法对这些叶子中所包含的发音语境(及其发音数据)不断进行调整，直至得到一种最优的混合密度.

## 3.1 训练语境无关音素模型

与以往的方法一样，在语境相关矢量量化之前，先用相应音素的全部发音语境的训练数据训练出一个离散密度的 HMM. 该 HMM 的矢量量化码本大小可以在210～256之间.

特征矢量包括一、二阶倒谱系数差分,这些倒谱差分拥有自己独立的矢量量化码本.此外,每一个 HMM 也拥有自己独立的上述矢量量化码本(由倒谱及倒谱差分的多量化码本构成).每一个音素模型的拓扑结构(即用于该音素的语境相关模型,也用于它的语境无关模型)如图1所示,有3个状态,每一个状态有一个前向转移和一个自转移.各个状态分别具有自己的混合密度.

语境无关音素模型的初始化训练数据是手工切分的.在初始化训练完成后,用各音素模型(语境无关模型)去自动切分和标注该音素在各种语境中的训练数据,这种自动切分和标注是一个迭代过程,直至收敛为止.

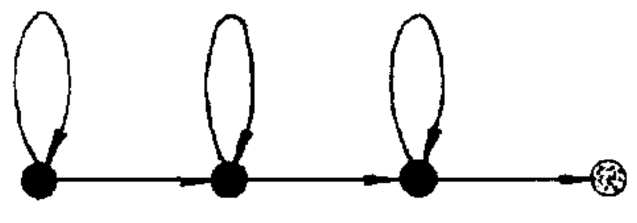

图1　音素模型的拓扑结构

在完成离散密度的(语境无关)HMM 的训练之后,用各音素的离散 HMM 对其在各种发音语境中的训练数据做 Viterbi alignment,将这些训练数据划分到模型的各个状态中.然后为每一音素模型的每一个状态建立一棵决策树.决策树的根节点拥有同一状态在全部发音语境中的训练数据.

### 3.2　基于 *k*-means 算法的混合密度的优化

决策树节点的分裂和对混合密度分量的优化(即对二值语言问题进行调整),是根据节点的分裂增益(goodness-of-split)及平均分布密度完成的.

对于要训练的音素模型 $P_0$的某一节点 $n$,给出一个语境问题 $q\in\boldsymbol{Q}$($\boldsymbol{Q}$ 是 $n$ 有关 $P_0$相邻音素的问题的集合),根据语境问题的答案的布尔值("是"或"否")将 $n$ 分裂成两个子节点(记作左子节点 $n_l$ 和右子节点 $n_r$).相应的分裂增益记作 $m_n(q)$.将 $n,n_l,n_r$ 的分布密度分别记作 $P_n,P_l,P_r$.可以用 Gaussian 分布表示为

$$P_x(y)=N(\boldsymbol{y},\mu_x,\Sigma_x). \tag{1}$$

上式中 $x=n,l,r;\mu_x,\Sigma_x$ 分别是 $n,n_l,n_r$ 的均值矢量和(对角)协方差矩阵.

分裂增益 $m_n(q)$可以根据下式计算[1]

$$m_n(q)=\sum_{k=1}^{N_l}\log P_l(\boldsymbol{y}_{l,k})+\sum_{k=1}^{N_r}\log P_r(\boldsymbol{y}_{r,k})-\sum_{k=1}^{N_n}\log P_n(\boldsymbol{y}_{n,k}). \tag{2}$$

上式中 $\boldsymbol{y}_{l,k},\boldsymbol{y}_{r,k},\boldsymbol{y}_{n,k}$分别是划分到 $n_l,n_r,n$ 的训练数据;$N_l,N_r$ 和 $N_n$ 是相应的特征矢量的数目;$q$ 表示使 $n$ 分裂的语境问题.

通过计算分裂增益,可以从音素模型 $P_0$的根节点的语境集合中,得到分裂增益最大的前 $K$ 个发音语境及其训练数据集合,作为决策树的初始叶子(混合密度的各个分量).确定初始叶子的目的,是因为混合密度分量的优化过程采用的是 k-means 算法,而 k-means 算法的性能的优劣,在很大程度上取决于初始化的分类集合的中心[3].

在确定决策树的初始叶子之后,利用 k-means 对混合密度进行优化,即对各个叶子中的发音语境进行调整,使决策树的整体平均分布密度达到最大.用下式计算决策树的平均分布密度

$$M=\frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{N_i}\log p_i(\boldsymbol{y}_{i,j}), \tag{3}$$

式中 $\boldsymbol{y}_{i,j}$是第 $i$ 片叶子(即混合密度的第 $i$ 个混合分量)的第 $j(j=1,2,\cdots,N_i)$个训练数据,$N=\sum_{i=1}^{K}N_i$.

## 4 实验结果

训练数据来自6个说话人的1000个句子. 每一帧训练数据(特征矢量)包括12阶 mel 倒谱系数,40ms 和80ms 倒谱系数的一阶差分,20ms 倒谱系数的二阶差分,以及瞬时能量的一阶差分,(log)基音频率及其一阶差分,帧长25ms,帧移间隔10ms,采样频率11kHz.

如前所述,使用文[1]的算法完成初始的语境相关矢量量化,将分裂增益最大的前 $K$ 个发音语境作为决策树的初始化叶子,其中最后一个叶子是多个发音语境的混合. 在确定初始化叶子之后,利用 $k$-means 算法对所有叶子中的语境逐一地进行调整,使决策树的整体平均分布密度达到最大. 实验结果如表1所示.

表1 分别采用 $k$-means 算法和叶子"融合"算法所得到的平均分布密度

| 决策树叶子(混合分量)的数目 | 平均 log 分布密度 ($k$-means 算法)的规整值 | 平均 log 分布密度 (叶子融合)的规整值 |
|---|---|---|
| 2 | −2.24 | −2.31 |
| 3 | −1.51 | −1.57 |
| 4 | −0.89 | −0.99 |

表中给出叶子数分别为2,3,4时决策树平均分布密度(规整值). 表中第2列给出的是采用 $k$-means 算法,对决策树中各叶子的发音语境逐一进行调整时得到的最大平均分布密度(即得到最优混合密度分量);第3列给出的是用文[1]的算法,仅对相似的决策树叶子进行"融合"时得到的最大平均分布密度. 可以看出,采用 $k$-means 算法对发音语境逐一地进行调整,所得到的最大平均分布密度均高于仅对相似的决策树叶子进行"融合"所得到最大平均分布密度. 由此可见,采用 $k$-means 算法对发音语境逐一地进行调整,可以更有效地对混合密度分量进行优化,从而得到更准确的混合密度.

**致谢** 本文研究工作得到国家模式识别实验室黄泰翼研究员的指正和鼓励,作者在此表示感谢.

## 参 考 文 献

1 Bahl L R, Souza P V De *et al*. Context dependent vector quantization for continuous speech recognition. In: Proc. 1993 IEEE Int. Conf. on Acoustic, Speech, and Signal Processing, Minneapolis MN May. 1993

2 Hwang M, Huang X, Alleva F A. Predicting unseen triphones with senones. *IEEE Trans. Speech and Audic Processing*. 1996, **4**(6): 412～419

3 Tou J T. Pattern Recognition Principles, Massachusetts: Addison-Wesley Publishing Company, 1974

**许晓斌** 北方交通大学信号与信息处理专业博士生. 主要研究方向为语音识别和人机对话系统.

**丁 丰** 北方交通大学信号与信息处理专业博士生. 主要研究方向为语音信号处理和自然语言处理.

**林碧琴** 1943年生,1965年毕业于北方交通大学,现为北方交通大学信息科学研究所副教授. 长期从事语音信号处理领域的教学和科研工作.

**袁保宗** 教授,博士生导师. 现任中国电子学会信号处理分会主席. 长期从事语音、图象、计算机视觉、自然语言处理、多模态信息处理、虚拟现实等领域的研究工作.