计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2018,27(5):41−48 [doi: 10.15888/j.cnki.csa.006367]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于深度神经网络的关键词识别系统①

孙彦楠, 夏秀渝

(四川大学 电子信息学院, 成都 610065)
通讯作者: 孙彦楠, E-mail: tzsyn@163.com

摘　要: 针对当前关键词识别少资源或零资源场景下的要求, 提出一种基于音频自动分割技术和深度神经网络的关键词识别算法. 首先采用一种基于度量距离的改进型语音分割算法, 将连续语音流分割成孤立音节, 再将音节细分成和音素状态联系的短时音频片段, 分割后的音频片段具有段间特征差异大, 段内特征方差小的特点. 接着利用一种改进的矢量量化方法对音频片段的状态特征进行编码, 实现了关键词集内词的高精度量化编码和集外词的低精度量化编码. 最后以音节为识别单位, 采用压缩的状态转移矩阵作为音节的整体特征, 送入深度神经网络进行语音识别. 仿真结果表明, 该算法能从自然语音流中较为准确地识别出多个特定关键词, 算法易于理解、训练简便, 且具有较好的鲁棒性.

关键词: 关键词识别; 语音分割; 矢量量化; 深度神经网络

引用格式: 孙彦楠,夏秀渝.基于深度神经网络的关键词识别系统.计算机系统应用,2018,27(5):41–48. http://www.c-s-a.org.cn/1003-3254/6367.html

Keyword Recognition System Based on Deep Neural Network

SUN Yan-Nan, XIA Xiu-Yu

(College of Electronic and Information Engineering, Sichuan University, Chengdu 610065, China)

Abstract: A new algorithm for keyword recognition based on audio automatic segmentation and depth neural network is proposed to identify the requirements of keyword recognition on the condition of low or zero resource. Firstly, an improved speech segmentation algorithm based on metric distance is used to divide the continuous speech stream into isolated syllables, and then the syllable is subdivided into short audio segments which are connected with the phoneme state. The segmented audio segment has the characteristics of large difference between the segments, and the characteristic variance of the segment is small. Then, an improved vector quantization method is used to encode the state features of the audio fragments, and the high precision quantization coding and the low precision quantization coding of the words are realized. Finally, the syllable is used as the recognition unit, and the compressed state transition matrix is used as the whole feature of the syllable. It is sent into the deep neural network for speech recognition. The simulation results show that the algorithm can identify many specific keywords from the natural speech stream, and the algorithm is easy to understand, the training is simple and the robustness is better.

Key words: keyword recognition; voice segmentation; vector quantization; Deep Neural Network (DNN)

1　引言

关键词识别 (Keyword Recognition, KWR) 是从自然声音流中检测并确认出一个或几个特定关键词的技术. 其广泛应用于语音检索、人机交互、语音监听等社会经济生活领域. 关键词识别与一般语音识别最大的不同是关键词识别时会遭遇大量的集外词 (Out-Of-

① 收稿时间: 2017-09-11; 修改时间: 2017-09-30; 采用时间: 2017-10-20; csa 在线出版时间: 2018-03-12

Vocabulary, OOV), 但不需对这些集外词的内在信息作具体的识别.

现阶段关键词识别的方法主要有 3 种[1]: 1) 利用动态时间规整算法基于滑动匹配思想的关键词检出方法. 它利用滑动窗口在连续语音流中进行搜索、匹配计算, 进而检出关键词. 这类方法关键词识全率和识准率均不是很高. 2) 利用隐马尔可夫算法基于垃圾模型的方法. 这种方法不仅需要对每个关键词建模, 还需要对多种集外词 (其他音节、自然声音等) 建立不同模型, 即垃圾模型. 然后用垃圾模型与关键词模型共同搭建网络, 最后采用维特比解码得出结果. 该方法需要一个较为全面的语料库建模, 识准率受关键词的规模影响较大, 模型训练和识别匹配的运算量巨大, 且当关键词发生变化时需要重新训练模型. 3) 基于文本的关键词检出方法, 该方法通过一个大词汇量连续语音识别系统识别待检音频, 再对结果进行搜索, 最终确定这段被测语音是否包含关键词. 这种方法需要大量的标注数据资源.

近几年以来, 少资源或零资源场景下的关键词识别由于其广泛的适用性得到广泛的关注. 少资源或无资源是指缺乏足够标注的目标样本语音数据, 并不具备训练一个鲁棒的大词汇量语音识别系统的条件[2]. 深度神经网络 (Deep Neural Network, DNN) 凭借其无监督学习的能力在连续语音识别技术领域得以广泛应用并取得了相比于以前更好的识别性能. 因此, 本文针对少资源或零资源情况提出了一种基于自动音频分割技术和深度神经网络的连续语音流关键词识别方案.

2 系统原理及方案设计

2.1 系统总体框架

关键词识别系统的总体框架如图 1 所示.

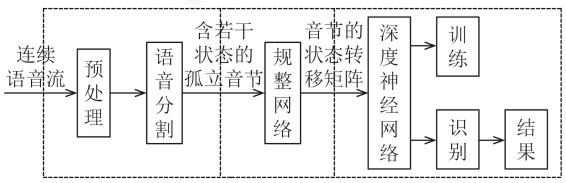


图 1 系统框架

整个系统分为 3 部分: 1) 语音预处理及分割模块, 该模块将连续语音流自动分割为不同大小的音频段, 涉及有声无声段的分割, 音节和音素状态的分割等. 2) 音节状态转移矩阵生成模块, 该模块在语音分割模块输出结果的控制下完成音节参数特征的提取和时间规整. 3) 基于深度神经网络的关键词识别模块, 该模块分为训练和测试两个模块, 训练模块完成网络权值的学习, 测试模块完成关键词的识别.

2.2 预处理及语音分割

针对自然连续语音流, 语音预处理及分割模块完成有声无声段, 音节和音素的自动分割. 具体框图如图 2 所示.

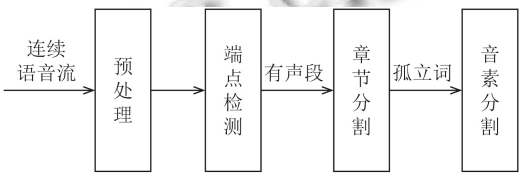


图 2 预处理及语音分割流程

预处理包括提升语音高频部分的预加重和分帧处理. 端点检测是将语音流中的无声段去除, 只保留有声部分. 目的是减少后续计算量, 同时还可以提高语音识别的正确率. 本文使用频谱方差作为端点检测的特征参数, 该特征参数在去除静音和噪声的同时, 可以避免将语音段中的轻音部分认定为是无声段. 采用双门限判定法[3]来检测语音, 用高门限判断是不是有声段, 用低门限确定有声段的起止点位置.

音节是人类听觉可以区分清楚的语音基本单位, 在汉语中一个汉字的读音就代表一个音节. 音节分割模块将连续有声段语音分割成一个个孤立的音节, 用于后续以音节为声学单元的关键词识别.

观察语音的语谱图, 我们可以很容易的区分每一个音节. 这是因为音节作为一个短时音频段落, 其听觉谱段间差异较大, 段内差异较小, 因此可以采用基于度量距离的算法实现分割.

我们提出一种综合数据段段间均值和方差的度量方法, 用来表征音频段落之间的差异, 简称为 *DIS*[4,5]:

$$DIS = \frac{(\mu_1-\mu_2)^{\mathrm{T}}(\mu_1-\mu_2)}{[b\cdot tr(\Sigma_1)+(N-b)\cdot tr(\Sigma_2)]/N} \tag{1}$$

式 (1) 中分子表示左右两段音频特征各自均值的差异, 分母反映两段音频特征方差的平均值. 当两段音频之间特征均值差异较大, 段内特征方差小时, *DIS* 越大, 表明两段音频段间距离越大.

假设特征向量参数各维度独立, 特征维度为 *D*, 协

方差矩阵简化为对角阵, 则:

$$tr(\Sigma)=\sum_{d=1}^{D}\sigma_d^2 \tag{2}$$

本文特征参数选用短时能量参数和 12 维 Mel 频率倒谱系数共 13 维. 为简化计算, 我们将式 (1) 就简化为:

$$DIS=\frac{(\mu_1-\mu_2)^{\mathrm{T}}(\mu_1-\mu_2)}{tr(\Sigma_1)+tr(\Sigma_2)} \tag{3}$$

音节分割分 3 步完成, 分别是计算 *DIS*、取极大值点和分割点确认. 具体做法为分窗计算 *DIS* 距离值, 逐帧滑动得到一系列 *DIS* 值, 提取距离值曲线上的极大值点, 为了消除音频失真带来的误差, 两个极大值点距离很近时只取一个, 然后利用阈值判断其是否为分割点, 当极大值点的 *DIS* 值超过预设门限 T-DIS 时, 判断为分割点, 否则舍去. 分割点确认后, 相邻分割点间的音频段落即为音节, 为减小信息丢失, 采取了重叠分段的思想, 左右两个分割点分别向左向右扩充 3 帧构成最终的分割段.

单个音节内部的听觉谱相对其他音节来说差异较小, 但仍然可以继续细分成更小的段落 (比如音素).因此, 我们可以继续利用公式 (3) 设置更小的滑动窗将音节划分成差异更加细微的音频段落, 这些段落可能是某个音素, 可能是某个音素到静音之间或者某两个音素之间的过渡部分. 在本文中, 我们将其称之为音素级段基元. 音素级段基元虽然长度各不相同, 但其内部各帧之间状态变化很小. 本文取其内部核心帧的特征参数均值作为该段音频的状态. 这样不仅压缩了数据量简化了后期的运算量, 同时也避免了同一种音节由于持续时间不同导致的特征参数差异较大的问题.

通过以上处理, 我们就将连续语音流分割成一个个含有若干状态的独立音节, 同时将连续语音识别问题转化为包含有未知信息的孤立词识别问题.

2.3 音节的状态转移矩阵

本文采用音节作为声学识别单元, 状态转移矩阵作为音节整体特征被送入后级神经网络进行识别. 生成状态转移矩阵的框图如图 3 所示.

经过音频分割, 每个音节可以用少量几个状态的特征参数表示, 为进一步压缩数据, 可以采用矢量量化对特征参数进行数据压缩. 应用聚类算法如 K-means 算法进行矢量量化时, 采用欧式距离作为相似性的评价指标. 传统的矢量量化以数据 x 和码字 $Y_j(j=1,2,\cdots,N)$ 的最小距离作为唯一评价指标, 以此确定区域边界, 寻找最佳划分 (胞腔). 这种方法对所有数据进行等精度的量化, 但对于关键词识别来说, 只需要识别少量的集内词, 对大量的集外词不需要作具体的识别. 如果由所有数据 (集内词和集外词) 来确定码字个数, 较多的码字相对于少量的集内词来说过于浪费, 而当码字的数量较少时量化精度得不到保证, 集内词之间就会缺乏区分性. 对于关键词识别, 我们希望集内词之间有较高的量化精度, 同时总的码字数量较少, 所以本文提出了一种改进型的矢量量化方法.

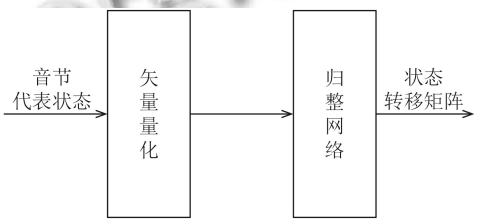


图 3 状态转移矩阵形成简图

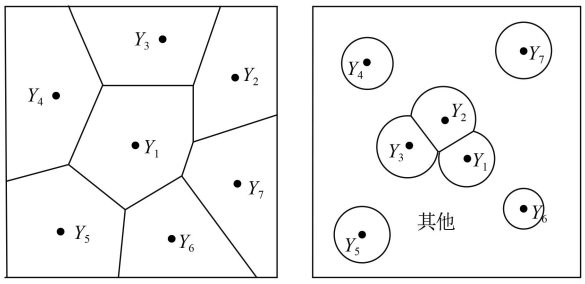
改进型矢量量化的码书生成只采用关键词数据 (集内词) 进行训练, 获取有意义的精度较高的码书. 确定各胞腔的最佳区域边界采用了两个评价指标, 分别是: 数据 x 和码字 $Y_j(j=1,2,\cdots,N)$ 间最小的距离值 $D_{\min}$, 码字 $Y_j(j=1,2,\cdots,N)$ 空间大小 T. 最佳区域边界的确定必须满足以下两个公式:

$$\begin{cases}D_{\min}=\min\{d(x,Y_1),\cdots,d(x,Y_N)\}\\ D_{\min}\le T\end{cases} \tag{4}$$

语音信号的特征参数每一维可以近似看作高斯分布[6], 我们依据 3σ 准则来确定距离范围, 取胞腔内所有数据的码字距离标准差的三倍 $T_{3\sigma}$ 作为距离范围指标. 本文实验发现, 只要训练使用的关键词音节样本数量足够, 最后得到的每个码字的边界大小 $T_{3\sigma}$ 大致相同.

图 4 为传统矢量量化与改进型矢量量化胞腔边界的比较图.

矢量量化的具体过程如下: 对于测试用关键词数据或集外词数据, 先计算其特征参数和各码字的欧式距离并确定最小距离, 然后再判断该最小距离值是否在 $T_{3\sigma}$ 范围内, 若满足条件则将其归属为某一个高精度有意义的量化中心. 如果超出该范围, 则认定其不属于该码书所在的胞腔, 将此种情况下的特征参数全部另归为其他一类, 也就是由集外词等垃圾信息确定的一个低精度码字.



(a) 传统矢量量化效果图 (b) 改进型矢量量化效果图

图 4 传统矢量量化与改进型矢量量化的比较

这种改进型的矢量量化只对关键词相关特征参数敏感, 对无关音节的特征参数不敏感, 保证了关键词的量化精度, 满足识别要求的同时大大减小了矢量量化的工作量.

语音具有时变性, 不同字发音不同, 不同人或同一人在不同环境下读同一个字的发音都不同, 包括声音大小, 重音位置, 持续时间都有所不同. 音频信号的丰富变化使得一个音节的音频特征参数序列的长度是可变的, 而大多数神经网络要求输入数据的结构固定[7]. 为了反映语音时变特征同时适应神经网络输入的要求, 本文将特征参数序列通过时间规整网络转换为状态转移矩阵, 状态转移矩阵能反映语音信号时变特征且维数固定.

通过前面矢量量化, 将关键词音素级段基元的特征参数归属于某个确定码字 (音素状态), 这样每一个音节都对应为一个状态序列 $O=\{O_1, O_2,\cdots, O_N\}$. N 是一个音节音素状态的总数量, O_i 是第 $i(i=1, 2,\cdots, N)$ 个状态的标号, O_i 的最大值为码本个数 K. 时间规整网络的输出是一个 $K\times K$ 的矩阵, 在文本, 我们用 TRM(m, n) 代表输出矩阵第 m 行第 n 列的元素 ($m=1, 2,\cdots, K$; $n=1, 2,\cdots, K$). 定义:

$$CO_j(m,n)=\begin{cases}1, & O_j=m \text{ 且 } O_{j+1}=n\\ 0, & \text{其他}\end{cases} \tag{5}$$

则:

$$\mathrm{TRM}(m,n)=\sum_{j=1}^{N-1} CO_j(m,n) \tag{6}$$

公式 (5) 反映了音节内任意两个相邻状态的状态转移情况, 通过公式 (6) 映射到输出矩阵 TRM 对应的坐标 (节点) 上, 统计音节所有状态转移的情况得到最终的输出矩阵. 通过这种方式, 就可以将长度不同的音节规整为格式统一的状态转移矩阵.

状态转移矩阵作为音节的整体特征矢量可以直接馈入下一级神经网络完成语音识别, 但这样一个多达 2500 维的输入依然显得过于庞大. 因为每个音节只有 2~7 个状态, 所以每个音节的状态转移矩阵是一个稀疏矩阵, 其大多数节点的值为 0. 观察图 5 所有训练用关键词音节的状态转移矩阵累加图, 我们发现大多节点的响应很小甚至为 0, 这也就意味着, 对关键词识别来说, 这些节点是多余的, 其输出不需要送入下级网络. 因此, 我们通过设定阈值筛选出有用的节点, 以减少后续神经网络的规模. 本文通过这种方式筛选出 157 个节点, 后续神经网络的规模大大减小了.

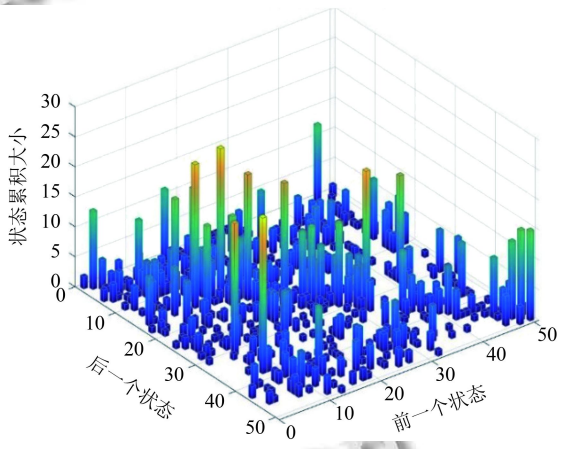


图 5 所有训练用关键词音节的状态转移矩阵累加图

本节以音节为识别单位, 生成压缩的状态转移矩阵作为音节的整体特征, 该特征反映了语音的部分时序特性, 同时完成了音节的时间规整. 音节的状态转移矩阵将作为后续识别神经网络的输入.

2.4 基于深度神经网络的语音识别

深度神经网络是一个有输入层, 超过两个的隐层和输出层的非线性转换单元的多层感知器[8], 如图 6 所示. 与“浅层”神经网络相比, 深度神经网络拥有更强大的建模和表征能力, 能够实现复杂函数的逼近. 本文选用深度神经网络作为语音识别模块, 可以有效提高系统分类识别性能.

深度学习是针对模型具有“深层”结构的网络权值学习算法, 能够有效解决仅采用反向传播算法所造成的训练容易陷于局部最优解的问题, 解决深层网络无法调整到神经网络低层参数而出现的性能急剧下降的问题, 可以有效的抑制训练过程中的过拟合现象等. 深

度学习核心的内容就是利用无监督性的学习, 消除信号中的冗余信息, 提炼具有高效分类能力的特征, 提高算法的识别率. 常用的深度神经网络模型有深度信念网络 (Deep Belief Network, DBN)、卷积神经网络 (Convolutional Neural Network, CNN)、稀疏自编码器 (Sparse Auto-Encoder, SAE)[9]等模型, 这些模型试用于不同的数据. 深度信念网络是一个概率生成模型, 其目的在于建立观察数据和标签之间的联合分布, 对于本文这种比较稀疏的数据, 学习容易收敛于局部最优解. 卷积神经网络是人工神经网络的一种, 它的权值共享网络结构降低了网络模型的复杂度, 减少了权值的数量. 但其在语音识别中, 通常使用若干帧梅尔倒谱系数作为数据, 有利于解决语音的时变性问题、降低学习复杂复. 稀疏自编码器试图找出每组输入数据的类似于线性代数中基的概念的一组基的线性组合, 所以其对较为稀疏的数据有着学习过程快, 学习性能优异且稳定的特点. 因此本文选用稀疏自编码器作为进行语音识别的模型.

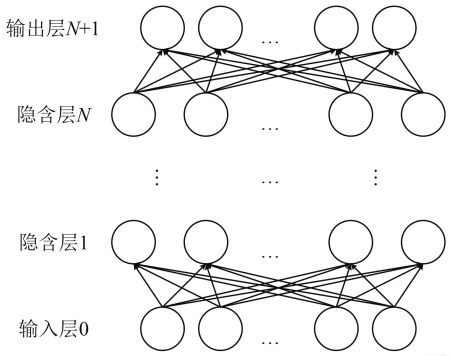


图 6 深度神经网络简图

稀疏编码算法是一种无监督学习方法, 其本质是本文上节提到的 K-means 算法的变体, 它寻找一组“超完备”基向量来更高效地表示样本数据. 超完备基的好处是它们能更有效地找出隐含在输入数据内部的结构与模式[10]. 稀疏自编码器的结构如图 7 所示, 稀疏自编码器在自编码器的基础上对网络编码层输出进行约束, 仅有少部分节点处于激活状态, 其余节点均处于未激活状态. 用最少个数的编码层神经元输出来表示输入数据, 对数据进行降维. 它利用低阶特征进行线性稀疏组合成高阶特征来表征原有信号, 筛选出信号中的显著性原子, 这对提高语音识别的识别率具有重要意义.

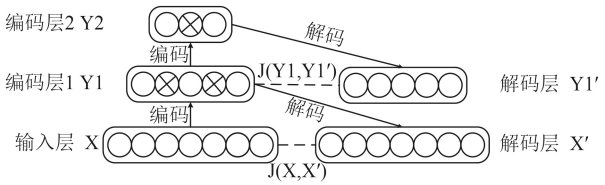


图 7 稀疏自编码器

本文利用深度稀疏自动编码机神经网络进行语音识别, 步骤如下:

a) 提取训练语音的音节状态转移矩阵, 将其作为网络输入训练第一个编码层的网络参数. 并将训练好的编码数据作为第一编码层的输出.

b) 把步骤 a) 的输出作为输入, 用同 a) 一样的方法训练第二编码层的网络参数. 类似的方式可以逐层训练出更多的隐层参数.

c) 把步骤 b) 的最后一级隐层输出作为 Softmax 回归模型的输入, 然后利用原始数据的标签 (硬分类为某个关键词或者为集外词) 监督性训练得到 Softmax 分类器.

d) 级联稀疏自动编码机和 Softmax 分类器, 生成全局网络. 计算整个网络的误差函数及其对各个参数的编导值, 更新权值.

e) 采用 LBFGS 算法进行整个网络的权值优化计算, 通过误差反向传播算法微调网络参数, 提高分类器的精准性.

f) 将测试数据的状态转移矩阵送入到训练好的神经网络中进行识别测试.

本节以音节的压缩状态转移矩阵作为神经网络的输入, 使用稀疏自编码器组合低阶特征组完成信号高阶特征表示, 然后使用 Softmax 分类器分类识别多个关键词和集外词. 该算法操作简单, 不同关键词特征区分明显.

3 系统实现及实验分析

3.1 实验设计

本文的音频数据由 3 部分构成, 第 1 组数据由微机上的声卡在实验室采集, 7 男 3 女共 10 人分别用慢语速, 正常语速和快语速连续录入普通话朗读 0~9 共 10 个音节, 每人每种语速重复两遍共 60 组数据, 30 组用于训练, 另外 30 组用于识别测试; 第 2 组数据是用于建立垃圾模型只包含集外词的语音流、非语言语音流 (雷鸣声、鸟叫、猫叫等自然界声音和人发出的咳

嗽、喘气等) 共 10 段, 每段平均长度为 3 min; 第 3 组数据是采集于广播传媒, 包括含有普通话数字 0~9 和集外词的语音流共 30 段 (每段时长平均为 10 min). 3 组数据采样频率均为 8000 Hz.

$$\bar{o}_i = o_i - \mu_i \tag{7}$$

在进行矢量量化之前需要对样本特征作归一化处理, 本文利用公式 (7) 进行倒谱均值归一化 (CMN)[9], 以减弱潜在的声学信道扭曲带来的影响.

公式 (7) 将每一个维度 i 的特征 O 归一化为均值为 0 的实数特征, 其中$\bar{o}_i$、o_i、μ_i 分别表示某帧第 i 维归一化后的特征参数、原始参数、均值.

首先利用我们提出的音频分割方法将音频数据分割成孤立音节; 接着利用微机采集的关键词数据 (第 1 组中的训练数据) 训练得到码书, 得到 50 个量化中心, 利用训练好的码书就可以对第 1 组中的测试数据、第 2 组数据和第 3 组数据的每个音节状态进行矢量量化了; 最后, 利用归整网络得到每个音节的状态转移矩阵, 压缩后共抽取 157 维构成神经网络的输入.

神经网络输入层结点为 157 个, 2 个隐含层神经元节点数量依次为 80, 40, 输出层神经元节点数量为 11 (10 个节点表示 10 种关键词, 另外 1 个节点表示集外词), 学习率为 0.8e–4.

3.2 语音分割实验

1) 连续语音流的音节分割

首先利用双门限判定法检测语音起止点, 结果如图 8 所示.

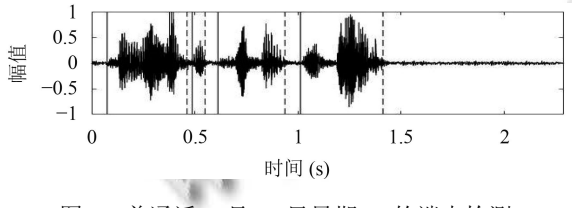


图 8 普通话“9 月 27 日星期二”的端点检测

图 8 中实线表示有声段起始时刻, 虚线表示有声段终止时刻. 然后对有声段进行音节分割, 结果如图 9 所示.

图 9 中在语音波形图和语谱图中虚线代表人工分割的位点, 实线代表 DIS 算法分割的位点. 使用观测窗长分别为 4、7、9 帧的 DIS 值按照 0.2、0.5、0.3 的加权系数加权求合, 确定用于音节分割的综合 DIS 值, 根据综合 DIS 值的平均值确定预设门限 T-DIS, 根据 T-DIS 在极大值中寻找分割点. 由图 9 可见, DIS 算法的音节分割点和人工分割点基本吻合.

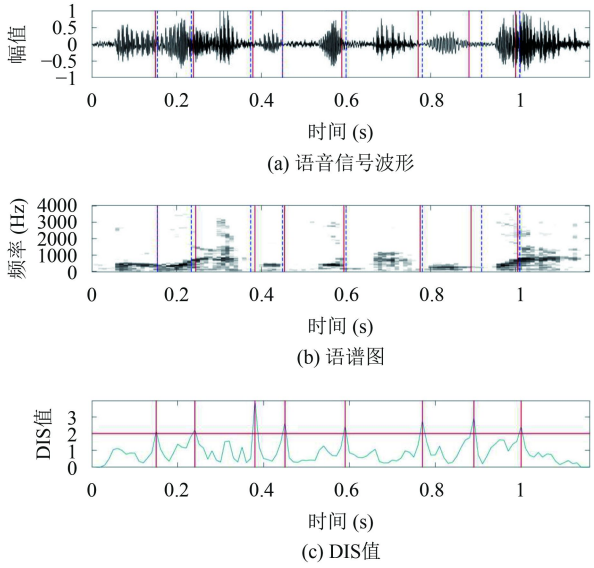


图 9 普通话“9 月 27 日星期二”的音节标注

我们以人工标注点为基准, 分析了音节自动分割的效果. 采用两个评价指标: 音节分割率和音节分割精准度.

音节分割率 α 定义为:

$$\alpha = 1 - \frac{|\nu_{\text{rel}} - \nu_{\text{DIS}}|}{\nu_{\text{rel}}} \tag{8}$$

ν_{rel} 表示人工标注分割的音节数目, ν_{DIS} 表示 DIS 算法分割的音节数目.

音节划分的精准度 β 定义为:

$$\beta = 1 - \frac{\sum_{k=1}^{M} \frac{|T_{\text{sk}} - T'_{\text{sk}}| + |T_{\text{ek}} - T'_{\text{ek}}|}{E_{\text{ck}}}}{M} \tag{9}$$

T_{sk}, T_{ek}, E_{ck} 分别表示人工标注的第 k 个音节的开始时间、结束时间和持续时长. T'_{sk}, T'_{ek} 分别表示 DIS 分割算法的第 k 个音节的开始时间、结束时间, M 表示实际音节的个数.

统计的结果如表 1 所示. 第 1 组数据是在安静的实验室环境下录制的普通话数字 0~9, 可以代表传统意义的孤立词. 第 2 组数据为广播传媒包含有普通话数字 0~9 和集外词的语音流.

实验表明, DIS 分割算法能够比较有效的将语音流中的音节单独分割开来.

2) 音节内部的状态分割

利用 DIS 算法, 采用大小为 4 帧的分析窗, 还可进一步将音节分割为不同的音素状态. 图 10 为数字 7 的音素状态分割情况.

表 1 DIS 算法音节分割效果

	分割率(α)(%)	精准度(β)
第1组	100	0.93
第2组	97	0.87

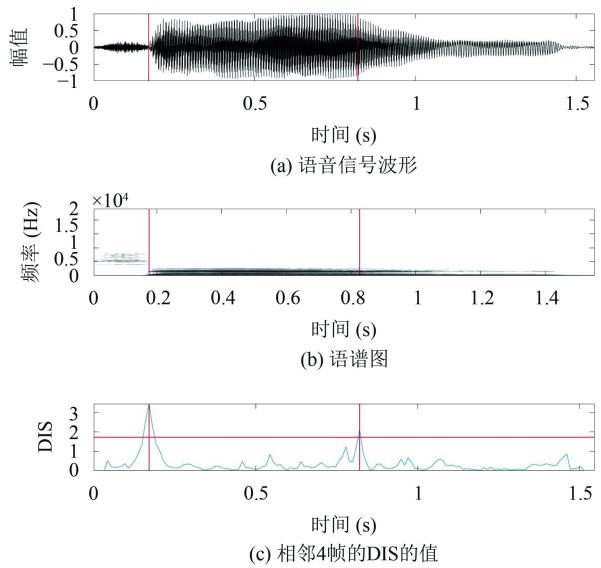


图 10 数字 7 的 DIS 分割图

从图 10 可见, 每个音节内部听觉谱随时间还是有所变化, 通过 DIS 算法将音节进一步细分为音素级段基元, 段基元内部各帧状态变化很小.

3.3 孤立词 (集内词) 识别实验

利用语音分割算法将连续语音流分割成孤立的音节, 就可以用类似孤立词识别的方法进行关键词 (集内词) 识别了, 本文基于神经网络进行语音识别. 语音特征参数送入神经网络时, 需要进行时间规整处理[11,12], 采用语音信号处理中常见的非线性时间规整方法, 以各帧参数间欧式距离作为参考量, 将音节特征参数序列规整为固定的帧数, 作为神经网络的输入. 本组实验进行集内词 (1~9) 识别实验. 首先通过人工分割或 DIS 自动分割出第 1 组和第 3 组数据中所有的集内词, 然后使用第 1 组中的训练数据进行训练, 使用第 1 组中的识别测试数据和第 3 组中的集内词数据进行识别测试.

表 2 比较了基于人工分割和 DIS 自动分割的音节识别结果, 对比了最小距离时间规整法和本文时间规整算法的语音识别结果, 其中语音识别均采用深度神经网络完成.

表 2 集内词识别结果 (单位: %)

	非线性时间规整识别率	本文算法识别率
人工分割	94.7	93
DIS分割	68.8	89

实验结果表明, 在人工分割的情况下, 本文方法和经典的最小距离法都可以克服不同说话人的干扰, 具有较高的识别率, 二者均能够良好的发挥深度神经网络的优势. 而在 DIS 自动分割的情况下, 由于分割是存在误差的, 导致经典的最小距离法识别率迅速下降, 而本文方法由于时间规整网络采用了状态转移矩阵, 比较好的保持了关键状态转移信息, 识别率下降不多, 对音节分割不精确表现出良好的鲁棒性.

3.4 关键词识别实验

关键词识别是从自然声音流中检测并确认出特定的关键词. 本组实验基于第 3 组数据完成, 关键词为 0~9. 与 2.3 节不同的是, 关键词识别时会遭遇大量的集外词, 这些词不需具体识别, 但需和关键词相区别, 所以相比实验二神经网络的输出端新增了一个输出端代表集外词.

在关键词识别技术领域, 常使用以下指标对系统性能进行评价:

$$\text{识全率} = \frac{\text{识别为关键词的数量}}{\text{待检关键词的数量}} \times 100\%$$

$$\text{识准率} = \frac{\text{正确识别的关键词数量}}{\text{待检关键词的数量}} \times 100\%$$

$$\text{虚警率} = \frac{\text{错误识别为关键词的信息量}}{\text{待检关键词的数量}} \times 100\%$$

$$\text{综合检出率} = \frac{\text{正确分类的信息数量}}{\text{总的信息数量}} \times 100\%$$

基于第 3 组数据, 关键词识别结果如表 3.

表 3 关键词识别结果

关键词总数	信息总数	识全率	识准率	虚警率	综合检出率
486	1828	95.6%	88%	4.4%	95.4%

实验结果表明, 相比于大词汇量的语音识别系统, 本文关键词识别方法得益于改进的矢量量化运用, 系统开销大大降低, 而识别性能尚可. 音频自动分割技术和状态转移规整网络解决了音频动态特征的表达问题,

再结合深度神经网络强大的分类能力, 系统花费的时间更短, 费效比更低. 本文关键词识别方法受说话人的影响较小, 在没有使用第 3 组中说话人数据训练的情况下而对其识别 (少资源场景), 对第 3 组数据中的关键词具有较高的识别率.

4 结论

本文提出了一种易于对接神经网络的关键词识别方法. 首先利用 DIS 算法将语音流分割成独立的音节, 然后通过规整网络找到能反映音频信号动态信息的特征参数, 较为准确的描述了关键词的语义信息. 深度神经网络算法简便易用, 识别结果较为准确, 时间花费度比较低, 可适于多种规模的关键词和集外词情况. 实验结果表明, 本文所展示的系统可以在少资源或者零资源 (low or zero-resource) 场景下较为准确在自然语音流的识别出多个特定关键词, 降低了说话人口音的影响, 且当集外词的规模增大时, 识全率和识准率只有很小的下降, 具有较好的鲁棒性.

参考文献

1 Xu H, Yang P, Xiao X, *et al*. Language independent query-by-example spoken term detection using N-best phone sequences and partial matching. Proceedings of 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). South Brisbane, Queensland, Australia. 2015. 5191–5195.

2 Chan W, Jaitly N, Le Q, *et al*. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. Proceedings of 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Shanghai, China. 2016. 4960–4964.

3 宋知用. MATLAB 在语音信号分析与合成中的应用. 北京: 北京航空航天大学出版社, 2013.

4 孙卫国, 夏秀渝, 乔立能, 等. 面向音频检索的音频分割和标注研究. 微型机与应用, 2017, 36(5): 38–41.

5 Kamper H, Jansen A, Goldwater S. A segmental framework for fully-unsupervised large-vocabulary speech recognition. Computer Speech & Language, 2017, (46): 154–174.

6 Bahdanau D, Chorowski J, Serdyuk D, *et al*. End-to-end attention-based large vocabulary speech recognition. Proceedings of 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Shanghai, China. 2016. 4945–4949.

7 Sharma P, Abrol V, Sao AK. Deep-sparse-representation-based features for speech recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2017, 25(11): 2162–2175.

8 Hinton G, Deng L, Yu D, *et al*. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Processing Magazine, 2012, 29(6): 82–97.

9 Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge, MA: The MIT Press, 2016.

10 Yu D, Deng L. Automatic speech recognition: A deep learning approach. New York, NY, USA: Springer Publishing Company, 2015.

11 张欣, 夏秀渝, 王雪君. 一种听觉显著图提取模型. 四川大学学报 (自然科学版), 2014, 51(2): 292–298.

12 侯靖勇, 谢磊, 杨鹏, 等. 基于 DTW 的语音关键词检出. 清华大学学报 (自然科学版), 2017, 57(1): 18–23.