计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210400041

基于 LSTM 神经网络的声纹识别

刘晓璇　季　怡　刘纯平
苏州大学计算机科学与技术学院　江苏 苏州 215006
(liuxiaoxuan_n@163.com)

摘　要　声纹识别利用说话人生物特征的个体差异性,通过声音来识别说话人的身份。声纹具有非接触、易采集、特征稳定等特点,应用领域十分广泛。现有的统计模型方法具有提取特征单一、泛化能力不强等局限性。近年来,随着人工智能深度学习的快速发展,神经网络模型在声纹识别领域崭露头角。文中提出基于长短时记忆(Long Short-Term Memory,LSTM)神经网络的声纹识别方法,使用语谱图提取声纹特征作为模型输入,从而实现文本无关的声纹识别。语谱图能够综合表征语音信号在时间方向上的频率和能量信息,表达的声纹特征更加丰富。LSTM 神经网络擅长捕捉时序特征,着重考虑了时间维度上的信息,相比其他神经网络模型,更契合语音数据的特点。文中将 LSTM 神经网络长期学习的优势与声纹语谱图的时序特征有效结合,实验结果表明,在 THCHS-30 语音数据集上取得了 84.31%的识别正确率。在自然环境下,对于 3s 的短语音,该方法的识别正确率达 96.67%,与现有的高斯混合模型和卷积神经网络方法相比,所提方法的识别性能更优。

关键词:声纹识别;长短时记忆;语谱图;神经网络;深度学习
中图法分类号　TP391.4

Voiceprint Recognition Based on LSTM Neural Network

LIU Xiao-xuan,JI Yi and LIU Chun-ping
School of Computer Science and Technology,Soochow University,Suzhou,Jiangsu 215006,China

Abstract　Voiceprint recognition determines the identification of the given speaker by voice,using the individual differences of biological characteristics. It has a wide range of use,with the characteristics of non-contact,simple acquisition,feature stability and so on. The existing statistical methods of voiceprint recognition have the limitations of single-source extracted feature and weak generalization ability. In recent years,with the rapid development of artificial intelligence and deep learning,neural networks are emerging in the field of voiceprint recognition. In this paper,a method based on Long Short-Term Memory (Long Short-Term Memory,LSTM) neural network was proposed to realize text-independent voiceprint recognition,using spectrograms to extract voiceprint features as the model input. Spectrograms can represent the frequency and energy information of voice signal in time direction comprehensively,and express more abundant voiceprint features. LSTM neural network is good at capturing temporal features,focusing on the information in time dimension,which is more consistent with the characteristics of voice data compared with other neural network models. The method in this paper combined the long-term learning of LSTM neural network with the sequential feature of voiceprint spectrograms effectively. The experimental results show that 84.31% accuracy is achieved on THCHS-30 voice data set. For three seconds short voice in natural environment,the accuracy of this method is 96.67%,which is better than the existing methods such as Gaussian Mixture Model and Convolutional Neural Network.

Keywords　Voiceprint recognition,Long Short-Term Memory,Spectrogram,Neural network,Deep learning

1　引言

声纹是用声学仪器显示出的携带信息的声波频谱。人的发声器官个体差异性很大,而每个人的声学特征具有相对稳定性。声纹识别,又称说话人识别[1],通过从说话人的语音中提取出声纹特征,来建立模型以识别说话人的身份。声纹相较虹膜、指纹、人脸等其他生物特征,具有非接触、易采集、特征稳定、不易盗取和模仿等特点。因此,声纹识别的应用领域十分广泛[2],在刑侦鉴定、金融安全、智能家居等领域的需求颇多。

在声纹识别领域,传统的统计模型和机器学习方法[3-7]还占有相当大的比重。通过提取声纹的特征参数,例如线性预测倒谱系数(Linear Prediction Cepstral Coefficients,LPCC)[3]、美尔频率倒谱系数(Mel Frequency Cepstral Coefficients,MFCC)[4]等,使用隐马尔可夫模型(Hidden Markov Model,HMM)[5]、高斯混合模型(Gaussian Mixture Model,

基金项目:秦惠箐与李政道中国大学生见习进修基金;国家自然科学基金面上项目(61773272);江苏省高等学校自然科学研究重大项目(19KJA230001)
This work was supported by the Hui-Chun Chin and Tsung-Dao Lee Chinese Undergraduate Research Endowment(CURE),National Natural Science Foundation of China(61773272) and Natural Science Foundation of the Jiangsu Higher Education Institutions of China(19KJA230001).
通信作者:季怡(jiyi@suda.edu.cn)

GMM)[6]、高斯混合模型-通用背景模型(Gaussian Mixture Model-Universal Background Model，GMM-UBM)[7]等统计模型进行识别，是常见的声纹识别方法。这类方法虽然经典，但提取的特征参数过于单一，难以适应复杂的应用场景。

随着人工智能浪潮的来临，许多深度学习方法在特征识别领域大有作为[8-10]，使得神经网络模型在声纹识别中成为研究的焦点。鉴于卷积神经网络(Convolutional Neural Network，CNN)在图像处理和分类识别中的出色表现[8]，语谱图兼具语音数据表征和图像形式处理的特点，表达出的声纹特征更加丰富，成为了深度学习方法中常见的声纹特征表达形式[11]。语音数据具有时序性，处理时需要充分考虑时间维度上的信息。循环神经网络(Recurrent Neural Network，RNN)因其擅长捕捉时序特征的特点，在语音识别[9]、语种识别[10]等语音任务中取得了优异的成果。

通过深入研究并分析各类神经网络在相关领域的应用特点，本文提出基于 LSTM 神经网络的声纹识别方法。该方法使用语谱图提取声纹特征，对其进行标准化处理后作为模型的输入，通过对 LSTM 神经网络模型进行特征训练和隐式学习，来实现文本无关的高正确率声纹识别。

本文第 2 节回顾了声纹识别和神经网络的相关工作；第 3 节具体介绍了语谱图、LSTM 神经网络和本文的声纹识别设计流程；第 4 节展示了实验结果并进行了对比分析；最后总结全文并展望未来。

2 相关工作

根据对语音的文本内容是否限定，声纹识别可分为文本相关和文本无关两种类型[12]。文本相关的声纹识别要求待测语音与训练语音具有完全相同的内容，而文本无关的声纹识别对此则不做要求。早期的声纹识别技术多为文本相关的[12-13]，训练和测试语音均为固定内容，类似于固定的密码，泛化性能差且应用场景具有局限性。随着动态口令的推广，安全性高且无需记忆的随机口令密码，要求声纹识别技术向着文本无关的方向发展[14]。文本无关的声纹识别难度更大、要求更高，但因其灵活性强、操作简便，实际应用范围更为广泛。

声纹识别的研究聚焦于特征参数与识别模型两方面。声纹特征用于表征说话人发声状态下的个性信息，早期的声纹特征主要基于频谱分析[4,15]，对原始输入信号只进行较少层次的处理。例如，美尔频率倒谱系数利用美尔倒谱频率变换，结合快速傅里叶变换和离散余弦变化技术，表征出人耳的非线性听觉特性[4]。但是，这类浅层信息不涉及相邻帧之间的联系，无法体现语音信号的动态变化，具有局限性。随着图像处理技术和深度学习模型的不断发展，声纹特征从单一的频谱特征参数逐渐向结合了时间、空间信息的结构化特征进行转变。

在识别模型方面，GMM 是经典的声纹识别模型[6]，易于构建且性能稳定，很长一段时间内的声纹识别模型的发展都是以 GMM 为基础进行拓展的。如 Reynolds 等提出的 GMM-UBM[7]，基于所有目标说话人的语音数据进行通用背景模型(Universal Background Model，UBM)的参数估计，并将其用于说话人 GMM 的参数训练之中，在减小训练量的同时，避免了过拟合的发生。然而，这类统计模型不能充分利用声纹中的时间特征，应用场景较为固定，泛化性能差。

近年来，顺应人工智能的发展趋势，传统的特征识别技术正向着神经网络和深度学习的方向靠拢[8-10,16-18]。以神经网络为代表的深层结构模型具有更强的特征学习和表达能力，更适合进行复杂语音信号的处理和建模。通过大量带标签语音数据的训练，结合深度学习知识，深度神经网络(Deep Neural Network，DNN)在声纹识别系统中有了显著的性能提升[16]。CNN 模型在图像识别领域取得了巨大成功，文献[17]使用 CNN 提取语音的深层次空间特征，以语谱图形式进行表征，从而实现声纹识别。

语音数据具有时序特性，相比其他网络模型，RNN 模型更擅长处理序列信息。文献[18]通过 RNN 提取语音上下文关联的时序特征，进行文本相关的声纹识别。LSTM 神经网络[19]作为一种应用广泛的 RNN，解决了传统 RNN 存在的梯度爆炸和梯度消失问题，在语音识别[9]、语种识别[10]、情绪识别[20]等语音任务中表现出色。

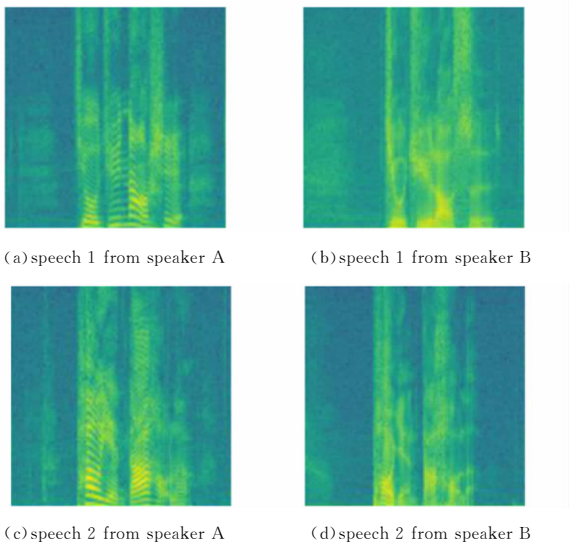
综上，利用语谱图综合表征语音数据的声纹特征，基于 LSTM 神经网络进行模型的构建与训练，能够将 LSTM 神经网络长期学习的优势与声纹语谱图的时序特征有效结合，实现文本无关的声纹识别，使其推广应用于更为一般的身份认证场景。

3 基于 LSTM 神经网络的声纹识别

3.1 语谱图

语音是由时间维度组成的三维信息，语谱图是语音信号的一种图像化表示方式。随着时间推移，在连续语音中，说话人的声纹特征会在前后时间片段的语音信息中产生关联，在时序中以某种形式展现出来。

值得注意的是，语谱图更注重对语音信号个性特征的表达，对语音的文本内容没有特别强调，这与本文“文本无关”的实验要求相符。图 1(a)和图 1(b)以及图 1(c)和图 1(d)的横向对比，展示了不同说话人在同一语料下的语谱图；图 1(a)和图 1(c)以及图 1(b)和图 1(d)的纵向对比，展示了同一说话人在不同语料下的语谱图。不同说话人亮纹的明显区别表明，声纹的个性特征与文本内容关系不大。



(a) speech 1 from speaker A　(b) speech 1 from speaker B

(c) speech 2 from speaker A　(d) speech 2 from speaker B

图 1　语谱图对比

Fig. 1　Comparison of spectrograms

语谱图是声音的时频域表示，相比单一时域表示的波形图，可以综合表征时间方向上的频率和语音能量信息，表达出更为深层的声纹特征，有利于模型的充分学习。因此，语谱图兼具语音数据表征和图像形式处理的特点，采用二维平面表达三维信息，可以综合运用图像处理方法进行分析。

3.2 LSTM 神经网络

人工神经网络由大量的人工神经元构成，通过参数和激活函数来模拟输入与输出的函数关系，是一种模仿生物神经网络结构和功能的模型。人工神经网络根据网络结构可分为前馈型神经网络、反馈型神经网络和记忆型神经网络。

前馈型神经网络每层的神经元根据收到的信息时序向前传播，接收上一层的输出并传递至下一层。反馈型神经网络的神经元在接收上一层的输出之余，还能获得自己的反馈信息。记忆型神经网络在反馈型的基础上引入记忆单元，神经元可以保存计算过程的中间状态，具有较强的记忆能力。

循环神经网络是一种反馈型神经网络，按照时间进行反馈，强调时序上的输入与上下文间的联系，擅长处理序列信息。但是，当反向传播规模过大时，循环神经网络易出现梯度消失和梯度爆炸的问题，导致训练的梯度无法在较长序列中传递，进而无法得到长距离的反馈信息。

长短期记忆神经网络是循环神经网络的进阶版本[19]，解决了传统循环神经网络在梯度上的问题，能够学习长期依赖信息。具体来说，用 LSTM 单元替代普通的网络神经元进行构建，并对每个 LSTM 单元引入遗忘门、输入门和输出门，能保护和控制信息。图 2 给出了 LSTM 单元的基本结构。

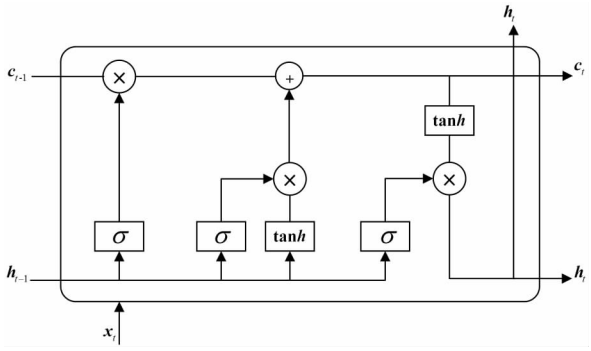


图 2 LSTM 单元

Fig. 2 LSTM cell

设 x_t 为当前时刻 t 的输入，h_{t-1} 为上一时刻的输出，c_{t-1} 为上一时刻的单元状态，则当前时刻的输出 h_t 和单元状态 c_t 的计算过程如下。这里，W 表示权值矩阵，b 表示偏置向量，σ 表示 Sigmoid 函数，tanh 表示 tanh 函数，$\circ$ 表示元素逐乘计算。

遗忘门决定了经过这一单元时需要舍弃的信息，计算式为：

$$f_t=\sigma(W_f[h_{t-1},x_t]+b_f) \tag{1}$$

输入门决定了需要更新至当前单元状态的信息，计算式为：

$$i_t=\sigma(W_i[h_{t-1},x_t]+b_i) \tag{2}$$

$$\tilde{c}_t=\tanh(W_c[h_{t-1},x_t]+b_c) \tag{3}$$

通过遗忘门和输入门，当前时刻的单元状态 c_t 更新为：

$$c_t=f_t\circ c_{t-1}+i_t\circ\tilde{c}_t \tag{4}$$

输出门决定了单元状态中当前可以输出的信息，计算式为：

$$o_t=\sigma(W_o[h_{t-1},x_t]+b_o) \tag{5}$$

$$h_t=o_t\circ\tanh(c_t) \tag{6}$$

语音数据具有时序性，语音个性特征在前后片段中均有所体现。因此，在对语音中的声纹特征进行学习时，应当充分考虑时间维度上的信息。基于上述分析，LSTM 神经网络对时序特征具有较强的捕捉能力，且规避了梯度消失和梯度爆炸问题，本文选择 LSTM 神经网络作为声纹识别模型，以学习声纹特征并对其进行识别。

3.3 声纹识别的设计流程

在声纹识别的应用阶段，通常是给出说话人的一段语音，对语音数据进行预处理后，以语谱图的形式提取出声纹个性特征，通过构建好的模型进行分类匹配，以确定该语音说话人的身份标签。完整的声纹识别流程如图 3 所示。

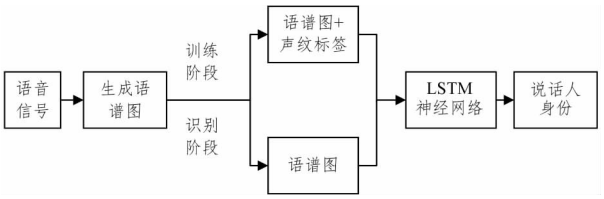


图 3 声纹识别流程

Fig. 3 Process of voiceprint recognition

3.3.1 语音信号预处理

自然场景下的语音信号带有环境和信道噪声干扰，需要进行预处理以还原声纹个性特征[21]。本文选用的预处理方法包括预加重、分帧和加窗。预加重对语音信号的高频部分进行加重，使频谱趋于平坦，一定程度上降低了噪声的影响。分帧将长语音划分成若干片段，即短语音帧。加窗操作将分帧后截断处的语音信号进行平滑过渡。

3.3.2 生成语谱图

使用传统滤波器提取到的特征参数，会对频域信息造成损失，而本文采用的语谱图是声音时频域的二维图像表示，能够表达出语音数据更为完整的个性特征。对预处理后的语音信号进行快速傅里叶变换(Fast Fourier Transform，FFT)，即可得到语音信号的能量密度谱，再对其进行彩色映射，从而生成对应的语谱图。为匹配神经网络，对语谱图进行标准化处理，裁剪为适合网络输入的相同尺寸。

综上，语谱图的生成流程如图 4 所示。

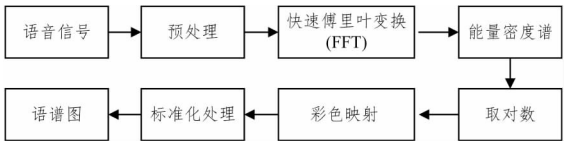


图 4 语谱图生成流程

Fig. 4 Generation of spectrograms

3.3.3 搭建并训练 LSTM 神经网络

LSTM 神经网络由一个或多个 LSTM 隐藏层构成，每一层由多个 LSTM 单元堆叠而成。网络的最后加入全连接层实现分类。模型训练采用有监督学习，输入标准化语谱图和对应的声纹标签，选择交叉熵函数作为损失函数，使用 AdaDelta 优化方法，通过大量数据学习及多次训练迭代，实现 LSTM 神经网络模型的训练。

4 实验

4.1 实验设置

实验的开发平台为 PyCharm，神经网络框架为 PyTorch。

实验环境为 Intel(R) Core(TM) m3-6Y30 CPU @ 0.90 GHz 1.50 GHz，安装内存为 4 GB。

一般情况下，录制得到的语音信号长度不同，生成的语谱图尺寸不一。为配合神经网络的输入层，将语谱图进行中心裁剪处理，裁剪为 161×161 大小的正方形彩色图像。语音录制的开始和结尾部分存在空白片段，所含信息量较少，采取中心裁剪可以最大限度地保留语音数据中含信息量多的部分。

本文构建的 LSTM 神经网络模型的结构如下：网络的第一层是输入层，输入数据为彩色语谱图，维度为 161×161×3；第二层是 LSTM 层，具有 400 个隐藏层神经元；第三层是全连接层，神经元数目为说话人标签数量；第四层是 Softmax 分类层，实现分类识别。实验中设置迭代次数为 100 次。

4.2　实验结果

首先在标准数据集上进行实验。本文使用的标准数据集为 THCHS-30，是由清华大学语音和语言技术中心发布的中文数据集[22]，在安静的办公室环境下录制而得，说话人绝大部分为女性，录音方式为说话人朗读新闻文本。每段音频长度在 10 s 左右，采样频率为 16 kHz，采样大小为 16 bits。

由于 THCHS-30 数据集存在标签标注错误的问题，对数据进行了排查和筛选，从中挑选 17 名说话人，每人 15 段语音，每个人的语音文本各不相同。对于每个说话人，随机选取 80%(12 个)的音频文件作为训练数据，将剩余 20%(3 个)的音频文件作为测试数据。对每个音频文件生成一张语谱图，即训练数据共有 204 个，测试数据共有 51 个。

本文的性能评价指标是识别正确率，即识别正确的语音数量与测试集中语音总数的比值。图 5 和图 6 分别给出了 LSTM 模型在 THCHS-30 训练集和测试集上的准确率(Acc)和损失值(Loss)的变化曲线。训练集上的准确率在 90 轮迭代后达到收敛，稳定在 98%附近；测试集上的最佳识别正确率达 84.31%。

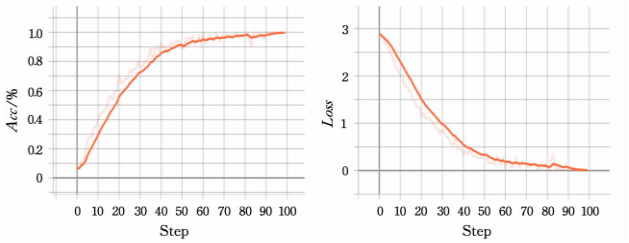


图 5　THCHS-30 训练集上准确率和损失值的曲线图

Fig. 5　Diagram of accuracy and loss on THCHS-30 train set

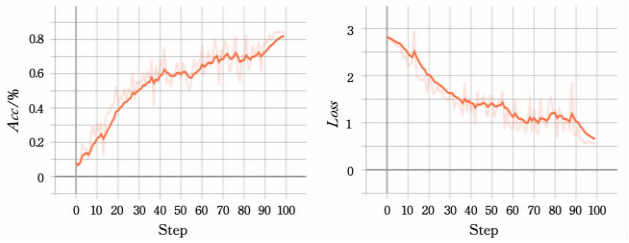


图 6　THCHS-30 测试集上准确率和损失值的曲线图

Fig. 6　Diagram of accuracy and loss on THCHS-30 test set

由图 5 和图 6 可知，LSTM 模型在训练集上的准确率随迭代次数的增加而平稳上升，但测试集上的准确率在上升过程中波动较大。本文选用的 LSTM 模型是一种结构较为复杂的神经网络模型，参数数量较多，且输入的语谱图特征维度较大，相比而言，使用的 THCHS-30 训练集中的样本数量过少。为了扩大数据规模、平衡性别比例、增强泛化性能，本文的 LSTM 模型在真实语音数据集上也进行了实验。

现实应用中，语音的采集大多在含背景噪声的自然环境下进行，且在门禁、考勤等身份验证领域，普遍使用动态口令式短语音，如随机四字短语。本文使用的真实语音数据集是在自然环境下采集的，包含 50 名 20 岁左右的说话人，男女性别比例一致。每人录音 15 次，每次录音长度为 3 s，采样频率为 16 kHz，采样大小为 16 bits，录音方式为说话人朗读指定的中文四字短语，每个人的短语文本各不相同，共有 750 个音频文件。与前一个实验相同，对于每个说话人，以 8∶2 的比例随机划分训练数据与测试数据，并对每个音频文件生成一张语谱图，即训练数据共有 600 个，测试数据共有 150 个。

图 7 和图 8 分别给出了 LSTM 模型在真实语音数据的训练集和测试集上的实验结果。可以看出，LSTM 模型在真实语音测试集上的最佳识别正确率达 96.67%，且收敛后基本不再波动。

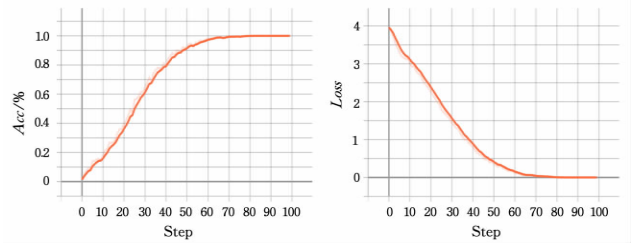


图 7　真实语音训练集上准确率和损失值的曲线图

Fig. 7　Diagram of accuracy and loss on real voice train set

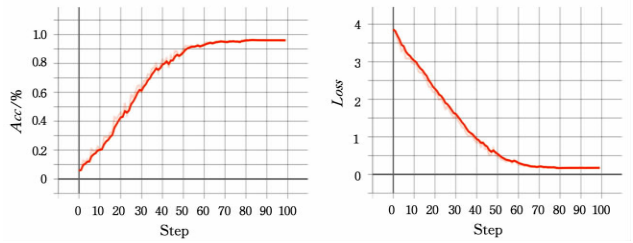


图 8　真实语音测试集上准确率和损失值的曲线图

Fig. 8　Diagram of accuracy and loss on real voice test set

为进行进一步对比，在同样的真实语音数据集上，对经典的统计模型 GMM-UBM 和卷积神经网络模型 ResNet18 进行了实验，表 1 列出了 GMM-UBM，CNN(ResNet18)和 LSTM 模型在真实语音测试集上准确率的对比。经比较可知，LSTM 神经网络模型的识别正确率相比 GMM-UBM 提升了 3.34%，相比 ResNet18 提升了 2.00%。

表 1　各模型的准确率对比

Table 1　Comparison with accuracy of each model

MODEL	ACC/%
GMM-UBM	93.33
CNN(ResNet18)	94.67
LSTM	96.67

实验结果印证了本文对声纹特征提取和声纹识别方法的论述，将 LSTM 神经网络长期学习的优势与声纹语谱图的时序特征有效结合，实现了真实场景下文本无关的高正确率声纹识别。

结束语　结合深度学习在声纹识别领域的研究成果，本

文提出了基于LSTM神经网络的声纹识别方法，利用语谱图提取声纹个性特征作为模型输入，并通过LSTM神经网络进行特征学习，实现文本无关的声纹识别。在50人3s短语音真实数据集上的实验中，取得了96.67%的识别正确率，高于现有的统计模型GMM-UBM和卷积神经网络ResNet18。基于深度学习和神经网络的发展，如何有效提升声纹识别的正确率、稳定性和泛化能力，将声纹识别技术推广至日常实际应用，或将成为声纹识别领域未来的研究方向。

参 考 文 献

[1] REYNOLDS D A. An overview of automatic speaker recognition technology[C]// IEEE International Conference on Acoustics. IEEE, 2011.

[2] FURUI S. Recent advances in speaker recognition[J]. Pattern Recognition Letters, 1997, 18(9): 859-872.

[3] ATAL B S. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification[J]. The Journal of the Acoustical Society of America, 1974, 55(6): 1304-1312.

[4] VERGIN R, O'SHAUGHNESSY D, FARHAT A. Generalized mel frequency cepstral coefficients for large-vocabulary speaker-independent continuous-speech recognition[J]. IEEE Transactions on Speech and Audio Processing, 1999, 7(5): 525-532.

[5] RABINER L R. A tutorial on hidden Markov models and selected applications in speech recognition[J]. Proceedings of the IEEE, 1989, 77(2): 257-286.

[6] REYNOLDS D A, ROSE R C. Robust text-independent speaker identification using Gaussian mixture speaker models[J]. IEEE Transactions on Speech and Audio Processing, 1995, 3(1): 72-83.

[7] REYNOLDS D A, QUATIERI T F, DUNN R B. Speaker verification using adapted Gaussian mixture models[J]. Digital Signal Processing, 2000, 10(1/2/3): 19-41.

[8] CHEN C, QI F. Review on Development of Convolutional Neural Network and Its Application in Computer Vision[J]. Computer Science, 2019, 46(3): 63-73.

[9] GRAVES A, MOHAMED A, HINTON G. Speech recognition with deep recurrent neural networks[C]// 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2013: 6645-6649.

[10] LOPEZ M I, GONZALEZ D J, PLCHOT O, et al. Automatic language identification using deep neural networks[C]// 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2014: 5337-5341.

[11] ZHENG C J, WANG C L, JIA N. Survey of Acoustic Feature Extraction in Speech Tasks[J]. Computer Science, 2020, 47(5): 110-119.

[12] ROSENBERG A E, SOONG F K. Evaluation of a vector quantization talker recognition system in text independent and text dependent modes[J]. Computer Speech & Language, 1987, 2(3/4): 143-157.

[13] FURUI S. Cepstral analysis technique for automatic speaker verification[J]. IEEE Transactions on Acoustics, Speech and Signal Processing, 1981, 29(2): 254-272.

[14] XIANG B, BERGER T. Efficient text-independent speaker verification with structural Gaussian mixture models and neural network[J]. IEEE Transactions on Speech and Audio Processing, 2003, 11(5): 447-456.

[15] LUCK J E. Automatic speaker verification using cepstral measurements[J]. The Journal of the Acoustical Society of America, 1969, 46(4B): 1026-1032.

[16] RICHARDSON F, REYNOLDS D, DEHAK N. Deep neural network approaches to speaker and language recognition[J]. IEEE Signal Processing Letters, 2015, 22(10): 1671-1675.

[17] HUANG J T, LI J, GONG Y. An analysis of convolutional neural networks for speech recognition[C]// 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2015: 4989-4993.

[18] HEIGOLD G, MORENO I, BENGIO S, et al. End-to-end text-dependent speaker verification[C]// 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016: 5115-5119.

[19] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[20] WU L Q, ZHANG D, LI S S, et al. Multi-modal Emotion Recognition Approach Based on Multi-task Learning[J]. Computer Science, 2019, 46(11): 284-290.

[21] HUA M, LI D D, WANG Z, et al. End-to-End Speaker Recognition Based on Frame-level Features[J]. Computer Science, 2020, 47(10): 169-173.

[22] WANG D, ZHANG X. Thchs-30: A free chinese speech corpus[J]. arXiv: 1512.01882, 2015.

LIU Xiao-xuan, born in 1999, undergraduate. Her main research interests include machine learning and pattern recognition.

JI Yi, born in 1973, Ph.D, associate professor, is a member of China Computer Federation. Her main research interests include pattern recognition and computer vision.