基于 HMM 的混响环境下语音识别研究

叶 硕,杜珍珍,彭春堂,贺 娟
(武汉邮电科学研究院,湖北 武汉 430000)

摘 要:语音识别是实现人机交互的关键技术之一。当语音信号处于狭小环境时,源信号将与延迟衰减后的信号叠加在一起,从而引起混响,导致信号失真、降低了语音的清晰度。为提高语音识别系统的性能,提出一种使用卷积同态滤波器去混响的方法,并用隐马尔可夫模型对语音的时序进行建模。隐马尔可夫模型是一种广泛用于语音识别的、用于描述随机过程统计特性的概率模型,使用前向后向算法降低计算复杂度,使用 Baum-Welch 算法得到重估模型参数,使用 Viterbi 算法找到最优的语音识别结果。实验结果表明,在无噪声环境下,该模型在识别正常语音时具有较高的可靠性,实现了短词汇非特定人的语音识别,并能有效解决语音混响问题。相较于未处理的混响语音,识别正确率提高了 4% ~5%,较好地实现了混响环境下的语音识别。

关键词:语音识别;混响;卷积同态滤波;隐马尔可夫模型
中图分类号:TP301 **文献标识码**:A **文章编号**:1673-629X(2019)08-0076-05
doi:10.3969/j.issn.1673-629X.2019.08.015

Research on Speech Recognition under Reverberation Environment Based on HMM

YE Shuo,DU Zhen-zhen,PENG Chun-tang,HE Juan
(Wuhan Research Institute of Posts and Telecommunications,Wuhan 430000,China)

Abstract:Speech recognition is one of the key technologies for human-computer interaction. When the speech signal is in a narrow environment,the overlapping of the delayed and attenuated signal and the source signal will cause the reverberation,which will lead to signal distortion and speech clarity reduction. In order to improve the performance of speech recognition system,we propose a method of using convolution homomorphic filter to remove reverberation with the hidden Markov model to model the time-series voice. The hidden Markov model is a probabilistic model widely used in speech recognition to describe the statistical characteristics of stochastic processes. In this paper,we use the forward-backward algorithm to reduce the computational complexity,Baum-Welch algorithm to revaluate model parameters and Viterbi algorithm to find an optimal speech recognition results. Experiment shows that in a noiseless environment,the method proposed has high reliability in the recognition of normal speech,realizes the speech recognition of short words of non-specific person and can effectively solve the problem of voice reverberation. Compared with the untreated reverberation speech,the recognition accuracy rate is improved by 4% ~5%,achieving the speech recognition under reverberation environment.

Key words:speech recognition;reverberation;convolution homomorphic filtering;hidden Markov model (HMM)

0 引 言

语音识别是当今的热门研究之一,自动语音识别(automatic speech recognition,ASR)技术是实现人机交互的关键[1]。在人机交互过程中,非特定人的语音识别具有广阔的应用前景。隐马尔可夫模型(hidden Markov model,HMM)是语音识别技术中的重要模型之一,根据不同需要对建模对象进行变化,可实现不同作用,比如说话人识别[2]、连续语音识别[3]、情绪识别等[4]。

语音识别环境多变,当人处于狭小环境或声源距离声音采集器较远时,由墙壁、物体反射的延迟且衰减后的声音信号与源信号叠加在一起引起混响。这种与原始信号叠加形成的干扰,会导致卷积失真,大大降低了语音的清晰可懂度。

在去混响领域,学者们做过许多研究,如文献[5]对混响声场进行了细致的分类,文献[6]分析了不同

收稿日期:2018-10-15 **修回日期**:2019-02-20 **网络出版时间**:2019-03-28
基金项目:2018 年度湖北省技术创新专项重大项目(2018AAA063)
作者简介:叶 硕(1993-),男,硕士生,研究方向为信号处理和语音识别。
网络出版地址:http://kns.cnki.net/kcms/detail/61.1450.TP.20190327.1633.076.html

手段去混响的效果。文中基于 HMM,提出一种使用卷积同态滤波器去混响的方法。在预处理阶段对混响语音进行降噪,提高语音在混响环境下的识别精度,并借助 MATLAB 完成仿真。

1　语音去混响方法研究

完整的识别过程如图 1 所示,可以分为语音采集、特征学习、识别匹配三个阶段。其中 In1 为训练模型所用的语音信号,In2 为待识别语音信号,Out 为输出识别语音。用于提取特征的语音,通常采集于无干扰的实验室环境,而待识别的语音,通常来自于存在干扰的环境。

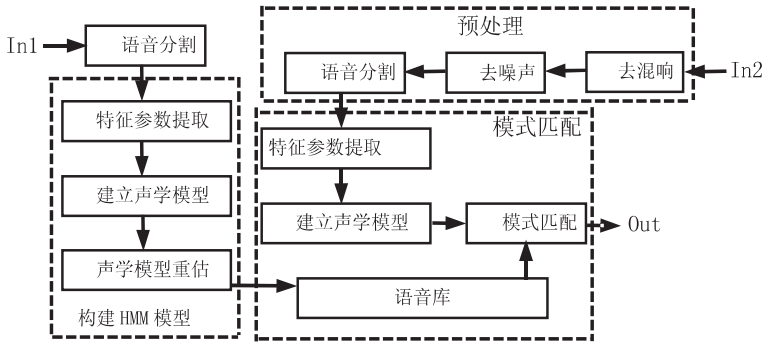

图 1　基于 HMM 的语音识别系统框图

在某些场合,适度的混响会在听觉上使人感到舒适,但在信号处理领域,混响不但会导致语音信号幅值变化、相位延时、共振峰偏移以及产生其他谱峰,其拖尾的混响声部分还会掩盖后面语音的弱能量音部分。在进行语音识别时,混响使得测试集参数和训练集参数发生匹配失真,识别系统性能发生急剧下降[5]。

混响环境下,接收到的信号可用如下数学模型表示[6]:

$$x(n) = s(n) + \sum_{k=1}^{M} a_k s(n - n_k)$$

其中,$s(n)$ 为有用信号;a_k 为反射系数;n_k 为反射波时延。$x(n)$ 又可表示为有用信号 $s(n)$ 与混响环境下的冲激响应 $h(n)$ 的卷积形式:

$$x(n) = s(n) * h(n)$$

$$h(n) = \delta(n) + \sum_{k=1}^{M} a_k \delta(n - n_k)$$

为解决混响问题,现行方法大致有三种:基于复倒谱滤波(complex cepstrum filtering,CF)、基于复倒谱均值减(complex cepstrum mean subtract,CMS)以及基于波束形成(beam forming,BF)。文中采用复倒谱滤波的方法,该方法本质上属于同态滤波,能将非线性结合的信号变换为加性结合的信号,从而可以使用线性滤波的手段去除混响[7]。卷积同态系统由三个子系统级联而成:特征子系统、线性系统、特征子系统的逆系统。

特征子系统实现原理如图 2 所示,其作用是将信号的时域卷积形式变换为时域的线性运算。输入信号 $x(n)$,经 Z 变换后变为两信号频域的相乘,取对数进而得到频域线性运算,再进行逆 Z 变换,得到原始信号的时域线性形式,即复倒谱:

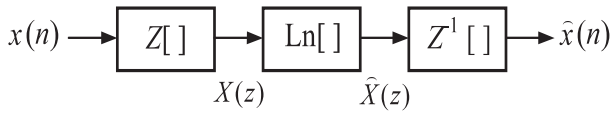

图 2　特征子系统实现

$$x(n) = Z^{-1}\{\ln\{Z[x(n)]\}\} = Z^{-1}\{\ln[S(z)]\} + Z^{-1}\{\ln[H(z)]\}$$

对 $x(n)$ 进行线性滤波,再通过特征子系统的逆系统便可提取出有用信号 $s(n)$。

2　HMM 的语音识别算法

2.1　HMM 的建立

孤立词或短词汇的语音识别算法一般分为两类:动态时间规整(dynamic time warping,DTW)和 HMM。

DTW 算法基于动态规划(DP)的思想,利用语音中逻辑的先后不可改变这一特性,能克服因发音习惯的不同而导致的语音信号与模板不匹配问题。但该算法只能识别特定说话人的特定文本,具有局限性,因此文中使用 HMM 来进行语音识别。

HMM 是一种用参数表示的、基于语音信号的时间序列结构建立的、用于描述其随机过程统计特性的概率模型,在处理离散时间序列的观察数据中应用广泛[8-9],能实现非特定说话人的语音识别。一般分为连续 HMM(CHMM)、半连续 HMM(SCHMM)以及离散 HMM(DHMM)[10],该模型表明,当前状态只与前一时刻所处的状态有关。

对比 DTW,HMM 的特点是:状态隐含、观察可测。将语音部分分割成极小的时间片段,那么该片段的特性近似稳定,总过程可视为从某一相对稳定特性到另一相对稳定特性的转移。

构建语音信号的 HMM,将它的语音分成上下两层,下层是不可测的、有限状态数的、马尔可夫链模拟的语音信号统计特性变化的隐含随机过程;上层引入概率统计模型,是与马尔可夫链的每一个状态相关联的观测序列的随机过程[11]。

语音信号是一种非平稳信号,一段完整的语音信号可以分为静音、语音、停顿、语音、静音五个部分[12]。其中语音部分,又可将每一个音节分成构筑其发音的最小单位—音素。多状态的 HMM 构成一个音素,多个音素的 HMM 串接构成一个字,将多个字的 HMM 串接起来,便可得到词汇的马尔可夫链。图 3 所示便是一个音素与观测序列的关系,其中 $O_1,O_2,\cdots,O_T$ 为观测得到的序列,若干个序列组成状态集 $S_1,S_2,\cdots$,而这个状态集便对应了语音的一个音素。

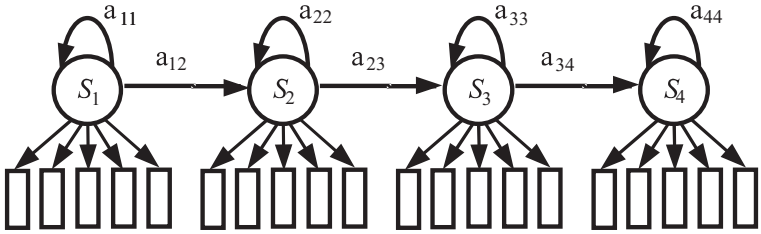

图 3 HMM 与语音参数关系

连续语音识别就是马尔可夫链和静音组合起来的 HMM,用概率密度函数计算语音特征参数对 HMM 模型的输出概率,通过搜索最佳状态序列,以最大后验概率为准则找到识别结果[13]。

2.2 HMM 的训练

一个 HMM 可由式 $\lambda=(\pi,\boldsymbol{A},\boldsymbol{B})$ 描述,其中 π 为初始状态概率,$\boldsymbol{A}$ 为状态转移概率矩阵,$\boldsymbol{B}$ 为观测概率矩阵。π 和 $\boldsymbol{A}$ 决定状态序列,$\boldsymbol{B}$ 决定观测序列。作为参数重估问题,HMM 需要解决三个问题[11,13-14]:

(1)输出概率计算问题。

该问题是语音信号的建模问题。已知观测序列 $O=\{O_1,O_2,\cdots,O_T\}$ 和隐马尔可夫模型 $\lambda=(\pi,\boldsymbol{A},\boldsymbol{B})$,将所求观察序列在 HMM 下出现的条件概率分成两部分,分别利用前向算法、后向算法将求得的条件概率进行乘积,进而得到整个观察序列的输出概率,以达到降低计算复杂度的目的。

定义 HMM 的前向概率为 $\alpha_t(i)=P\{O_1,O_2,\cdots,O_T;q_t=i\mid\lambda\}$,表示在给定 HMM 参数 λ 的前提下,观测序列为 $\{O_1,O_2,\cdots,O_T\}$ 在 t 时刻处于隐藏状态 i 的概率。前向概率 $\alpha_t(i)$ 的递推公式如下:

初始化:$\alpha_1(i)=\pi_i b_i(O_1),i\in[1,N]$。

迭代计算:$\alpha_{t+1}(j)=\left[\sum_{i=1}^{N}\alpha_t(i)a_{ij}\right]b_j(O_{t+1})$,$t\in[1,T-1]$,$j\in[1,N]$,其中 $\alpha_t(i)a_{ij}$ 表示在时刻 t 观察到 $O_1,O_2,\cdots,O_T$ 并在 t 时刻处于状态 q_i,在 $t+1$ 时刻到达状态 q_j 的联合概率。

终止计算:$P(O\mid\lambda)=\sum_{i=1}^{N}\alpha_T(i)$。

与前向概率相对应,定义后向概率为 $\beta_t(i)=P\{O_{t+1},O_{t+2},\cdots,O_T;q_t=i\mid\lambda\}$,表示在给定 HMM 参数 λ 的前提下,观测序列 $\{O_{t+1},O_{t+2},\cdots,O_T\}$ 在 t 时刻处于隐藏状态 i 的概率。

后向概率 $\beta_t(i)$ 的递推公式如下:

初始化:$\beta_T(i)=1$,$i\in[1,N]$。

迭代计算:$\beta_t(i)=\sum_{j=1}^{N}\beta_{t+1}(j)a_{ij}b_j(O_{t+1})$,$t=T-1,T-2,\cdots,1$;$i\in[1,N]$。

终止计算:$P(O\mid\lambda)=\sum_{i=1}^{N}\pi_i b_i(O_1)\beta_1(i)$。

经过分析,可得输出概率计算公式为:

$$P(O\mid\lambda)=\sum_{i=1}^{N}\alpha_t(i)\beta_t(i)=\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j),\ t\in[1,T-1]$$

反复迭代直到 HMM 参数不再发生明显变化为止。

(2)状态序列解码问题。

该问题是寻找最优匹配问题。Viterbi 算法是一种广泛用于通信领域的动态规划算法,即用动态规划求概率最大路径,它克服了全概率公式无法找到最优状态转移路径的问题。给定观察序列和 HMM,通过 Viterbi 识别算法确定一个最优的状态转移序列,并得到该路径所对应的输出概率。

设最优路径为 $I^*=(i_1^*,i_2^*,\cdots,i_T^*)$。

初始化:$\delta_1(i)=\pi_i b_i(O_1)$,$\psi_1(i)=0$,$i\in[1,N]$。

迭代计算:

$\delta_t(j)=\max_{1\leqslant i\leqslant N}[\delta_{t-1}(i)a_{ij}]b_j(O_t)$,$i\in[1,N]$,$j\in[1,N]$,$t\in[2,T]$

$\psi_t(j)=\arg\max_{1\leqslant i\leqslant N}[\delta_{t-1}(i)a_{ij}]$,$i\in[1,N]$,$j\in[1,N]$,$t\in[2,T]$

最后计算:$p^*=\max_{1\leqslant i\leqslant N}[\delta_T(i)]$,$i\in[1,N]$,$I_T^*=\arg\max_{1\leqslant i\leqslant N}[\delta_T(i)]$,$i\in[1,N]$。

路径回溯:$i_t^*=\psi_{t+1}(i_{t+1}^*)$,$t=T-1,T-2,\cdots,1$。

其中 $\delta_t(i)$ 为 t 时刻第 i 状态的累计输出概率，$\psi_t(j)$ 为 t 时刻第 i 状态的前续状态信号，为最优状态序列中 t 时刻所处的状态，为最终的输出概率。实际使用中，通常用对数形式的 Viterbi 算法，这样将避免进行大量的乘法计算，减少了计算量，同时还可以保证较高的动态范围，避免由于过多的连乘而导致溢出问题。在识别阶段，如果 HMM 模型为整词模型，就没有必要保存前续节点矩阵和状态转移路径，可以进一步减少计算量。

(3)模型参数的估计问题。

该问题是模型的修正问题，使 HMM 能够做到非特定人的语音识别。Baum－Welch 算法是极大似然(ML)准则的一个应用，利用该算法对初始化的 HMM 参数进行训练重估，即多个不同人对同一条命令重复多次录入，分别计算各自的特征参数序列，得到重估模型参数，使 $P(O \mid \lambda)$ 概率最大。

π 的重估公式：$\pi_i = \gamma_1(i) = \dfrac{\alpha_1(i)\beta_1(i)}{\sum_{i=1}^{N}\alpha_1(i)\beta_1(i)}$。

a_{ij} 的重估公式：$A_{ij} = \dfrac{\sum_{t=1}^{T-1}\xi_t(i,j)}{\sum_{t=1}^{T-1}\gamma_t(i)} = \dfrac{\sum_{t=1}^{T-1}\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)/P(O \mid \lambda)}{\sum_{t=1}^{T-1}\alpha_t(i)\beta_t(i)/P(O \mid \lambda)}$。

$b_j(O_k)$ 的重估公式：$B_j(O_k) = \dfrac{\sum_{t=1}^{T}\gamma_t(j)_{O_k=v_k}}{\sum_{t=1}^{T}\gamma_t(j)} = \dfrac{\sum_{t=1}^{T}\alpha_t(i)\beta_t(i)/P(O \mid \lambda)_{O_k=v_k}}{\sum_{t=1}^{T}\alpha_t(i)\beta_t(i)/P(O \mid \lambda)}$。

由于该系统采用从左至右、无跳转、单向结构的 HMM 模型，初始概率恒等于 $\pi_1 = 1$、$\pi_i = 0$，$i \in [2, N]$，因此不需进行重估。

3　实验测试

文中的数据库为自建库，采用普通麦克风录制。测试中发现，当使用 16 kHz 及以上时，语谱图在 4 000 Hz 以上的频段依然存在大量数据，而人类发声系统一般发出的语音处于 300 Hz ~ 3 400 Hz 之间。分析原因，是由语音信号采样率引起的，因此文中使用 8 kHz 的采样率。说话语音为普通话。

自建库包括两部分，一部分为训练语音：20 个说话人录制 10 条时长为 2 s 的正常语音；另一部分为测试语音：30 个说话人录制 10 条长为 2 ~ 4 s 的正常语音，再在狭小回廊环境录制 10 条长为 2 ~ 4 s 的语音作为混响语音。训练语音与测试语音文本相同。

实验开始对语音信号进行降噪处理，滤除高频噪声，然后进行端点检测等处理，截取语音信号中存在语音的部分。

图4展示了三组经过预处理后的图，分别为同一

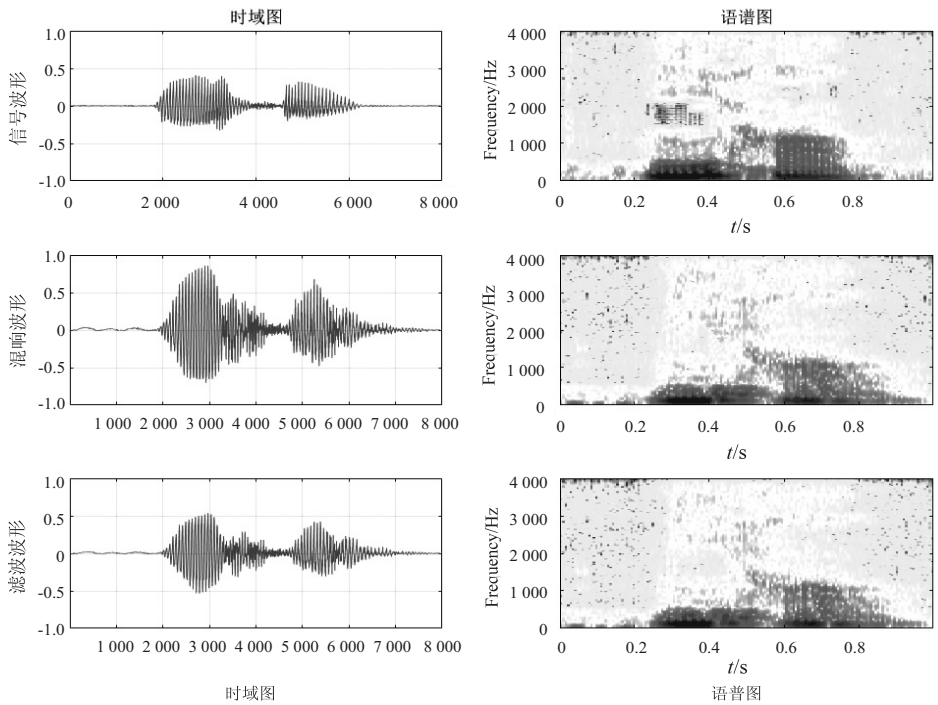

图 4　同态滤波去混响方法的时域图与语谱图

词汇的纯净语音、混响语音以及去混响后的语音;每组图片由两张图构成,分别为语音的时域图与语谱图。语谱图是语音的频谱图,其横坐标是语音持续的时间,纵坐标为语音的频率,坐标点值为语音的能量。通过观察可以发现,经过复倒谱滤波后,时域信号的振幅有所降低,对比语谱图可以发现,0.2 ~0.4 s 的语音混杂程度有所降低。

系统分训练和识别两个阶段。在语音识别中 Mel 频率倒谱系数(Mel frequency cepstrum coefficient, MFCC)将线性频标转化为 Mel 频标,能屏蔽大部分高频噪声的干扰,有利于识别信息。因此在训练阶段提取训练语音信号的 MFCC 参数,对 10 条语音分别建立 HMM。

在识别阶段,将带混响的语音信号输入采集器,使用卷积同态滤波去除混响,提取语音信号的 MFCC 参数,建立待识别语音的 HMM,将测试语音的 HMM 模型与训练库中的各个模型应用 Viterbi 算法进行搜索,找到输出概率最大的训练模型。

表 1 展示了基于 HMM 的语音识别结果,测试了该模型对正常语音、混响语音的识别情况。其中混响语音测试又分为 a,b 两个步骤,步骤 a 单纯测试该模型对混响语音的识别情况,步骤 b 则是将混响语音通过卷积同态滤波器后再进行识别。

表 1 基于 HMM 的语音识别结果

指标	正常语音	混响语音 a	混响语音 b
测试次数	100	100	100
识别率/%	87	76	81

可以看出,该模型在识别正常语音时具有较高的可靠性,实现了短词汇非特定人的语音识别,对比使用卷积同态滤波器前后的结果,能发现该系统对于混响语音的处理也具有一定的效果,识别准确率有所提升。

4 结束语

文中基本实现了混响语音的语音识别,实验结果表明,单纯使用复倒谱滤波的方法去混响,效果并不是特别明显。虽然 HMM 在语音识别方面用途广泛,但其浅层学习结构在海量数据下性能会受到限制[15-16],因此单纯使用 HMM 进行语音识别也遇到了很大阻力。随着机器学习的兴起,神经网络声学建模与传统手段相结合,将进一步推动语音识别技术的发展。

参考文献:

[1] 王海坤,潘 嘉,刘 聪. 语音识别技术的研究进展与展望[J]. 电信科学,2018,34(2):1-11.

[2] ZHANG Jing,CHEN Xiaomei. A research of improved algorithm for GMM voiceprint recognition model[C]//Control and decision conference. Yinchuan,China:IEEE,2016:5560-5564.

[3] HUANG Jing,VISWESWARIAH K. Improved decision trees for multi-stream HMM-based audio-visual continuous speech recognition[C]//IEEE workshop on automatic speech recognition & understanding. Merano, Italy: IEEE, 2009:228-231.

[4] HU Hao,XU Mingxing,WU Wei. GMM supervector based SVM with spectral features for speech emotion recognition[C]//International conference on acoustics,speech and signal processing. Honolulu,HI,USA:IEEE,2007:413-436.

[5] 孔 荣. 混响环境下孤立词识别的研究[D]. 苏州:苏州大学,2013.

[6] 苏先礼. 语音去混响研究[D]. 成都:四川大学,2006.

[7] 姚天任,孙 洪. 现代数字信号处理[M]. 武汉:华中科技大学出版社,1999:170-171.

[8] NAKAMURA S,TAKIGUCHI T,SHIKANO K. Noise and room acoustics distorted speech recognition by HMM composition[C]//International conference on acoustics, speech, and signal processing. Atlanta, GA, USA: IEEE, 1996: 69-72.

[9] BUYUK O,ARSLAN L M. HMM-based text-dependent speaker recognition with handset-channel recognition[C]//Proceeding of the 18th IEEE signal processing and communications applications conference. Diyarbakir, Turkey: IEEE, 2010:383-386.

[10] RABINER L,JUANG B H. An introduction to hidden Markov models[J]. IEEE ASSP Magazine,1986,3(1):4-16.

[11] 于晓明,柏 松. 基于前向-后向 HMM 的连续语音识别系统的研究[J]. 计算机工程与设计,2009,30(18):4339-4341.

[12] 叶 硕,彭春堂,杜珍珍,等. 基于 DTW 的孤立词语音识别系统设计[J]. 长江大学学报:自科版,2018,15(17):33-37.

[13] 郭圣权,连晓峰. MATLAB 环境下的基于 HMM 模型的语音识别系统[J]. 计算机测量与控制,2004,12(5):470-472.

[14] 李 航. 统计学习方法[M]. 北京:清华大学出版社,2012:184-186.

[15] 张晴晴,刘 勇,潘接林,等. 基于卷积神经网络的连续语音识别[J]. 工程科学学报,2015,37(9):1212-1217.

[16] 林巧民 齐柱柱. 基于 HMM 和 ANN 混合模型的语音情感识别研究[J]. 计算机技术与发展,2018, 28(10):74-78.