陈浩南,王坤赤,汤敏. 基于深度学习的声纹识别的研究综述[J]. 智能计算机与应用,2025,15(4):132-138. DOI:10.20169/j.issn.2095-2163.24110401

基于深度学习的声纹识别的研究综述

陈浩南, 王坤赤, 汤 敏
(南通大学 信息科学技术学院, 江苏 南通 226019)

摘 要: 声纹识别受到环境噪声、说话人的情绪状态、身体状况等因素影响。为了提高声纹识别的可靠性,利用特征学习的深度学习方法得到广泛关注。深度学习是基于深层神经网络模型和方法的机器学习,在众多领域都展现出了令人瞩目的应用潜力,例如语音识别、语音情感识别、语音分离等。本文聚焦于深度学习在声纹识别领域的应用研究。首先,概述声纹识别研究的基础知识,包括声纹特征提取、准确率的计算方法等。接着,探讨深度学习中基于 CNN、RegNet、RNN 的各种神经网络框架,并对其当前发展状况进行总结与分析。最后,还对声纹识别的未来发展趋势进行展望,包括语谱图实现和神经网络模型构建。目前,相较于传统的声纹识别,基于深度学习的声纹识别模型在识别准确率上已经获得显著提升,平均准确率能够达到 90%以上。

关键词: 声纹识别; 语音特征提取; 语谱图; 深度学习; 神经网络
中图分类号: TP391.4 **文献标志码**: A **文章编号**: 2095-2163(2025)04-0132-07

Research summary of voiceprint recognition based on deep learning

CHEN Haonan, WANG Kunchi, TANG Min
(School of Information Science and Technology, Nantong University, Nantong 226019, Jiangsu, China)

Abstract: Voiceprint recognition is affected by multiple factors such as environmental noise, speaker ′s emotional state and physical condition. In order to improve the reliability and accuracy of voiceprint recognition, deep learning, a machine learning based on deep neural network models and methods, shows its wide application potential in voiceprint recognition, emotion recognition and speech separation. This paper focuses on the application of deep learning in the field of voiceprint recognition. This paper firstly outlines the basic knowledge of voiceprint recognition research, including voiceprint feature extraction, accuracy calculation methods, etc. Then, various neural network frameworks based on CNN(Convolutional Neural Network), RegNet(Regularized Network) and RNN(Recurrent Neural Network) in deep learning are discussed, and their development in voiceprint recognition is summarized and analyzed. Finally, the article also looks forward to the future development trend of voiceprint recognition, including spectrogram improvement and neural network model optimization. At present, compared with the traditional voiceprint recognition, the voiceprint recognition model based on deep learning has been significantly improved in recognition accuracy, and the average accuracy can reach more than 90 %.

Key words: voiceprint recognition; speech feature extraction; spectrogram; deep learning; neural network

0 引 言

数字化时代中,声纹识别(Voice Print Recognition, VPR),作为一种新兴的生物特征识别技术,正逐步融入人们的日常生活。声纹识别之所以重要,则是因其能够利用每个人声音的独特性,包括音调、节奏、发音和语言习惯等。这些特征在一定程度上不仅稳定,而且也难以被复制。这种独特的差异性使得声纹成为一种高效可靠的说话人身份鉴别工具。声纹识别技术的发展历程可以追溯到 2000 年,当时主要采用 GMM-UBM[1](Gaussian Mixture Model-Universal Background Model)方法,随

基金项目: 国家自然科学基金(62371261); 南通市科技计划基金(JC2023076); 江苏省研究生科研与实践创新计划项目(KYCX24_3643); 农业重大技术协同推广计划项目(2024-ZYXT-11)。
作者简介: 陈浩南(2004—),男,本科生,主要研究方向:图像处理及分析。
通信作者: 汤 敏(1977—),女,博士,副教授,硕士生导师,主要研究方向:图像处理及分析。Email:tangmnt@163.com。
收稿日期: 2024-11-04

后发展为基于 i-vector[2] 的技术，并广泛应用于语音识别和安全验证等多个领域，成为保护个人和组织安全的重要工具。

作为机器学习和人工智能领域的关键分支，深度学习的核心优势在于能够构建多层级的神经网络模型，从而自动从海量数据中提取复杂的特征表示。这一技术的发展基于神经网络的演进，最早的神经网络模型可追溯至 1943 年，由 Warren McCulloch 和 Walter Pitts[3] 提出的 MCP（McCulloch-Pitts）模型。随后，神经网络的隐含层不断优化，直至 1998 年，LeCun 及其团队[4] 提出专为图像数据处理设计的卷积神经网络。CNN 通过卷积层捕捉局部特征，并利用池化层降低特征维度。在声纹识别领域，深度学习的应用推动了技术的进步。2014 年，基于深度神经网络的 d-vector[5] 方法被提出，专门用于处理文本相关的声纹识别任务。此后，声纹识别技术持续发展，引入了 end-to-end loss[6]、Siamese network[7]、xvector[8] 等创新方法，显著提升了声纹识别的性能。直到 2020 年，Villalba 等学者[9] 进一步引入了 ResNet34 神经网络作为编码器，并采用了 angular Softmax 和 additive angular margin Softmax 作为损失函数，以替代传统的 Softmax，从而进一步提升了声纹识别的性能。

本文综述基于深度学习的声纹识别研究，将从如下方面进行总结：

（1）语音特征提取，包括声学特征如 MFCC（Mel-frequency Cepstral Coefficients，MFCC）和梅尔滤波器组系数（Mel Filter Bank，FBank）等，并讨论了在声纹识别中各自的重要性。

（2）深度学习在声纹识别的应用，包括 CNN、RegNet 神经网络和 RNN 等。这些深度学习模型能够自动从语音数据中学习复杂的特征表示，提高声纹识别的准确性和鲁棒性，平均准确率 90%以上。

（3）探讨声纹识别技术的发展趋势，并提出未来的前景展望。例如，可以通过改进语谱图的特征提取方法，或者对现有神经网络结构进行优化，进一步提升声纹识别性能。

1　声纹识别研究基础

1.1　语音特征提取

语音样本的特征提取在声纹识别领域发挥至关重要的作用。理想的特征提取方法应能够捕捉说话人的关键信息，并且捕捉质量直接关系到最终实验结果的准确性。尽管众多学者已经提出多种特征提取技术，但迄今为止，尚未有一种方法被公认为语音样本特征提取的通用标准。因此，声纹识别领域亟需一种效果显著的特征提取技术。目前，声学特征是声纹识别领域中广泛采用的语音特征类型。

声学特征[10] 包括倒谱（cepstrum）、线性预测编码（Linear Predictive Coding，LPC）、MFCC、Mel 缩放功率谱图（Mel Spectrogram）、FBank、伽马频率倒谱系数（Gammatone Frequency Cepstrum Coefficient，GFCC）等，其中常用的是 MFCC 与 FBank。

梅尔滤波器组[11] 是将语谱图的频率从 Hz 刻度转换为 Mel 刻度，使用一组等高的三角滤波器，每个滤波器的起始点在上一个滤波器的中点处。梅尔滤波器组系数是通过将语谱图与梅尔滤波器组进行点乘得到的，代表每个滤波器在不同帧上的能量。

MFCC 是在 Mel 滤波器作用于语谱图后，通过离散余弦变换（Discrete Cosine Transform，DCT）压缩梅尔谱得到的一组不相关的系数，通常只保留前 13 个系数。

MFCC 基于听觉系统提出：人类听觉系统所感知到的声音频率 $Mel(f)$ 与该声音实际物理频率 f 之间的对应关系并不是完全线性的，而是在一定范围内呈对数关系，描述了 *Mel* 频率与实际频率非线性的关系，可以近似为如下公式：

$$Mel(f) = 2\,595\log(1+f)/700 \tag{1}$$

对一段输入语音信号 $x(n)$，MFCC 和 FBank 的提取过程如图 1 所示。在对语音信号进行预加重、分帧、加窗等操作后进行快速傅里叶变换，将时域信号转换为频域信号，变换公式如下：

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j\frac{2\pi}{N}k_n}, k = 0,1,\cdots,K-1 \tag{2}$$

然后，进行 Mel 滤波，所包含的滤波器个数为 M，中心频率是 $f(m)$。m 的取值为 $1,2,\cdots,M$，这里 M 为 22 ~ 26 内的数值。$f(m)$ 之间的距离是不同的，与 m 值同步变化，并且满足如下公式：

$$\begin{aligned} &Mel(f(m)) - Mel(f(m-1)) = \\ &Mel(f(m+1)) - Mel(f(m)) \end{aligned} \tag{3}$$

Mel 滤波器[12] 的频率响应定义为：

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

其中，$\sum_{M=0}^{M-1} H_m(k) = 1$。

接着，对 Mel 滤波器组的输出取对数运算得到 FBank，最后，进行离散余弦变换得到 MFCC。

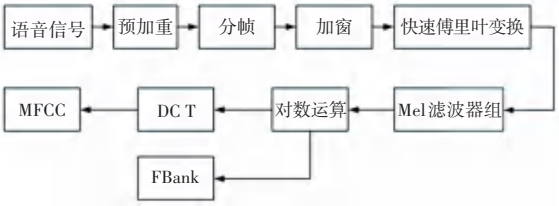

图 1 MFCC 与 FBank 特征提取流程

Fig. 1 MFCC and FBank feature extraction process

1.2 深度学习

基于深度学习的声纹识别方法[13]是使用神经网络来提取语音特征，从而实现声纹识别。神经网络具有自适应性强、识别准确率高、鲁棒性高、泛化能力强等特点。常用的神经网络包括 CNN[14]、RegNet 网络以及 RNN[15]网络等。

CNN 的核心理念在于运用卷积运算来捕捉图像特征。这种运算依赖于局部感知能力，通过在输入图像上滑动卷积核并执行卷积运算，从而识别图像中的局部模式。卷积运算的结果，即特征图[16]（Feature Map），能够通过不同的卷积核捕捉到多样化的特征。一个典型的 CNN 架构包括输入层（Input Layer）、若干卷积层（Convolution Layer）、池化层（Pooling Layer）、激活函数以及全连接层（Full-Connected Layer）。卷积层负责提取图像特征，池化层用于缩减特征图的维度，而全连接层则负责执行分类或回归任务。在全连接层，所有特征被整合成一个向量，并通过 Softmax 函数进行分类，Softmax 因其在处理多类分类问题中的高效性而获得广泛采用。

RegNet 网络[17]可以分为 stem，body 和 head 三个部分，如图 2 所示。在 RegNet 网络结构中，Body 部分为网络的主体部分，主体部分包含一系列的网络阶段（RegStage）。每个 RegStage 中块的数量（d_i）、输出特征矩阵的通道数（w_i），以及块中每个组的宽度（g）是网络的主要参数，可以根据需要进行调整。Stem 是一个普通的卷积层，其主要作用是对输入图像进行预处理，通常包含一个步长为 2，卷积核大小为 3×3 的卷积层，用于生成初始的特征图。Head 层则负责对从主体部分提取的特征进行分类，通常包含全局平均池化、随机失活神经元函数（Dropout）和全连接层（Fully Connected layer，FC）。灵活调整网络的深度、宽度和分组卷积，从而实现性能优化。

语音信号是时间序列数据，RNN 因其能处理时间序列数据而经常用在语音识别领域中。但 RNN 训练时可能会遇到梯度消失或爆炸的问题，因为梯度会随时间步长指数增长或衰减，导致模型难以训练和数值不稳定。为了解决这些问题，随即提出了 LSTM（Long Short-Term Memory）。LSTM 通过记忆细胞和门控机制（遗忘门、输入门、输出门）来控制信息流动，有效处理长序列数据，在语音识别等领域表现优异。

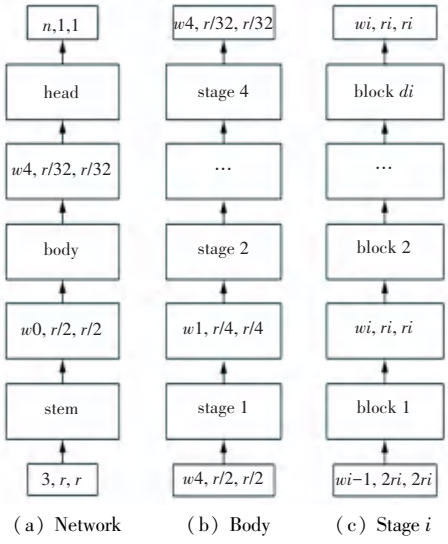

图 2 RegNet 网络结构图

Fig. 2 RegNet network structure diagram

LSTM 是一种特殊的循环神经网络[18]，对于序列数据等带有时间维度的问题，这种结构能够很好地处理长期依赖关系的建模，其网络架构如图 3 所示。

图 3 中，每个方框表示一个细胞，x_t 是当前架构输入，h_{t-1} 表示上阶段的状态，考虑过去的状态，在这个过程中遗忘门的输入是 h_{t-1} 和 x_t，输出是一个 0~1 之间的数值，代表对每个在细胞状态 C_{t-1} 的遗忘指数，0 表示完全丢弃，1 表示完全保留，其计算表达式如下：

$$f_t = \sigma(\boldsymbol{W}_f \cdot [h_{t-1}, x_t] + b_f) \tag{5}$$

其中，f_t 表示函数输出；σ 表示激活函数；$\boldsymbol{W}_f$ 表示权重矩阵。

其次，LSTM 的输入门是控制输入信息对于当前时刻重要性的机制，将当前时刻的输入 x_t 和上一时刻的隐藏状态 h_{t-1} 作为输入，通过激活函数计算出输入门的激活值 i_t，参见下式：

$$i_t = \sigma(\boldsymbol{W}_i \cdot [h_{t-1}, x_t] + b_i) \tag{6}$$

通过 tanh 函数计算候选记忆细胞 C_t，如下所示：

$$C_t = \tanh(\boldsymbol{W}_C \cdot [h_{t-1}, x_t] + b_C) \tag{7}$$

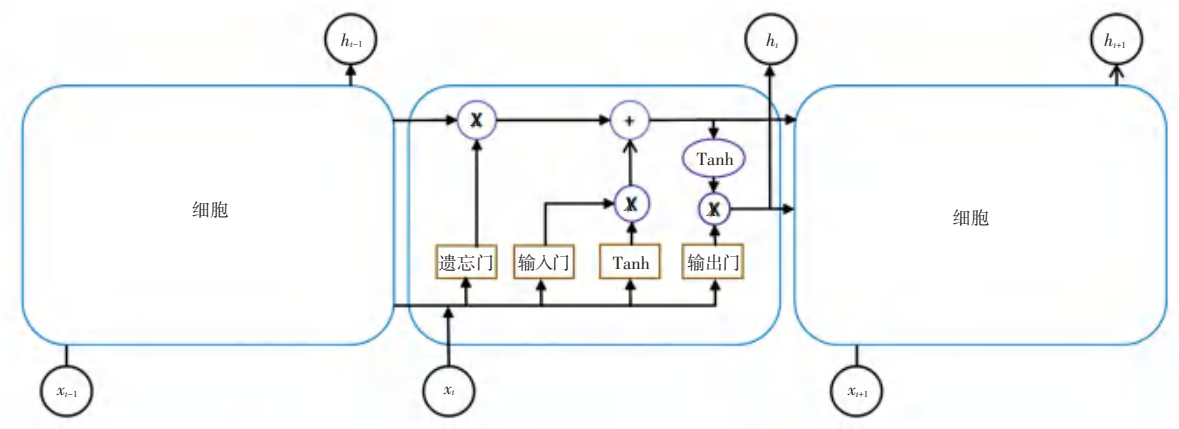

图 3　LSTM 网络

Fig. 3　LSTM network

将上一时刻的记忆细胞 C_{t-1} 和当前时刻的候选记忆细胞 C_t 以及输入门的激活值 i_t 组合起来，通过按元素乘法和加法计算出新的记忆细胞，用符号 C_t 表示，定义公式如下：

$$C_t = f_t \times C_{t-1} + i_t \times C_t \tag{8}$$

最后，输出门基于当前细胞的状态决定输出值，即控制隐藏状态中哪些信息可以传递到下一个时刻，使用一个 Sigmoid 函数来确定需要输出的细胞状态，推得公式如下：

$$o_t = \sigma(\boldsymbol{W}_o \cdot [h_{t-1}, x_t] + b_o) \tag{9}$$

接下来，将当前时刻的隐状态和细胞状态输入到一个 tanh 函数中，得到一个在−1 到 1 之间的值，这个值会被乘以输出门的输出，最终得到 LSTM 当前时刻的输出，可用下式求得：

$$h_t = o_t \times \tanh(C_t) \tag{10}$$

1.3　准确率验证

1.3.1　声纹识别评价指标

在声纹识别的多样化应用场景中，评估模型性能需依据具体情境选择合适的评价指标。在多分类的声纹识别任务中，准确率（Accuracy，*ACC*）成为了一个更合适的评价标准，可反映模型正确识别样本的能力，准确率越高，说明模型性能越出色。此外，接收者操作特征曲线（Receiver Operating Characteristic Curve，*ROC*）也是一个评估模型性能的重要工具。*ROC* 曲线展示了错误接受率（False Accept Rate，*FAR*）与错误拒绝率（False Reject Rate，*FRR*）之间的权衡关系。通过这些评价指标，可以更全面地评估和比较不同声纹识别模型的性能。

错误接受率指的是系统错误地将 2 个实际上属于不同类别的样本识别为同一类别的情况，对应的识别结果即为错误接受案例。而错误拒绝率则是指系统错误地将 2 个实际上属于同一类别的样本识别为不同类别的情况，对应的识别结果即为错误拒绝案例。错误接受率和错误拒绝率的计算方法公式具体如下：

$$P_{\mathrm{FAR}} = \frac{FP}{FP + TN} = P_{\mathrm{FPR}} \tag{11}$$

$$P_{\mathrm{FRR}} = \frac{FN}{TP + FN} = 1 - P_{\mathrm{TPR}} \tag{12}$$

其中，P_{FAR} 表示错误接收率；P_{FPR} 表示假正率；P_{FRR} 表示错误拒绝率；P_{TPR} 表示真正率。

通过这些指标，可以精确地衡量和优化声纹识别系统的性能。真正例（True Positive，*TP*）、假正例（False Positive，*FP*）、真负例（True Negative，*TN*）和假负例（False Negative，*FN*）的关系见表 1。

表 1　分类结果混淆矩阵

Table 1　Classification result confusion matrix

真实情况	预测结果	
	正例	反例
正例	真正例（*TP*）	假反例（*FN*）
反例	假正例（*FP*）	真反例（*TN*）

准确率（Accuracy，*ACC*）的计算公式如下所示：

$$ACC = \frac{TP + TN}{TP + FP + FN + TN} \times 100\% \tag{13}$$

1.3.2　损失函数

对于声纹识别模型，一般采取交叉熵损失函数[19]（Cross Entropy Loss）作为损失函数，其定义公式如下：

$$L = -\sum_{h,w}\sum_{c} Y_n^{(h,w,c)} \log(S(X_n)^{(h,w,c)}) \tag{14}$$

其中，L 表示交叉熵损失函数；Y_n 表示标签图像；X_n 表示输入图像；S 表示分割图像；h、w、c 分别表示图像的高度、宽度和通道数。交叉熵损失会对

那些分类错误的实例施加更大的惩罚,对于分类错误的实例,梯度将会更大,进而使得模型在训练过程中有更快的收敛速度。

2 声纹识别模型

2.1 声纹识别研究基础

声纹识别通常分为基于传统统计特征的方法和基于深度学习的方法两大类,随着对精确度要求的提高,学界和业界越来越倾向于采用深度学习方法进行声纹识别。在基于深度学习的声纹识别方法中,输入数据通常采用语谱图的形式。这种方法能够有效地捕捉到语音信号中的关键特征,从而提高识别的准确性。但是,语谱图现在产生很多变体,对近期相关文献采用的语音特征及其特点归纳总结见表 2。文献[19-21]使用 FBank 特征作为神经网络的输入,类似于人耳对音频处理的方式,拟合人耳接收声音信号的特性,因此提取到的声纹特征更多地保留声音信号的本质。另一方面,文献[22-26]则采用 MFCC 作为神经网络输入,在 FBank 基础上进行离散余弦变换得到 MFCC。在此基础上,文献[17]采用 MFCC + Mel Spectrogram + Contrast 的语音特征,极大地提高了识别准确率。此外,文献[27]结合 MFCC 和 GFCC,形成梅尔伽马倒谱系数(MGCC),增强说话人的区分性。

表 2 声纹识别方法汇总

Table 2 Summary of various methods of voiceprint recognition

神经网络	文献序号	语音特征	特点
CNN	[20]	MFCC	结合了 CNN 和受限玻尔兹曼机(RBM),用于处理小样本数据集的声纹识别问题,在 TIMIT 语音数据库的识别准确率达到了 97.80%
	[21]	spectrogram	通过维纳滤波去噪、词嵌入语谱图降维,以及结合 CNN 和 LSTM 的深度学习模型来进行声纹识别,在标准语音数据集测试集的识别准确率为 97.42%
	[22]	MFCC	比较了自定义 CNN 与多个预训练网络使用语音图和 MFCC 图的图像输入。还考虑了基于声学特征和使用朴素贝叶斯模型的多类分类的机器学习方法。自定义的、较浅的 CNN 在灰度语音图上训练获得最准确的结果为 90.15%
	[23]	FM spectrogram	结合了一维 CNN 和焦点模块(FM),用于提取说话人特征,并减少时域和空域中的异质性,从而加快层处理速度。此外,提出了一种基于 Softmax 损失和平滑 $L1$ 范数的联合监管方法,以提高效率。识别准确率达到 99.21%
	[24]	MFCC + GFCC	通过线性加权的方式结合了 MFCC 和 GFCC,以增强说话人的区分性,应用压缩激励机制的 CNN 和双向门控循环单元网络(BiGRU)的集成网络(CNN-SE-BiGRU),平均识别率达到了 96.05%
	[25]	FBank	通过分组卷积和 CBAM(Convolutional Block Attention Module)改进 CNN 网络,在标准数据集上识别准确率达到了 99.07%
	[17]	MFCC + Mel Spectrogram + Contrast	研究使用了 4 种不同的特征:MFCC、Mel Spectrogram、Chroma、频谱对比度(Contrast)和音调中心(Tonnetz),相互混合。并运用 CNN 模型,分类准确率达到 99.52%
RegNet	[19]	FBank	使用神经网络架构搜索(NAS)技术,通过 RegNet 神经网络寻找最佳参数组合,在 THCHS-30 数据集的识别准确率达到 91.95%
RNN	[26]	FBank	利用 CNN 处理图像的优势从语谱图中提取语音信号的个性特征,再将这些特征输入到深度 RNN 中完成声纹识别,识别准确率达到了 95.42%
	[27]	MFCC	通过 PCA(Principal Component Analysis)进行特征降维并通过 LSTM 网络的自训练和隐式学习,在 THCHS-30 数据集识别准确率为 91.84%。
	[28]	MFCC	保留了语音数据中的静音段,利用 RNN 能够提取与上下文相关的特征,包括说话方式、节奏等,使得特征信息更加完备,识别准确率达到 90%以上
	[29]	MFCC	采用时延神经网络,通过多层循环神经网络或双向循环神经网络[30]的方式进一步进行拓展。并采用改进型的循环神经网络模型。识别准确率为 98.63%

2.2　应用场景

随着电信网络、通信行业的发展，声纹识别将在保护用户隐私方面具有重大作用。通过对声纹特征的精确识别，可以有效辨别出单一特征生成的伪造语音[28]。此外通过声纹鉴定，形成语音证据[29]，可以有效防止电信诈骗，对于提升通信安全、保护用户隐私具有积极现实意义。

2.3　声纹识别的难点

在目前的语音处理领域，提取语音特征的技术尚未达到精细化水平，这导致许多关键信息未能被充分捕捉。此外，背景噪声的干扰也是一个不容忽视的问题，可能会严重影响语音识别的准确性。更复杂的是，个体的声音特征并非一成不变，可能会因情绪波动、健康状况等因素影响而产生波动，这些变化进一步削弱了声纹的稳定性和识别的准确性。因此，开发更为精确的语音特征提取方法，提高声纹识别的可靠性和稳定性，显得尤为关键。

面对个体声音特征的易变性，声纹识别领域还面临着数据稀缺的挑战。如何设计和优化合理的神经网络模型，使其能够在有限的数据条件下避免过拟合，并实现对声纹的泛化识别，是该领域亟待解决的问题。这不仅需要深入研究声音特征的复杂性，还需要在模型设计上进行创新，以适应不断变化的声音模式。

2.4　声纹识别数据库

声纹识别数据库是专门用于收集、管理和维护声纹数据的系统，在声纹识别技术的研究与应用领域起到基础性的支撑作用。构建声纹识别数据库的过程，涵盖声纹数据的采集、存储、管理以及应用等多个环节。全球范围内，许多机构都在致力于声纹数据库的整理与共享，常用的声纹识别数据库及其特点参见表 3。

表 3　声纹识别数据库汇总

Table 3　Summary of voiceprint recognition databases

数据库名称	网址	特点
TIMIT	https://catalog.ldc.upenn.edu/LDC93S1	由德州仪器(TI)、麻省理工学院(MIT)和 SRI International(SRI)合作构建。来自美国 8 个主要方言地区的 630 位说话者，每人说 10 个句子。文献[17]、[22]、[25]中使用。
Librispeech	https://www.openslr.org/12	大规模英语语音识别数据集，包含大约 1 000 h 的 16 kHz 采样率的英语有声读物录音。这些录音来源于 LibriVox 项目。文献[10]、[21]中使用。
VoxCeleb	https://www.robots.ox.ac.uk/~vgg/data/voxceleb/	大规模的说话人识别数据集，包含从 YouTube 视频中提取的大量真实场景下的语音片段，包含超过 150 000 条语音片段，来自 1 251 位不同的说话人，这些说话人涵盖了不同的种族、口音、职业和年龄，数据集性别平衡。文献[7]、[8]、[9]、[25]、[26]、[30]中使用
THCHS-30	https://www.openslr.org/18/	由清华大学语音与语言技术中心(CSLT)发布的一个开放式中文语音数据库，总时长超过 30 h，采样频率 16 kHz，采样大小 16 bits。文献[19]、[28]中使用
Free ST Chinese Mandarin Corpus	https://www.openslr.org/38/	由 Surfingtech(www.surfing.ai)提供的免费中文普通话语料库，包含 855 名说话者的 102 600 个发音。文献[21]中使用
Aishell-1	https://www.aishelltech.com/kysjcp	开源的中文普通话语音数据集，由北京希尔贝壳科技有限公司发布。包含约 178 h 的语音数据，由 400 名来自中国不同地区、具有不同口音的人的声音组成。文献[22]、[26]中使用

3　声纹识别技术展望

随着技术的不断进步，声纹识别将在提高识别准确性和鲁棒性方面持续突破[31-32]。深度学习模型将在声纹识别中发挥越来越重要的作用，特别是在特征提取和模型训练方面。未来的研究将集中于以下方面：

(1)开发新的语谱图提取方法，以生成更精确的 FBank 图和 MFCC 图，从而更好地捕捉说话人的关键信息。

(2)探索新的智能算法，提升声纹识别的智能化和自动化水平，也是当前研究的重点之一。

(3)在环境噪音和语音变化等的推动下，声纹识别技术面临更高的标准和要求。

为了应对这些挑战，必须持续进行技术攻关，以研发出更准确、更可靠的声纹识别系统。此外，由于

声纹在语音交互中占据的基础性地位,则将更多地融入语音识别、语音分离、语音增强和语音转换等技术中,发挥其核心作用。

4 结束语

本文深入探讨声纹识别技术的核心价值、发展历程、以及国内外研究动态。首先分析语音特征提取的关键技术,接着探讨深度学习在声纹识别领域的应用,并结合应用对不同的声纹识别模型展开了分析。研究表明,尽管声纹识别技术已取得显著进步,但仍面临若干挑战和限制。因此,未来的研究应致力于更深层次的探索,以期提升声纹识别技术的实用性和准确性。同时,随着人工智能和机器学习技术的持续发展,预计声纹识别技术将在不久的将来实现更多创新突破。

参考文献

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

[2] KUMARI J T, JAYANNA H. I-vector-based speaker verification on limited data using fusion techniques[J]. Journal of Intelligent Systems,2018,29(1):565-582.

[3] MCCULLOCH W S, PITTS W. A logical calculus of the ideas immanent in nervous activity [J]. Bulletin of Mathematical Biophysics, 1943, 5(4): 115-133.

[4] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[5] DEHAK N, KENNY P J, DEHAK R, et al. Front-end factor analysis for speakerverification[J]. IEEE Transactions on Audio, Speech and Language Processing, 2010, 19(4): 788-798.

[6] 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). Piscataway,NJ:IEEE, 2016: 5115-5119.

[7] SNYDER D, GHAHREMANI P, POVEY D, et al. Deep neural network-based speaker embeddings for end-to-end speaker verification [C]//2016 IEEE Spoken Language Technology Workshop (SLT). Piscataway,NJ:IEEE, 2016: 165-170.

[8] SNYDER D, GARCIA-ROMERO D, POVEY D, et al. Deep neural network embeddings for text-independent speaker verification[C]//Interspeech. Stockholm, Sweden:ISCA,2017: 999-1003.

[9] VILLALBA J, CHEN N, SNYDER D, et al. State-of-the-art speaker recognition with neural network embeddings in NIST SRE18 and speakers in the wild evaluations [J]. Computer Speech & Language, 2020, 60: 101026.

[10]彭绪鹏. 基于深度学习的声纹识别系统的研究与实现[D]. 北京:北京交通大学,2023.

[11] DAVISS B, MERMELSTEIN P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1980, 28(4): 357-366.

[12] HUNG W. Derivation of robust mel frequency cepstral features based on SNR-dependent adaptive filter bank analysis [J]. Electronics Letters,2001,37(22):1369-1370.

[13]蒲立力. 基于深度神经网络的声纹识别方法研究[D]. 桂林:桂林理工大学,2023.

[14]周飞燕,金林鹏,董军. 卷积神经网络研究综述[J]. 计算机学报,2017,40(6):1229-1251.

[15]杨丽,吴雨茜,王俊丽,等. 循环神经网络研究综述[J]. 计算机应用,2018,38(S2):1-6.

[16]尹宝才,王文通,王立春. 深度学习研究综述[J]. 北京工业大学学报,2015,41(1):48-59.

[17] YÜCESOY E. Genderrecognition based on the stacking of different acoustic features[J]. Applied Sciences, 2024, 14(15): 6564.

[18]GERS F A, SCHMIDHUBER J . LSTM recurrent networks learn simple context-free and context-sensitive languages [J]. IEEE Transactions on Neural Networks, 2001,12(6):1333-1340.

[19]黄一舟,崔璨,李增. 基于 RegNet 神经网络的声纹识别[J]. 智能计算机与应用,2024,14(8):197-202.

[20]王茂,何勇. 基于 FBank 特征与改进 CNN 的声纹识别研究[J]. 智能计算机与应用,2024,14(8):40-47.

[21]余玲飞,刘强. 基于深度循环网络的声纹识别方法研究及应用[J]. 计算机应用研究,2019,36(1):153-158.

[22] SUN Cunwei, YANG Yuxin, WEN Chang, et al. Voiceprint identification for limited dataset using the deep migration hybrid model based on transfer learning [J]. Sensors, 2018, 18(7): 2399.

[23]COSTANTINI G, CESARINI V, BRENNA E. High-Level CNN and machine learning methods for speaker recognition [J]. Sensors, 2023, 23(7): 3461.

[24] POEMOMO A, KANG D K. Biased dropout and crossmap dropout: learning towards effective dropout regularization in convolutional neural network [J]. Neural Networks, 2018, 104: 60-67.

[25]冯毅夫,王华锋,徐雷,等. 一种基于 RNN 的声纹识别方法[P]. 中国:CN107731233B,2024-10-20.

[26]和椿皓. 基于时延和循环神经网络的声纹识别算法研究[D]. 保定:河北大学, 2024.

[27]高锦凤,陈玉,魏永明,等. 一种改进 YOLOv3 的学校场所目标识别方法 [J]. 中国科学院大学学报, 2023,40(4):531-539.

[28]张宇翔,李茁,陆镜泽,等. 基于声纹特征的伪造语音检测[J]. 声学学报,2025,50(1):201-210.

[29]赵凯,李御瑾,付顺顺,等. 声纹鉴定在电信网络诈骗治理中的应用研究[J]. 中国人民公安大学学报(自然科学版),2024,30(3):70-76.

[30]董莺艳. 基于深度学习的声纹识别方法研究[D]. 重庆:重庆理工大学,2019.

[31] VELAYUTHAPANDIAN K, SUBRAMONIAMSP. A focus module-based lightweight end-to-end CNN framework for voiceprint recognition[J]. Signal, Image and Video Processing, 2023, 17: 2817-2825.

[32] CHUNG J S, HUH J, MUN S, et al. In defence of metric learning for speaker recognition[J]. arXiv preprint arXiv,2003.11982,2020.