计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200200006

语音识别中单音节识别研究综述

张 经 杨 健 苏 鹏
大理大学数学与计算机学院 云南 大理 671003
(zhang_gold@163.com)

摘 要 声学模型建模可实现对语音信号的处理和特征抽取,是语音识别过程中必不可少的基础性工作,同时也是影响语音识别整体性能的一个重要因素。在语音识别中,选择合适的建模基元能使后续系统获得更高的准确率和更强的鲁棒性。音节是汉语等汉藏语系的最小发音单位,针对其发音特点,研究使用音节作为汉藏语系语音识别的建模基元,再提取相应的特征进行识别就有着尤为重要的意义。针对单音节识别目前的研究进展,首先介绍了基于有限状态矢量量化的算法,以及其改进算法在单音节识别中的研究成果;然后介绍了基于隐马尔可夫模型的算法,并详细介绍了将隐马尔可夫模型与其他算法相结合的音节识别研究成果;接着介绍了基于神经网络的算法;最后总结并提出了单音节识别研究未来发展的重要方向。

关键词:语音识别;单音节识别;矢量量化;隐马尔可夫模型;人工神经网络
中图法分类号 TN912.34

Survey of Monosyllable Recognition in Speech Recognition

ZHANG Jing, YANG Jian and SU Peng
School of Mathematics and Computer Science, Dali University, Dali, Yunnan 671003, China

Abstract Acoustic model modeling realizes the processing of speech signals and feature extraction, which is an essential basic work in the process of speech recognition and an important factor affecting the overall performance of speech recognition. In speech recognition, selecting appropriate modeling primitives can make subsequent systems obtain higher accuracy and stronger robustness. Syllable is the smallest pronunciation unit of Sino-Tibetan languages such as Chinese. According to its pronunciation characteristics, it is of great significance to study the use of syllable as the modeling element of Sino-Tibetan language speech recognition and to extract the corresponding features for recognition. In view of the current research progress of monosyllabic recognition, this paper first introduces the algorithm based on finite state vector quantization and the research results of its improved algorithm in monosyllabic recognition. Then the algorithm based on hidden Markov model is introduced, and the syllable recognition research results combining hidden Markov model with other algorithms are introduced in details, and then the algorithm based on neural network is introduced. Finally, the important development direction of monosyllabic recognition research in the future is summarized and proposed.

Keywords Speech recognition, Monosyllable recognition, Vector quantization, Hidden Markov model, Artificial neural network

1 引言

语音识别技术主要分为语音特征提取、声学模型建模和语言模型建模三大技术领域。其中,声学模型建模可实现对语音信号的处理和特征抽取的工作。不同语音识别方法的声学模型建模基元的选择各不相同,建模基元根据时间粒度大小可分为音素基元、声韵母基元、半音节基元、音节基元、词基元等[1]。建模基元的选择是语音识别的基础,也是影响语音识别整体性能的一个重要因素。汉语等汉藏语系语言是单音节结构语言,音节是其语音语义中最小的发音单位[2]。但目前在语音识别领域中,印欧语系语言通常把音素作为最小单位,因此选择音素为识别基元的研究较多,而面向音节识别的研究较少。本文主要从矢量量化算法、隐马尔可夫模型、神经网络算法3个方面对以音节为基元的语音识别研究进行介绍。

2 基于矢量量化的算法

矢量量化(Vector Quantization,VQ)[3]是一项信源编码技术,20世纪70年代末,Linda等成功地实现了VQ码本的生成并将其应用于语音编码中。该技术通过抑制信号量化过程中产生的冗余信息量,从而实现了高效率的语音信号压缩。

VQ应用于语音识别中,使用与系统词库中的每一个字或词相对应的码本(也就是VQ码本)作为该字或词的参考模板。识别时,对于输入的任意语音特征矢量序列,计算出语音每一帧的特征矢量与码本的失真之和除以该语音的帧数,即

本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:云南省哲学社会科学规划项目项目(YB2017072);云南省地方高校联合基金面上项目(2018FH 001-064)
This work was supported by the Yunnan Philosophy and Social Sciences Planning Project (YB2017072) and General Project of Joint Fund of Local Colleges and Universities in Yunnan Province(2018FH 001-064).
通信作者:杨健(sbjc1215@126.com)

总平均失真矢量误差，误差最小的码本所对应的字或词就是识别结果。识别过程如图 1 所示。

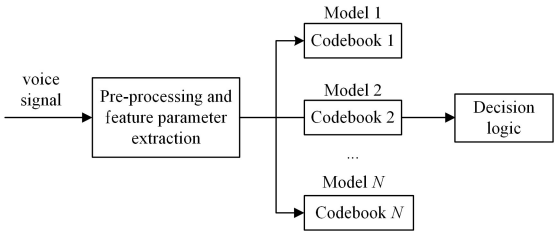

图 1　基于矢量量化的语音识别流程图

Fig. 1　Flow chart of speech recognition based on vector quantization

有限状态矢量量化(Finite State Vector Quantization，FSVQ)改进了传统的 VQ 方法，在量化矢量时，不仅利用了当前输入的帧信息，还利用了前续帧的信息，通过利用语音帧间相关性的方法提高了量化性能。文献[4]利用语音特征参数中个人发音特征的信息，提出了 FSVQ 算法，为 0～9 这十个数字的音节设计局部码本，形成了不同说话人的参考模板，并应用在训练和识别过程中，使识别讲话者的准确率达到了 95%。FSVQ 量化器公式如下：

$$U_n = Q(X_n, S_n) \tag{1}$$

$$S_{n+1} = F(U_n, S_n) \tag{2}$$

$$Y_n = U(U_n, S_n) \tag{3}$$

其中，$F(U_n, S_n)$是状态转移函数，该函数结合了当前帧与前续帧的信息，使得性能比无记忆的 VQ 大大提高。

矢量量化方法最关键的一步就是设计出最优码本，LBG 算法是设计码本的一种基本算法，但这种算法容易陷入局部最优解，且依赖于初始码本的选取。针对传统算法的这一缺陷，文献[5]结合了 Chu 等提出的猫群算法[6]，这种群体智能算法对全局最优解有着较强的搜索能力。通过猫群算法优化的 LBG 算法经训练后得到了与全局最优更接近的码本，进而完成了汉语音节识别。在部分测试样本的识别准确率上，传统的 VQ 算法达到了 93.03%，而采用猫群算法优化后的 VQ 法达到了 95.98%。

3　基于隐马尔可夫模型的算法

隐马尔可夫模型(Hidden Markov Model，HMM)是传统语音识别的主流模型，是由短时间内看作平稳变化的声学信号模型串联构成的马尔可夫链组成的[7]，表示一个双重随机过程：一重随机过程描述状态的转移；另一重随机过程描述状态和观察值之间的统计对应关系。文献[8]利用人工切分数据，对散文《师恩难忘》中的每个汉字分别建模，在统计单词并建立词典后将语音切分为单个音节，然后提取其 MFCC 参数并利用 HMM 进行识别，最后利用词典匹配算法在每个音节识别出的 3 个最有可能的结果中找出最终结果，识别的正确率达到了 85.6%。文献[9]将单音节识别应用到单词识别，首先将单词完整发音分割成音节片段，然后识别单元使用训练好的 HMM 来对这些音节片段进行识别，最后匹配单元匹配识别出的音节及其序列并获得识别单词。识别流程如图 2 所示。

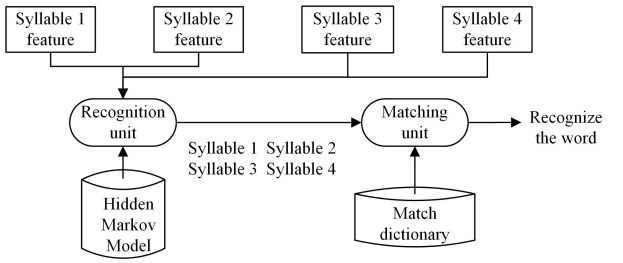

图 2　单音节识别流程

Fig. 2　Monosyllabic recognition process

文献[10]考虑到泰米尔语和其他印度语言基于音节的发音特征，使用短期能量作为幅度谱，将连续语音信号分割为音节单元，然后使用无监督增量聚类方法对相似的音节进行聚类，并手动对其进行标记，最后为音节簇创建基于 HMM 的声学模型，训练这些音节簇并用于转录连续语音，识别精度达到了 94.55%。

在经典的 HMM 中，假设模型中的状态转移具有齐次 Markov 性，状态 i 的驻留概率为常数 $a_{i,i}$，系统进入状态 i 后在该状态驻留的时间 T(即段长)服从式(4)所示的几何分布：

$$P(T) = a_{i,i}^{T-1}(1-a_{i,i}),\ T \geqslant 1 \tag{4}$$

这种模型对于描述语音的段长特征能力较弱，为了增强经典 HMM 对段长信息的表达能力，文献[11]提出了一种基于段长分布的 HMM。在这个模型中，转移概率与段长概率一一对应，且由段长概率唯一确定。将基于段长分布的 HMM 应用于音节识别中时，首先把模型的状态与语音中的单音节相对应，然后把单音节的语音信号特征作为音节单元的观测量。在识别非特定人汉语音节时，该模型的识别错误率比经典 HMM 模型下降了 17.8%，显示出了良好的识别性能。其结构如图 3 所示。

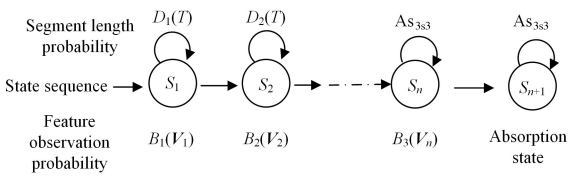

图 3　基于段长分布的 HMM 模型结构

Fig. 3　HMM model structure based on length distribution

为了缓解汉语中相邻音节间存在的协同发音问题，文献[12]引入了一种音节间的过渡模型。首先，使用 HMM 语音识别系统对每一个待识别的句子生成一个列表，然后针对列表中的每一条路径，找到所有相邻音节的过渡区域，提取特征序列并计算产生该序列的概率得分，最后将这些概率得分进行加权计算，得到路径总得分并进行排序，从而得到识别结果。该模型对于缓解音节间的协同发音问题起到了积极作用，在加入音节间过渡模型后，错误率下降了 12.13%。

在音节识别的研究中，有很多研究将 HMM 与其他算法相结合以提高识别准确率。文献[13]改进了传统的 VQ/HMM 算法，提出了一种 VQ 与 HMM 分级识别算法：先训练出汉语 1 200 个发音的全局码本，再针对 400 个基本音节设计并训练音节码本，第一级识别完成局部 VQ 到全局 VQ 标号序列的转变，第二级识别训练 400 个相对应音节的 HMM。这种方法将 4 个声调作为一个模型来训练，减少了储存量和计算量，并增强了识别的稳定性，使汉语 400 个基本音节的识

别率达到了96%以上。文献[14]结合SVM良好的分类性能和HMM稳定的识别性能，提出了基于HMM+SVM的音节识别算法。首先将汉语语音分类，并进行音节分割、特征提取等一系列处理，然后通过SVM将处理后的语音分到相应的音节类中，最后使用HMM进行识别。这种算法在加快识别速度的同时也提高了识别的准确率，最终识别率可达到93.5%。

HMM使用一个双重的随机过程来描述语音信号时间序列，对于HMM的状态输出概率，可采用多种概率分布函数来对其进行描述，传统的声学模型主流方案是GMM-HMM（高斯混合-隐马尔可夫模型），GMM用以对HMM每个状态描述的语音特征分布建模，其优点在于能平滑近似的模拟任意形状的密度分布，只要混合高斯分布数量足够，高斯混合模型能够逼近任何精度概率分布，并且非常适合用于描述短时和静态语音模式，该模型表示如式(5)所示：

$$P(O_t \mid S_j)=\sum_{n=1}^{N} W_n \cdot \frac{1}{\sqrt{(2n^D)\mid \sum_n}} \cdot e^{\left[-\frac{1}{2}(O_t-U_n)\sum_n^{-1}(O_t-U_n)^T\right]} \tag{5}$$

其中，状态S_j是对观测序列O_t的输出概率表示，U_n，$\sum_n$，W_n分别表示高斯混合分量的均值、方差、权重。其结构如图4所示。

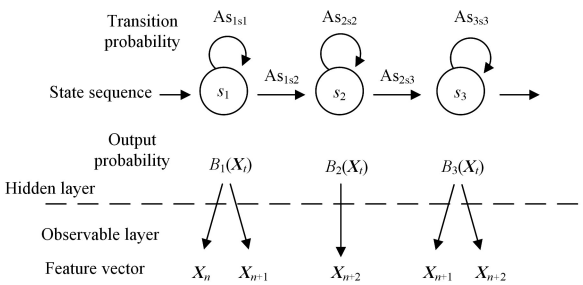

图4 GMM-HMM结构

Fig. 4 GMM-HMM structure

文献[15]提出了基于BN-GMM-HMM和BN+MFCC-GMM-HMM的声学建模。前者是提取Bottleneck(BN)特征代替传统的MFCC语音特征，并用于训练传统的GMM-HMM模型。相比传统的MFCC特征，BN特征不仅具有语音长时相关性，而且具有高度抽象表征信号的能力。后者是采用深度神经网络(Deep Neural Network，DNN)提取的具有时长性的39维瓶颈特征与传统的39维MFCC特征复合为一个78维的高维特征参数，然后降维处理为39维特征参数并用于GMM-HMM声学建模。以藏语音节识别错误率为准则，基于BN-GMM-HMM的声学模型的音节识别错误率降低了2.63%，基于BN+MFCC-GMM-HMM声学模型的音节识别错误率降低了4.12%。

4 基于神经网络的算法

DNN作为一种深层网络，比GMM有着更强的分类能力和描述复杂信号的能力，文献[16]综合利用了DNN强大的分类与描述复杂信号的能力以及HMM的时序建模能力，提出将DNN-HMM混合模型应用于音节识别中，实验结果表明该模型的识别能力优于传统的GMM-HMM模型。模型结构如图5所示。

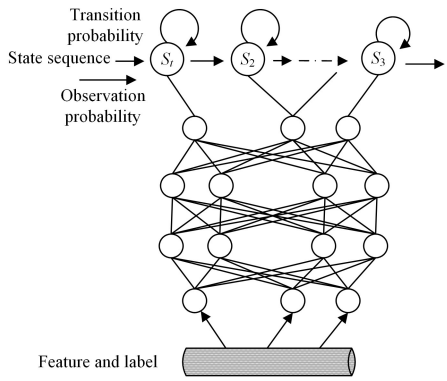

图5 DNN-HMM结构

Fig. 5 DNN-HMM structure

文献[17]提出了时间标签循环神经网络(Time-Tag Recurrent Neural Network，TTRNN)，对汉语拼音的整个音节进行建模识别，解决了声母韵母精确提取比较困难和相邻音节间协同发音等问题。TTRNN实际上是一种增强的循环神经网络，在网络的设计上，给每一帧增加了一个时间标签，使其能够离散不同时间的相同特征，解决了传统的RNN不能正确分辨处于不同时间的同一组帧的问题。在训练过程中，该算法常常陷入局部极小值，因此采用Back-Propagation Through Time[18]方法来进行训练。同时，由于各个语音的长度是不同的，而TTRNN不能处理长度不同的模式，因此本文借鉴Zhu[19]提出的方法，对各个语音的特征长度进行了规整。实验结果表明，该方法在孤立音节识别分类方面表现出了很好的识别效果和较强的鲁棒性，在识别数字语音时，该方法对训练数据的识别率达到了99.3%，对孤立音节测试集的识别率达到了98.2%，对孤立音节外测试集的识别率达到了96.5。时间标签循环神经网络结构如图6所示。

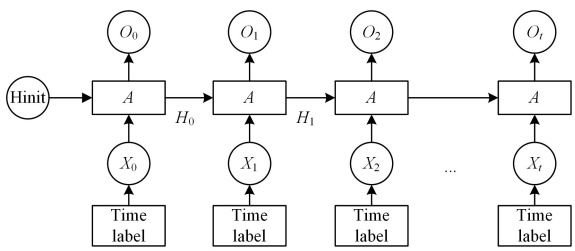

图6 时间标签循环神经网络结构

Fig. 6 Time label cyclic neural network structure

结束语 本文针对单音节识别，介绍了3个大类的发展现状及其主流模型。矢量量化能够实现语音信号的高效压缩，但其应用于语音识别时，随着系统用词量的增加，需要的码本数量将是天文数字，这种算法不足以实现大词汇量语音识别的高效性。HMM作为语音信号的一种统计模型，凭借其强大的时序建模能力，成为语音识别的主流研究途径。

随着深度学习的发展，大量研究表明，深度学习对于语音信号的特征提取、语音识别的声学建模有着出众的能力。将深度学习应用到语音识别当中以获得更高的准确率和更强的鲁棒性，是国内外众多研究人员孜孜以求的目标。

但目前在语音识别领域中，以音节为识别基元的研究较少，且在音节识别的现有研究中，将深度学习应用到单音节识别中的研究还远远不够。对于单音节结构语言，如汉语及一

(下转第203页)