研究与开发

基于 CNN 的连续语音说话人声纹识别

吴震东,潘树诚,章坚武
(杭州电子科技大学,浙江 杭州 310018)

摘　要:近年来,随着社会生活水平的不断提高,人们对机器智能人声识别的要求越来越高。高斯混合—隐马尔可夫模型(Gaussian of mixture-hidden Markov model, GMM-HMM)是说话人识别研究领域中最重要的模型。由于该模型对大语音数据的建模能力不是很好,对噪声的顽健性也比较差,模型的发展遇到了瓶颈。为了解决该问题, 研究者开始关注深度学习技术。引入了 CNN 深度学习模型研究连续语音说话人识别问题, 并提出了 CNN 连续说话人识别(continuous speaker recognition of convolutional neural network, CSR-CNN)算法。模型提取固定长度、符合语序的语音片段,形成时间线上的有序语谱图,通过 CNN 提取特征序列,经过奖惩函数对特征序列组合进行连续测量。实验结果表明,CSR-CNN 算法在连续—片段说话人识别领域取得了比 GMM-HMM 更好的识别效果。

关键词:连续语音;语谱图;GMM-HMM;深度学习
中图分类号:TP393　　**文献标识码**:A
doi: 10.11959/j.issn.1000-0801.2017046

Continuous speech speaker recognition based on CNN

WU Zhendong, PAN Shucheng, ZHANG Jianwu
Hangzhou Dianzi University, Hangzhou 310018, China

Abstract: In the last few years, with the constant improvement of the social life level, the requirement for speech recognition is getting higher and higher. GMM-HMM (Gaussian mixture-hidden Markov model) have been the main method for speaker recognition. Because of the bad modeling capability of big data and the bad performance of robustness, the development of this model meets the bottleneck.In order to solve this question, researchers began to focus on deep learning technologies. CNN deep learning model for continuous speech speaker recognition was introduced and CSR-CNN model was put forward. The model extracts fixed-length and right-order phonetic fraction to form an ordered sound spectrograph. Then input the voiceprint extract from CNN model to a reward-penalty function to continuous measurement. Experimental results show that CSR-CNN model has very good recognition effectin continuous speech speaker recognition field.

Key words: continuous speech, sound spectrograph, GMM-HMM, deep learning

收稿日期:2017-01-22;**修回日期**:2017-02-13
基金项目:浙江省自然科学基金资助项目(No.LY16F020016);国家重点研发计划经费资助项目(No.2016YFB0800201);浙江省重点科技创新团队项目(No.2013TD03)
Foundation Items: Zhejiang Natural Science Foundation of China (No.LY16F020016), National Key Research and Development Program of China (No.2016YFB0800201), Zhejiang Province Science and Technology Innovation Program (No.2013TD03)

1 引言

随着移动互联网、物联网等技术的高速发展，实现人与电子产品之间的自由交互越来越受到人们的重视。声纹识别技术在实现这一目标中扮演着非常重要的角色。语音识别技术正在走向实用。苹果公司于 2011 年收购了 Siri 公司，并在 iPhone 4 上应用了语音识别功能，但当时识别体验不理想。2013-2015 年，苹果公司相继收购了拥有识别整个短语的语音识别技术的 Novauris 公司和英国语音技术初创公司 VocalIQ。与此同时，谷歌在 2011 年收购了语音通信公司 Say Now 和语音合成公司 Phonetic Arts，2015 年入资中国以导航为主的问问公司，并推出带有语音识别技术的智能手表。Amazon 在 2011-2013 年，相继收购语音识别领域的 Yap 语音识别公司、Evi 语音技术公司和 Ivona Software 语音技术公司。Facebook 于 2013 年后，相继收购了 Mobile Technologics 和 Wit.ai 语音识别公司，实现了用户可以通过语音来控制应用程序、穿戴设备和控制机器人等功能。微软的 Cortanna 和微软小冰在记录用户使用习惯和智能对话等功能，使人们生活更加智能化。国内百度语音、科大讯飞等科技公司在语音识别领域也在进行大量的应用基础及应用性研究。

与语音识别技术发展阶段相似，声纹识别技术也在走向实用。现有技术在长文本、低噪声声纹识别时，已达到较高识别率。但是在片段语音环境下，常用的线性预测频率倒谱系数（linear prediction cepstrum coefficient）和 Mel 频率倒谱系数（mel frequency cepstrum coefficient）等声学特征，识别率明显下降。在模式识别方面，静态说话人模型包括：高斯混合模型[1]（Gaussian mixture model）、高混合通用背景模型（Gaussian mixture model-universal background model）和支持向量机[2]（support vector machine，SVM）。这些静态模型在用特征描述目标说话人的时候有很好的效果。一般来说，传统的重要模型包括 Douglas Reynolds 提出的 UBM-MAP-GMM 模型、Patrick Kenny 提出的 Joint Factor Analysis[3-6] 和 NajimDehak 提出的 i-vector[7,8]。在一定程度上，可以把现有短语音声纹识别模型视为不充分的声学特征，该模型尚不能很完美地描述说话人声纹特点。

2006 年，深层结构模型在识别领域开始发光发热，可以说是语音、图像识别领域突破性发展的重要一年。2006年之前，研究者们通过各种方式来搭建深层的架构来实现语音和图像的识别，但是都得不到好的结果。因为训练一个有深度的前馈神经网络，用浅层的学习方法往往得不到理想的效果。当层数越深时，深层网络的梯度就会变得很不稳定，这使得深层次的梯度对之前层的关联度几乎丧失，从而导致模型训练效果急剧恶化。就在这一年，Hinton 等人[9]首次提出了非监督贪心逐层训练算法生成的模型——深度置信网络[10]（deep belief network，DBN），每一层都用训练数据来初始化深度神经网络，优化了深度网络结构，一定程度上解决 BP 算法[11]带来的局部最优解问题。卷积神经网络[12]（convolutional neural network，CNN）被提出并大量应用于图像特征挖掘。

2009 年以来，深度学习方法逐渐被引入声纹识别领域，用以对语音的深层特征加以挖掘，构建更充分的声纹识别模型，如递归神经网络[13]（recurrent neural network，RNN）以及 RNN 的各种变型（LSTM 模型）[14]。RNN 模型的网络结构可以表达前后信息相关的时序效果，所以在语音识别方面有很大的优势。但目前而言，RNN 及 LSTM 模型均未达到超越 GMM 模型的声纹识别能力。

本文将目前更为成熟的深度神经网络 CNN 模型引入声纹识别，构建连续—片段语音，基于有序语谱图的 CNN 声纹识别系统取得了比 GMM 模型更好的声纹识别能力。

2 模型建立

声纹识别是生物特征识别[15]中的一种，也被称作说话人识别，可分为说话人辨别和说话人确认两类。前者是在很多说话人的情况下判断是其中哪个说话人所说的，是“多对一”的过程；后者是判断为某个说话人说所的。根据不同的任务需求和应用场景，选取不同的声纹识别技术，如在支付交易或者远程登录的时候需要确认技术，而在缩小目标范围的时候则需要辨别技术。

传统的声纹识别模型一般都是在隐马尔可夫模型（hidden Markov model，HMM）[16]的基础上建立的，而 HMM 是一种基于统计的特征识别方法。换句话说，是根据声学模型和语言模型，通过最大后验概率来识别。现阶段基于深度学习的语音识别，模型通过对大量数据的训练，自动地学习数据中的特征。表现一个人声学层面的特征有好多种，包括解剖学声学特征（倒频谱、共振峰等）、语法特征、韵律特征、通道信息、语种、语调和习语等[17]。传统的声纹识别方法需要研究者对这些声学特征进行人工分类。而在深度学习中，研究者不用知道声学特征的相关信息，机器会自动地学习数据中的声纹特征信息。显著提高了研究者的研究

效率,并且经过对大量数据的学习,机器能够学到更加完备的特征,效果比人工分类更好。

2.1 声纹识别系统

一个完整的说话人识别系统由声学特征提取、统计模型和分值计算组成,如图 1 所示。系统训练的过程是从原始的波形信号中提取语音的声学特征,如词、音节、音素及声韵母等,并经过训练得到一个声学模型,这个模型作为识别语音声学特征基元的模板,模型结合研究者通过对人类声学特征研究得到的语言模型,经过解码器的处理输出相应的识别结果。

图 1 说话人识别系统结构

现有的语音识别模型运用效果最好的为高斯混合模型(GMM),其基本过程为提取语音 MFCC 特征序列,运用统计模型对输出序列进行概率评分,依据评分结果进行识别判断。具体过程如下。

2.2 特征提取

MFCC 的整个提取过程如图 2 所示。其中,帧周期持续 10~25 ms,在这期间,声音被认为是静止的。帧周期取 20 ms 的时候,移码一般取 10 ms。

预修正的部分是高通滤波器。数学表达式如下:

$$H(z)=1-az^{-1} \tag{1}$$

其中,a 是预修正系数,一般取 0.95~0.97。频率弯折能够让声音有更好的表现特性,比如在声频压缩方面。

汉明窗口能够平滑帧信号的边缘:

$$\omega(n)=\left[0.54-0.46\cos\left(\frac{2\pi n}{M-1}\right)\right]R_M(n) \tag{2}$$

在音频处理中,Mel 频率倒谱系数[18]表示声音短期的功率谱。将功率谱取对数带入 Mel 频率倒谱系数中:

$$F_{\text{mel}}(f)=2\,595\lg\left(1+\frac{f}{700}\right) \tag{3}$$

Mel 频率倒谱系数从音频片段的倒谱表示中派生而来,Mel 倒谱系数和倒谱系数的区别在于,Mel 频率倒谱的频带划分在 Mel 刻度上是等距的,这比一般的对数倒谱更加符合人类的听觉系统。音频的响应函数如下:

$$H_m(k)=\begin{cases}0, & k<f(m-1)\\ \dfrac{k-f(m-1)}{f(m)-f(m-1)}, & f(m-1)<k<f(m)\\ \dfrac{f(m+1)-k}{f(m+1)-f(m)}, & f(m)<k<f(m+1)\\ 0, & k>f(m+1)\end{cases} \tag{4}$$

其中,M 是三角滤波器的总数,m 的取值范围是 $0\leqslant m<N$。$f(m)$是 Mel 带通滤波器组的第 m 个滤波器,其数学表达式如下:

$$f(m)=\left(\frac{N}{f_s}\right)F_{\text{mel}}^{-1}\left(F_{\text{mel}}(f_l)+m\frac{F_{\text{mel}}(f_h)-F_{\text{mel}}(f_l)}{M+1}\right) \tag{5}$$

其中,N 是 FFT 的长度,f_h 和 f_l 分别是滤波器的最大频率和最小频率。F_{mel}^{-1}是 F_{mel} 的反函数,反函数的作用是把 Mel 频率转换到 Hz 频率。数学表达式如下:

$$F_{\text{mel}}^{-1}(b)=700(\mathrm{e}^{b/1125}-1) \tag{6}$$

2.3 统计模型

GMM 和 UBM 的训练过程如图 3 所示。

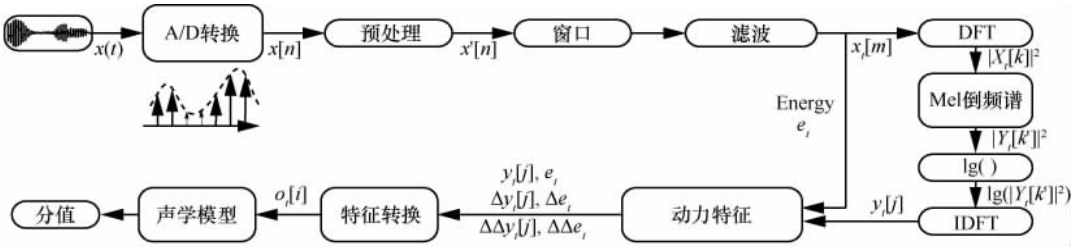

图 2 MFCC 提取过程

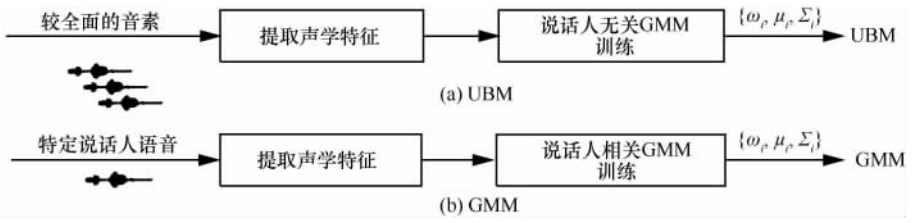

图 3 GMM 和 UBM 的训练过程

给出一连串的特征向量 $X=\{x_1,\cdots, x_t,\cdots, x_m\}$ 和说话人模型的依赖参数 $\lambda=\{\omega_i, \mu_i, \Sigma_i\}$，这些参数各自的迭代公式如下：

$$\omega_i=\frac{1}{m}\sum_{t=1}^{m}Pr(i\mid x_t,\lambda) \tag{7}$$

$$\mu_i=\frac{\sum_{t=1}^{m}Pr(i\mid x_t,\lambda)x_t}{\sum_{t=1}^{m}Pr(i\mid x_t,\lambda)} \tag{8}$$

$$\Sigma_i=\frac{\sum_{t=1}^{m}Pr(i\mid x_t,\lambda)(x_t-\mu_i)(x_t-\mu_i)'}{\sum_{t=1}^{m}Pr(i\mid x_t,\lambda)} \tag{9}$$

其中，第 i 次的后验概率为：

$$Pr(i\mid x_t,\lambda)=\frac{p(x_t\mid i;\mu,\Sigma)p(i;\omega)}{\sum_{l=1}^{k}p(x_t\mid l;\mu,\Sigma)p(l;\omega)} \tag{10}$$

$$Pr(i\mid x_t,\lambda)=\frac{\omega_i b_i(x_t)}{\sum_{l=1}^{k}\omega_l b_l(x_t)} \tag{11}$$

经分值估算，达到某概率阈值之上，判定输入语音为说话者语音。概念估算计算式如下：

$$\lg L=\lg(p(X\mid\lambda))-\lg(p(X\mid\lambda_u)) \tag{12}$$

3 CNN 连续说话人识别算法 CSR-CNN

语音方面的深度学习模型一般都是 RNN 模型及它的变形 LSTM 模型。因为 RNN 模型引入了定向循环，能够处理输入之间前后关联的问题。这种识别技术一般应用在机器翻译、图像描述生成等领域。在说话人识别领域，RNN 模型的识别效果并不是很理想。卷积神经网络在很多识别问题上已经有了很好的识别效果，比如手写字体的识别、人脸识别、交通标志分类、行人检测、图像标注和行为检测[19-24]。因为 CNN 模型在图像领域的优越表现，本文想通过图像的方法来达到连续—片段语音说话人识别的目的。本文结合 CNN 模型和声纹的频谱图特征，在说话人识别领域提出连续—片段语音说话人识别（continuous speech recognition of convolutional neural network，CSR-CNN）算法。

3.1 算法结构

CSR-CNN 由 CSR 和 CNN 两个模型构成。CSR 是连续—片段说话人识别模型，CNN 为特征提取模型，其结构如图 4 所示。

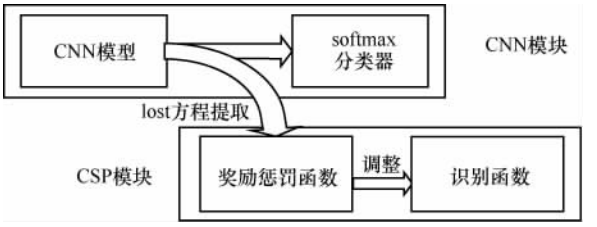

图 4 CSR-CNN 模型结构

系统先将时域上的说话人语音信息转换为语谱图（语音在时域上的表示是没有任何声学特征的），然后用训练数据训练一个 CNN 模型，并用测试数据检测模型正确率。训练好这个模型，将待检测人的语谱图分片传入该模型，并提取它输出特征向量。通过特征向量和标签特征向量得到一个 lost 方程，如果 lost 方程计算评分大于给定的一个阈值，那么给出一个惩罚函数值，反之给出一个奖励函数值。这两个函数最终决定着说话人识别函数的结果。当说话人识别函数达到某个阈值时，就判定身份验证成功，反之验证失败。

3.2 CNN 模型

卷积神经网络可分为输入层、卷积层、池化层和输出层，如图 5 所示，其中卷积层和池化层是卷积神经网络特有的。多个卷积核滤波器对原始输入图像卷积来提取多个抽象特征（线条、边缘等），池化层对卷积层进行池化处理，使提取的特征更加紧凑并减少神经元个数。使用多个卷积层和池化层的组合可以提取更加具像的特征（眼睛、鼻子等）。最后，通过 softmax 分类器和全连接层输出结果。卷积神经网络有 3 个主要的特征：局部感知域、权值共享和池化层。

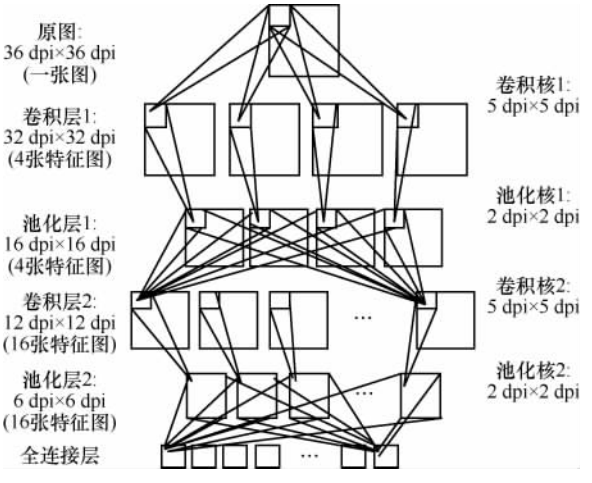

图 5 CNN 模型结构

3.2.1 局部感知域

卷积神经网络中，本文把很小的邻近区域作为输入，如图 6 所示，5 dpi×5 dpi 的卷积核窗口和输入图像做卷积，得到下一层图像的一个像素点。其中被卷积部分就是局部感知域，每一个局部感知域在下一隐层中都有一个神经元与之对应。

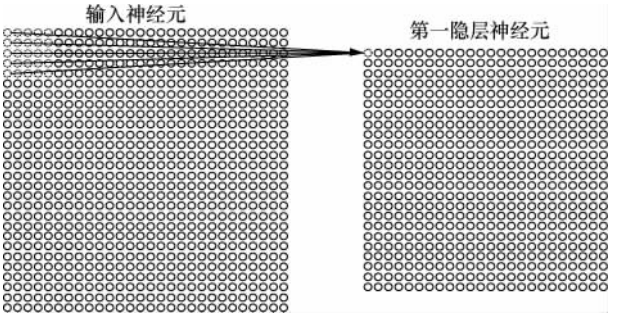

图 6　局部感知区域示意

3.2.2 权值共享

如图 7 所示，每个卷积核都带有一组自己的权值和 bais 值并会自左向右、自上向下依次和输入图像做卷积。这就说明该卷积核特征映射图的每一个神经元都在检测同一特征，只是这些特征位于图片的不同地方，这使得识别目标在不断移动时也能被识别。

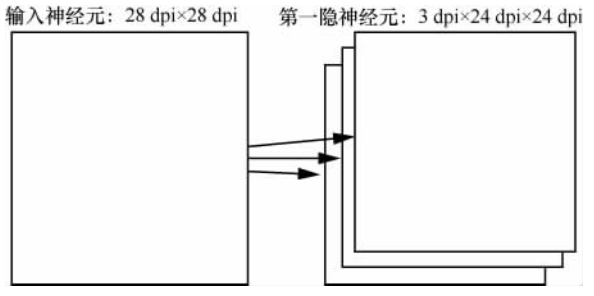

图 7　卷基层特征图提取示意

图 7 举例了 3 个特征映射图，每一张特征映射图都是通过一个权值共享的卷积核和输入图像卷积所得到。

每个卷积核只能提取一种特征，训练中需要初始化多个卷积核。就计算量而言，以取 20 个特征为例，其需要 520 个参数，和全连接神经网络 23 550 个参数相比，大大降低了计算量。

系统搭建了一个有 L 个卷积隐层的 CNN。其中 $X=(x_0, x_1, \cdots, x_N)$是输入向量，$H=(h_0, h_1,\cdots,h_L)$是中间层的输出向量，$Y=(y_0, y_1,\cdots,y_M)$是模型的实际输出，$D=(d_0, d_1,\cdots, d_M)$是目标输出，$V_{ij}$ 是前一层输出单元 i 到隐层单元 j 的权重，W_{jk} 是隐层单元 j 到前一层输出单元 k 的权重。另外，θ_k 和 ϕ_j 分别前一层输出单元和隐层单元的阈值。

输入数据和特征提取窗口做卷积，并通过一个激活函数(ReLU)得到下一层的特征图。卷积表达式如下：

$$h_j = f\left(\sum_{i=0}^{N-1} V_{ij} x_i + \phi_j\right) \tag{13}$$

得到的特征图作为下一个池化层的输入，进行降维处理。降维处理对系统有 3 个作用：让特征更加紧凑，特出显著特征；减少系统的训练参数，n 尺寸的池化层可以减少 n^2 倍的参数；增加系统的顽健性。

池化层的数学表达式如下：

$$y_k = f\left(\sum_{j=0}^{L-1} W_u h_j + \theta_k\right) \tag{14}$$

其中，$f(*)$ 是激活函数，系统中使用的激活函数是 ReLU，其数学表达式如下：

$$f(x) = \begin{cases} x, & x > 0 \\ \alpha x, & x \leqslant 0 \end{cases} \tag{15}$$

经过多个卷积层和池化层后，提取到的特征经过最后一个全连接层得到一组特征向量，并通过分类器实现最后的分类。

3.3 CSR 模型

引入 CSR 模型的目的是实现在连续语音的情况下，能够不间断地确定目标说话人的身份。CSR 模型结构如图8 所示。

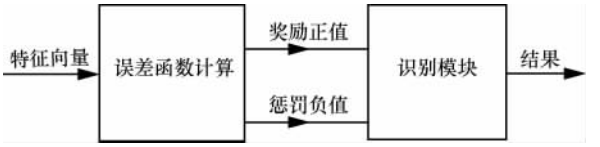

图 8　CSR 模型结构

CSR 模型中，设置一个奖惩函数，数学表达式如下：

$$f(lost^n) = \begin{cases} 1, & lost^n \leqslant b \\ -1, & b \leqslant lost^n < 1 \end{cases} \tag{16}$$

其中，$lost^n$ 是第 n 个待检测语音数据在通过 CNN 模型训练后得到的归一化特征向量和目标特征向量的误差函数，b 是根据模型识别率给定的误差阈值。

由式(16)可以看出，当 $lost^n$ 的值低于给定阈值的时候，给予说话人识别函数一个奖励函数，反之给予一个惩罚函数。

系统识别函数的数学表达式如下：

$$\varphi_n = \varphi_{n-1} + f(lost^n) \tag{17}$$

其中，φ_n 是判断第 n 时刻的系统状态，$f(lost^n)$是第 n时刻的奖惩函数。

设定说话人识别函数 φ_n 取值区间为[c,d],即当说话人识别函数达到最大值或者最小值时，它的值就不会改变,并且给出一个识别阈值 w。

当 $\varphi_n>w$ 时,则表示目标说话人身份鉴定成功;当 $p<w$ 时,则表示目标说话人身份鉴定失败。当语音数据源源不断输入,该模型可以不间断地确认说话人的身份。从 φ_n 值的设定可以看出,当识别率 P 值处在峰值时,即使因为周围语音环境发生短暂性的变化以及可能的误判,也可以持续地确认说话人身份。

CSR 模型对单独的 CNN 模型的识别率有很高的提升。CSR-CNN 模型的识别率数学表达式如下：

$$P \approx 1-\left(P_n \times \alpha(a)+P_m \times \beta(a)\right) \quad (18)$$

其中,P_n 是识别函数在识别阈值上侧的最小值出现的概率,P_m 是识别函数在识别阈值下侧的最大值出现的概率,$\alpha(a)$是错误接受率,$\beta(a)$是错误拒绝率。在一般的模型中,P_n、P_m、$\alpha(a)$和 $\beta(a)$的取值一般为百分之几,所以识别率 P 接近于 1。所以 CSR-CNN 模型在连续—片段说话人识别领域有很好的识别效果。

4 实验及结果分析

本文实验中所使用的数据库包含目标说话人在实验室环境下随机朗读 200 个短语(每个短语持续 1~2 s)以及目标说话人 40 s 的长语句和攻击者 15 s 的长语句，咬字清晰,使用手持麦克风录制语音。

本文首先要对原始的语音信号进行预处理。将时域上的语音信号进行频域上的转换,生成 200 个频谱图,并调整为 258 dpi×258 dpi 的大小,作为模型的输入。将预处理后的频谱图作为输入传入 CSR-CNN 模型中，实验中设置的迭代步数为 5 000 步，在 5 000 步时,lost 方程趋于平稳,模型趋于最优,最终 lost 的值为 0.03。当步长在 0.02时,本实验的 CNN 模型的识别率比较高,最终识别率为 96%。训练完 CNN 模型,将连续说话人识别模型和 CNN 模型集合进行识别。将目标说话人和攻击者的长语句,进行 1~2 s 的切片,分别得到 30 个和 15 个短语块,并进行频域的转换。将这 35 个短语块按图 9 和图 10 的序列,分别组成语音序列 1 和语音序列 2。

提取每个短语块的输出特征向量，结合目标特征向量得到每个短语块归一化处理的 $lost^n$ 值。将 $lost^n$ 的值输入 CSR 模型,CSR 模型通过对 $lost^n$ 的判断来决定输出一个奖励函数还是惩罚函数,并输入最后的系统判决函数。

运行自己搭建的 CNN 实验模型，对准备数据库的说话人识别率为 92%,达到了一个较高的识别率水平。再结合 CSR 模型,对准备的长语句数据进行识别,说话人函数的输出函数如图 11 所示。

图 11 中，当说话人函数的数值在虚线标识区域的上方时,即函数值大于 2.5 的时候,系统就认为目标说话人被识别,反之则为识别失败。通过对表 2 和实验结果图 11 的对比可以发现，表 1 和表 2 中都有 15 个攻击者说话人语音片段，而实验结果的图 11 中语音序列 1 和语音序列 2 分别有 17 次和 16 次的函数下降过程。这说明实验中语音序列 1 和语音序列 2 分别有 17 次和 16 次的识别结果为非目标说话人,即其中分别有 2 个说话人语音片段和 1 个说话人语音片段被误判为攻击者语音。经过数据比较,本文发现在语音序列 1 中,第 38 和 39 个目标说话人语音片段经过 CNN 模型被误判为攻击者语音;在语音序列 2 中,

真	真	真	真	真	真	真	真	真	假
假	真	真	真	真	真	假	真	真	真
真	真	真	真	假	假	假	假	假	假
假	假	真	真	真	真	真	真	真	真
假	真	假	假	假					

图 9 检测语音 1 的语音片段序列

真	真	假	假	真	真	真	真	真	真
真	真	真	真	真	真	真	真	假	真
假	真	真	真	真	真	假	真	假	真
假	真	假	假	假	假	真	真	真	假
假	假	假	真	真					

图 10 检测语音 2 的语音片段序列

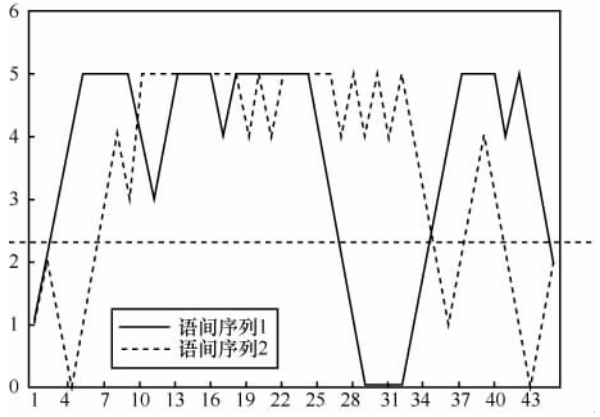

图 11　语音序列 1 和语音序列 2 的系统输出函数

第 9 个目标说话人语音片段经过 CNN 模型被误判为攻击者语音。但是将 CNN 结果输入 CSR 模型后，这个误判没有影响系统的整体的结果。该系统对 CNN 模型的误判率有一定的容错率，这提高了单 CNN 模型的识别率。

5　结束语

本文主要介绍了声纹识别的发展进程和目前应用比较广泛的几个深度学习模型，并阐述了这几个模型在语音识别领域中的应用和发展现状。最后通过结合语谱图、CNN 模型，在连续声纹识别中提出了 CSR-CNN 算法。

语音作为人机交互的一个关键接口，在人工智能方面有非常广泛的实际应用前景。这几年的研究表明，深度学习技术在声纹识别领域能够明显提高声纹识别系统的准确率。

虽然深度学习技术在语音领域取得了很大的成果，但是为了能够实现更加高效的人际关系，还有很多技术难点要克服。比如：很深层训练网络的梯度精确度问题、在实际应用中的噪声顽健性问题等。其中，噪声顽健性问题是现在语音识别中非常热门的话题。现阶段实际应用中，带噪声的语音识别率一般都不是很高。未来对于语音识别系统的研究方向应该更加倾向于仿人脑听觉系统，随着生物解剖学的发展，使模型不断接近人脑的语音识别特性，将在这一领域持续研究。

参考文献：

[1] SU D, WU X, XU L. GMM-HMM acoustic model training by a two level procedure with Gaussian components determined by automatic model selection[C]// 2010 IEEE International Conference on Acoustics Speech and Signal Processing, March 14–19, 2010, Dallas, TX, USA. New Jersey: IEEE Press, 2010: 4890-4893.

[2] JOACHIMS T. Making large-scale SVM learning practical [J]. Technical Reports, 1998, 8(3): 499-526.

[3] REYNOLDS D A, QUATIERI T F, DUNN R B. Speaker verification using adapted gaussian mixture models [J]. Digital Signal Processing, 2000, 10(1-3): 19-41.

[4] HEBERT M. Text-dependent speaker recognition[M]. Heidelberg: Springer, 2008: 743-762.

[5] VOGT R J, LUSTI C J, SRIDHARAN S. Factor analysis modeling for speaker verification with short utterances [J]. Journal of Substance Abuse Treatment, 2008, 10(1): 11-16.

[6] VOGT R, BAKER B, SRIDHARAN S. Factor analysis subspace estimation for speaker verification with short utterances [C]// INTERSPEECH 2008, Conference of the International Speech Communication Association, Sept 6–10, 2008, Brisbane, Australia. [S.l.: s.n.], 2008: 853-856.

[7] KANAGASUNDARAM A, VOGT R, DEAN D, et al. i-Vector based speaker recognition on short utterances[C] // INTERSPEECH 2011(DBLP), August 27–31. 2011, Florence, Italy. [S.l.: s.n.], 2011.

[8] LARCHER A, BOUSQUET P, KONG A L, et al. i-Vectors in the context of phonetically-constrained short utterances for speaker verification[C] // ICASSP, March 25–30, 2012, Kyoto, Japan. New Jersey: IEEE Press, 2012: 4773-4776.

[9] HINTON G E, SALAKHUTDINOV R R. Reducing the dimensionality of data with neural networks[J]. Science, 2006, 313(5786): 504-507.

[10] ZOU M, CONZEN S D. A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data[J]. Bioinformatics, 2005, 21(1):71-79.

[11] RUMELHART D E, MCCLELLAND J L. Parallel distributed processing[M] // Cambridge: The MIT Press, 1986: 45-76.

[12] ZORRIA SSATINE F, TANNOCK J D T. A review of neural networks for statistical process control [J]. Journal of Intelligent Manufacturing, 1998, 9(3): 209-224.

[13] CHEN S H, HWANG S H, WANG Y R. An RNN-based prosodic information synthesizer for Mandarin text-to-speech[J]. IEEE Transactions on Speech & Audio Processing, 1998, 6(3): 226-239.

[14] TAN T, QIAN Y, YU D, et al. Speaker-aware training of LSTM-RNNS for acoustic modeling [C]// 2016 IEEE International Conference on Acoustics, Speech and Signal Processing, March 20–25, 2016, Shanghai, China. New Jersey: IEEE Press, 2016: 5280-5284.

[15] GALES M J F. Maximum likelihood linear transformations for HMM-based speech recognition [J]. Computer Speech & Language, 1998, 12(2): 75-98.

[16] RAMASWAMY G N, GOPALAKRISHAN P S. Compression of acoustic features for speech recognition in network environments [C]//1999 IEEE International Conference on Acoustics, Speech and Signal Processing, May 15, 1998, Seattle, WA, USA. New Jersey: IEEE Press, 1998: 977-980.

[17] PAN J, LIU C, WANG Z, et al. Investigation of deep neural networks (DNN) for large vocabulary continuous speech recognition: why DNN surpasses GMMS in acoustic modeling [C]//2012 International Symposium on Chinese Spoken Language Processing, Dec 5-8, 2012, Kowloon Tong, China. New Jersey: IEEE Press, 2012: 301-305.

[18] HUANG Z, TANG J, XUE S, et al. Speaker adaptation of RNN-BLSTM for speech recognition based on speaker code[C] // IEEE International Conference on Acoustics, Speech and Signal Processing, March 20-25, 2016, Shanghai, China. New Jersey: IEEE Press, 2016: 5305-5309.

[19] SAATCI E, TAVASANOGLU V. Multiscale handwritten character recognition using CNN image filters [C] // 2002 International Joint Conference on Neural Networks, May 12-17, 2002, Honolulu, HI, USA. New Jersey: IEEE Press, 2002: 2044-2048.

[20] LIU K, ZHANG M, PAN Z. Facial expression recognition with CNN ensemble [C] // International Conference on Cyberworlds, Sept 28-30, 2016, Chongqing, China. New Jersey: IEEE Press, 2016: 163-166.

[21] JURISIC F, FILKOVIC I, KALAFATIC Z. Multiple-dataset traffic sign classification with OneCNN[C] // Iapr Asian Conference on Pattern Recognition, Nov 3-6, 2015, Kuala Lumpur, Malaysia. New Jersey: IEEE Press, 2015: 614-618.

[22] ZHANG L, LIN L, LIANG X, et al. Is faster R-CNN doing well for pedestrian detection? [M]. Heidelberg: Springer-Verlag: 443-457.

[23] ZHENG Y, LI Z, ZHANG C. A hybrid architecture based on CNN for image semantic annotation [M] // SHI Z Z, VADERA S, LI G. Intelligent Information Processing VIII, Heidelberg: Springer, 2016: 81-90.

[24] PARMAKSIZOGLU S, ALCI M. A novel cloning template designing method by using an artificial bee colony algorithm for edge detection of CNN based imaging sensors[J]. Sensors, 2011, 11(5): 5337-5359.

[作者简介]

吴震东(1976-),男,杭州电子科技大学网络空间安全学院讲师,主要研究方向为生物特征识别、生物密钥、网络安全、自然语言处理、人工智能等。

潘树诚(1991-),男,杭州电子科技大学通信工程学院硕士生,主要研究方向为基于深度学习的声纹、人脸识别研究等。

章坚武(1961-),男,杭州电子科技大学通信工程学院教授、博士生导师,主要研究方向为移动通信系统、多媒体通信技术、网络安全等。