计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190800054

基于帧级特征的端到端说话人识别

花 明 李冬冬 王 喆 高大启
华东理工大学信息科学与工程学院 上海 200237
(961564330@qq.com)

摘 要 现有的说话人识别方法仍存在许多不足。基于话语级特征输入的端到端方法由于语音长短不一致需要将输入处理为同等大小,而特征训练加后验分类的两阶段方法使得识别系统过于复杂,这些因素都会影响模型的性能。文中提出了基于帧级特征的端到端说话人识别方法。模型采用帧级语音作为输入,同等大小的帧级特征有效解决了话语级语音输入长度不一致的问题,且帧级特征可保留更多的话者信息。与如今主流的两阶段法识别系统相比,端到端的识别方法将特征训练和分类打分一体化,简化了模型的复杂性。在训练阶段,每段语音被分帧成多个帧级语音输入到卷积神经网络(Convolutional Neural Networks,CNN)用于训练模型。在评估阶段,训练好的CNN模型对帧级语音进行分类,每段语音基于多个帧的预测得分计算该条语音数据的预测类别。每段语音的类别通过取各帧最多预测类别和各帧预测值平均的方法来计算。为了验证方法的有效性,使用普通话情感语音语料库(MASC)的语音数据进行训练和测试。实验结果表明,与现有方法相比,基于帧级特征的端到端识别方法的性能表现更佳。

关键词:说话人识别;端到端;卷积神经网络;帧级特征;话语级语音
中图法分类号 TP301

End-to-End Speaker Recognition Based on Frame-level Features

HUA Ming,LI Dong-dong,WANG Zhe and GAO Da-qi
School of Information Science and Engineering,East China University of Science and Technology,Shanghai 200237,China

Abstract There are still many shortcomings in the existing speaker recognition methods. The end-to-end method based on utterance-level features requires to process the input to be the same size due to the inconsistency of the speech length. The two-stage method of feature training with posterior classification makes the recognition system too complex. These factors affect the performance of the model. This paper proposed an end-to-end speaker recognition method based on frame-level features. The model uses frame-level speech as input,and the same size frame-level features effectively solve the problem of inconsistent speech-level speech input length,and the frame-level features can retain more speaker information. Compared with the mainstream two-stage identification system,the end-to-end identification method integrates feature training and classification,which simplifies the complexity of the model. During the training phase,each speech is segmented into multiple frame-level speech inputs to a Convolutional Neural Network (CNN) for training the model. In the evaluation phase,the trained CNN model classifies the frame-level speech,and each segment of speech calculates the prediction category of the speech data based on the prediction scores of multiple frames. The maximum predicted category of each frame and the average prediction value of each frame are adopted to calculate the class of each segment of speech respectively. In order to verify the validity of the work,the speech data of the Mandarin Emotional Speech Corpus (MASC) were used for training and testing. The experimental results show that the end-to-end recognition method based on frame-level features achieves better performance than the existing methods.

Keywords Speaker recognition,End-to-end,Convolutional Neural Networks,Frame-level features,Utterance-level speech

1 引言

声音是包含个人信息的重要载体,从中可以分辨出说话人的身份、情感等多种信息[1]。语音在声学仪器中所显示的声波频谱即为声纹,它是由频率、强度等多种特征维度组成的生物特征。相较于指纹、人脸、虹膜等其他生物特征,声纹具有非接触、易接受、收集成本低、伪造困难等优势。

声纹识别,也称说话人识别(Speaker Recognition,SR),

到稿日期:2019-08-13 返修日期:2019-11-28 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金项目(61806078);国家重大新药开发科技专项(2019ZX0921004);上海市教育发展基金会和上海市教育委员会"曙光计划"(61725301)
This work was supported by the Natural Science Foundation of China (61806078), National Major Scientific and Technological Special Project for "Significant New Drugs Development"(2019ZX09201004) and "Shuguang Program" supported by Shanghai Education Development Foundation and Shanghai Municipal Education Commission (61725301).
通信作者:李冬冬(ldd@ecust.edu.cn)

其从语音中提取高层声学特征，并建立模型以得到识别结果[2]。Fbank 特征(FilterBank)是常见的声学特征，其利用快速傅里叶变换(Fast Fourier Transformation，FFT)将时域信号转换为频域信号，再利用对数操作得到信号的时域和频率间的关系信息。梅尔频率倒谱系数(Mel Frequency Cepstral Coefficients，MFCC)[3]在 Fbank 的基础上进一步做离散余弦变换(Discrete Cosine Transform，DCT)，它采用 Mel 刻度频率来描述人耳对频率感知的非线性特征。Fbank 特征间的相关性比较高，MFCC 则具有更好的判别度[3]。在识别模型方面，Dr. Reynolds 提出的高斯混合模型-通用背景模型(Gaussian Mixture Model-Universal Background Model，GMM-UBM)[4-5]是 SR 中的经典模型。它先训练与说话人无关的特征，接着使用登记的说话者训练语音来调整 UBM 参数以生成说话者适应的 GMM，GMM 则通过多个加权高斯概率密度来拟合语音空间分布的概率密度。概率线性判别分析(Probabilistic Linear Discriminant Analysis，PLDA)[6]结合 i-vector[7]的方法考虑了信道因素对识别结果的影响，提升了系统的鲁棒性。

如今，深度学习受到越来越多的关注[8-10]。神经网络也被广泛应用于声纹识别领域，其主要分为两大类：以文献[11-12]为代表的端到端的识别模型和文献[13-14]所研究的两阶段识别方法。文献[15-16]使用卷积神经网络(CNN)对语音特征进行端到端的识别，整条语音使用语谱图作为特征输入，CNN 相较于全连接神经网络(Deep Neural Network，DNN)性能表现更佳。文献[17]使用 DNN 作为特征提取器，由 Fbank 特征得到更高级的特征表示。神经网络将帧级特征映射到对应的说话者身份目标，取训练完成的识别模型的最后一层隐层的输出为 d-vector 特征。后验分类模型再根据 d-vector 特征得到识别结果。文献[18]使用 CNN 替代 DNN 网络来完成这一任务，后验模型采用余弦距离(Cosine Distance)进行分类。该模型取得了比 DNN 更好的实验结果。然而，上述两类方法都存在些许不足。端到端的方法多采用话语级的特征输入，但由于各语音之间存在时长差异，若要输入为等大小的特征，就要将输入剪裁或补零为同等尺寸。这会造成数据失真，进而影响模型的识别效果。而采用帧级别的两阶段方法使用帧级特征作为输入，虽然解决了输入大小不一致的问题，但该方法过于繁琐。两阶段的方法只是利用神经网络来进行 embedding 的学习，得到高维的特征表示，并不能够直接得到识别结果。

基于上述研究背景，本文提出了基于帧级特征的端到端识别方法。该方法采用 CNN 网络作为识别模型，端到端的识别方法一次性完成了特征训练和分类打分两个步骤，极大地简化了传统模型的复杂性。该方法利用帧级特征有监督地训练 CNN 模型。模型训练完毕后，测试帧级语音的预测标签，每句话的预测标签基于语音中的多帧预测值来计算，计算方法为最多预测类别法和预测值平均法。最终，得到整句话的预测得分和预测标签。实验表明，本文构建的基于 CNN 的端到端方法的识别性能优于已有的识别方法。

2 卷积神经网络的基本原理

在 CNN 网络中，模型由 3 种网络层组成：卷积层、池化层和全连接层。卷积是一种线性、平移不变性的运算，由输入信号上执行局部加权的组合构成。池化层为输入的位置和尺寸的改变带来一定程度的不变性。全连接层的所有神经元与上一层的所有神经元连接。卷积层具体可表示为：

$$h_{jk}^{l}=\sum_{j}\sum_{k}\omega_{jk}^{l}\cdot x_{jk}^{l-1}+b^{l} \tag{1}$$

$$x_{jk}^{l}=\sigma(h_{jk}^{l}) \tag{2}$$

其中，ω_{jk}^{l} 代表第 l 层中 (j,k) 区域的卷积核权重，b^{l} 代表第 l 层中添加到卷积核中的偏置，x_{jk}^{l-1} 表示第 $l-1$ 层中 (j,k) 区域的值，h_{jk}^{l} 表示第 l 层中第 j 个单元的输出值。同时，对卷积后的值进行非线性激活，采用修正线性单元(Rectified Linear Unit，Relu)[19]来激活，即为式(2)中的 $\sigma(x)$，可用下式表示：

$$\sigma(x)=\begin{cases}x, & x\geqslant 0\\ 0, & \text{others}\end{cases} \tag{3}$$

池化层可利用特征的局部相关性来减小特征的维度，常用的池化方法是平均池化和最大池化。这里，采用使用频率更高的最大池化，最大池化层生成的特征表示如下：

$$x_{mn}^{l+1}=\max(x_{ij}^{l}) \tag{4}$$

其中，x_{ij}^{l} 表示第 l 层中池化所覆盖的区域 i 和 j，x_{mn}^{l+1} 表示池化之后对应的输出。经过卷积、池化之后，特征会被展开成向量形式输入到全连接层中。最后，softmax 层激活得到各类的预测值，输入的预测标签取预测值最大的类别。softmax 的计算方法如下：

$$S_i=\frac{e^{i}}{\sum_{j}e^{j}} \tag{5}$$

$$y=\arg\max_{i}S_i \tag{6}$$

其中，i 表示第 i 类别，j 表示预测的总类别数，S_i 为第 i 个类别的预测得分。函数采用元素的指数形式来计算得分。CNN 模型根据计算得分得到预测结果 y，取预测值最大的类别作为结果类标。

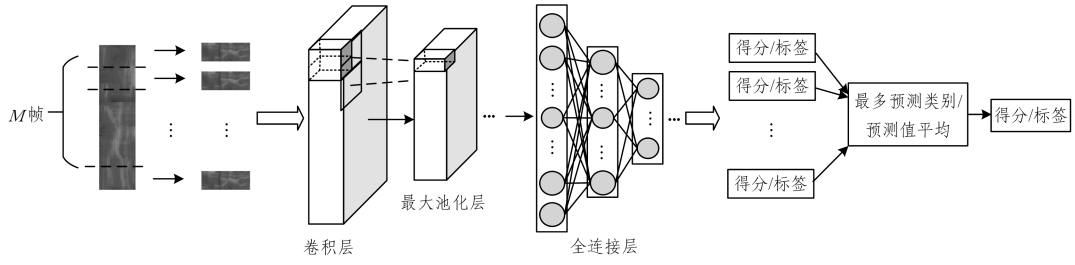

图 1 端到端识别方法的总体框架图

Fig. 1 Overall framework of end-to-end identification method

3　基于帧级特征的端到端说话人识别

本文提出的基于帧级特征的端到端说话人识别方法对现有两类方法的不足进行了改进。其输入采用帧级特征的形式，识别模型又是端到端的一体化方法。模型的整体框架如图 1 所示，CNN 网络对每帧特征得到相应的得分和标签。由于神经网络可以更好地利用特征间的相关性[17,20]，因此采用 Fbank 特征作为输入。最后，根据最多预测类别和预测值平均的方法得到整句话的预测结果。

3.1　帧级特征提取

每段语音的长度是不一致的，分帧可将每段语音划分为同等大小的帧级数据，只是每段语音的帧的数量会有所不同。帧级语音采用 Fbank 特征参数，神经网络可充分利用 Fbank 特征间的相关性。特征的具体提取过程如图 2 所示。

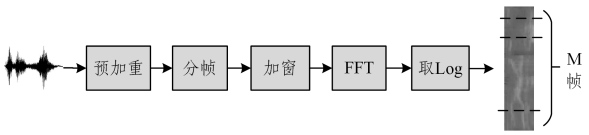

图 2　帧级特征的提取原理图

Fig. 2　Schematic diagram of frame level feature extraction

首先对语音进行预加重以加强信号的高频部分：

$$H(x(l))=1-\mu\cdot x(l)^{-1} \tag{7}$$

其中，μ 表示预加重系数，通常取值 0.97；然后利用信号的短时平滑的特性进行分帧处理，帧大小为 t，帧滑大小为 s。那么，一段采样率为 f kHz、信号长度为 l 的语音数据可以用一个 $M*N$ 大小的矩阵 $H=(H_1,H_2,\cdots,H_M)^{\mathrm{T}}$ 来表示，计算方法如下：

$$M=\left(\frac{l-t}{s}\right)+1 \tag{8}$$

$$N=f*t \tag{9}$$

其中，M 表示语音中的帧数，N 表示帧的长度。分帧之后做加窗处理，采用汉明窗以平滑每帧信号的边缘：

$$\omega[H_i]=0.54-0.46*\cos\left(\frac{2H_i\pi}{n}\right) \tag{10}$$

接着，FFT 处理 $\omega[H]$ 信号得到频域信号：

$$P_i=\frac{|FFT(\omega[H_i])|^2}{N_{size}} \tag{11}$$

其中，N_{size} 表示 FFT 分析的长度，$\omega[H_i]$ 表示第 i 帧信号。最终，对语音取对数操作：

$$x_i=\log P_i \tag{12}$$

即，语音可被提取成特征 $x=(x_1,x_2,\cdots,x_M)^{\mathrm{T}}$，取帧级特征作为识别模型的输入。

3.2　基于 CNN 的端到端识别模型

如上文所述，本文采用基于 CNN 的端到端的识别模型。网络具体架构如表 1 所列，模型的输入为 26×1 大小的帧级 Fbank 特征，提取方法如 3.1 节所述。卷积层采用 1×1 尺寸的卷积核，核的数量为 128；池化层采用最大池化，大小为 1×2；之后接入四层全连接层，每层的节点数逐层递减，分别为 1024，512，256 和 128；最后一层 softmax 用于预测分类，节点数为 18，对应最终预测的说话人类别数。

表 1　CNN 识别模型的具体网络结构

Table 1　Specific network structure of CNN identification model

网络层	参数设置
输入层	26×1
卷积层	1×1，128
池化层	1×2
全连接层 1	1024
全连接层 2	512
全连接层 3	256
全连接层 4	128
softmax 层	18

具体地，现有说话人 i 的 n 句话。那么第 i 个说话人的语音集合为 $S_i=\{u_1,u_2,\cdots,u_n\}$。由于模型的输入为帧级特征，则第句话的集合为 $u_j=\{x_1,x_2,\cdots,x_m\}$，表示语音数据的帧特征。端到端识别模型的预测类别计算方式分别采用最多预测类别和预测值平均。最多预测类别的计算方式为：

$$y=\underset{i}{\mathrm{mode}}\ y_i \tag{13}$$

其中，y_i 为一段语音中第 i 帧的预测类别，mode 函数取各帧中预测类别的众数。预测值平均的计算方式为：

$$y=\arg\max_i\left[\frac{1}{m}\left(\sum_j S_{ij}\right)\right] \tag{14}$$

其中，S_{ij} 为第 j 帧对第 i 类别的预测得分。由于语音中有 m 帧数据，则 $\frac{1}{m}\left(\sum_j S_{ij}\right)$ 表示该句话对第类别的预测得分，最终取预测值最大的类别为预测标签。

4　实验及结果分析

4.1　数据集介绍

实验数据集选用普通话情感语音语料库[21]（Mandarin Affective Speech Corpus，MASC）来训练和评估所提出的模型。MASC 已被语言数据联盟（Linguistic Data Consortium，LDC）所收录，目录编号为 LDC2007S09[22]。该数据库的数据质量优良，音频文件多为在实验室等较为安静的环境下录制，避免了语音数据包含太多背景噪声的问题。语音数据以 8kHz 的速率采样。帧长度和移位长度分别设置为 256 和 128。大小为 128 的汉明窗用于 FFT 分析。语料库共包括 68 名说话人（45 名男性和 23 名女性）的语音，包括中立、愤怒、兴高采烈、恐慌和悲伤 5 种情绪状态。每个说话人对于每种情感有 60 句话语（每个句子大约 2s）。由于选用了 PLDA 做对比实验，在训练说话人无关的通用背景模型 UBM 时也需要训练语料，因此我们对 18 个说话人做分类任务。具体地，前 50 个人的所有语音用于训练 UBM 模型，后 18 个人的所有语音用于模型分类任务。每人每种情感前 70% 的数据用于训练模型，剩余的后 30% 用于测试。即训练语音有 18×5×60×70%=3780 句，测试语音有 18×5×60×30%=1620 句，且每句语音中帧的数量与语音的具体时长相关。

4.2　实验设置

端到端的识别模型结构详见 3.2 节。Fbank 帧级特征被转置成（26，1）大小的向量形式。CNN 模型中卷积层的卷积步幅设置为 1，模型中除输入、输出（含 softmax）层以外的网络层使用 Relu[19] 激活函数。为了防止过拟合，我们还在模型中添加了批量标准化层（batch normalization）[23] 和丢失层（dropout）[24]。模型的损失函数采用交叉熵来计算，使用初始

学习率为 0.001 的 Adam 优化器[25]来优化模型损失函数。由于采用帧级特征作为输入，训练的数据量呈指数级增加，数据的维度也大幅下降，因此训练过程中最小批量的大小设置为 12800，总共执行 200 轮迭代。本文的实验环境为 Windows64 位操作系统，CNN 网络模型训练采用 Keras 深度学习框架。

此外，选取 PLDA、CNN(utterance-level)话语级特征的端到端模型和 CNN-Cosine 两阶段模型作为本文的基线实验。PLDA 模型训练采用 1024 阶的高斯混合模型，PLDA 的输入为 i-vector 特征，i-vector 是根据训练好的 UBM 模型再执行 5 次 EM 算法迭代，最终得到固定 400 维度的特征向量。PLDA 模型训练采用微软的 Matlab 说话人识别工具包[26]。CNN(utterance-level)模型使用了语谱图特征，为了方便模型的训练，语谱图被统一固定为 128×128 的大小，并且做均值-方差归一化处理。CNN-Cosine 两阶段模型的设置与本文所提方法相同，在计算余弦距离时对该说话人的得分进行 L2 得分归一化处理。

4.3 结果与分析

本文采用准确率(Accuracy，ACC)和等错误率(Equal Error Rate，EER)作为系统性能的评价指标。ACC 值是指预测正确的样本数占所有预测样本总数的比例，其可以评价样本分类是否正确。EER 值是错误拒绝率(False Rejection Rate，FRR)和错误接受率(False Acceptance Rate，FAR)相等时的错误率。相对于 ACC 值，EER 可衡量样本具体是如何被错分的，即目标说话人样本是被分成了非目标说话人样本，还是非目标说话人样本被错分了。

根据之前对 CNN 模型参数的设置训练得到模型损失值与训练轮数的关系。损失值 loss 可以衡量模型训练的好坏，便于观察以调整训练网络模型。loss 值越小，模型预测越准确，对应的准确率越高、等错误率越低。如图 3 所示，带方格的线为训练集曲线，不带方格的线为验证集曲线。验证集的 loss 值高于测试集是因为模型对测试集的拟合程度不如对训练集的拟合程度。整体上，随着训练轮数的增加，loss 曲线呈下降的趋势，说明模型正训练得越来越好。训练初期曲线下降得很快，在训练到 100 轮时 loss 趋于平缓。而在 150 轮时，验证集的 loss 基本维持稳定，测试集中 loss 减少的程度也非常小，这说明神经网络模型已基本训练完毕。loss 曲线图有助于观察模型的训练情况及选择合适的训练参数，如果要节省训练时间和降低训练成本，本实验可以将训练轮数减少到 150，这既可以节省模型训练的开支，同时对模型的性能也不会有很大的影响。

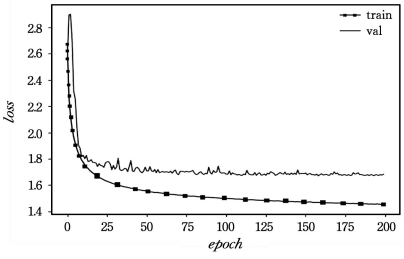

图 3 模型训练过程中损失值的曲线图

Fig. 3 Diagram of loss values during model training

为了得到模型的 EER 值，本文绘制了能体现分类系统误码率的检测错误权衡(Detection Error Tradeoff，DET)曲线图。在 DET 图中，FRR 和 FAR 会随着判断阈值的变化而变化，而模型的 EER 值即为曲线上横纵坐标相等时点的取值。如图 4 所示，实粗线为所提模型的 DET 曲线，虚线、虚点线、点横线分别为 PLDA、CNN 话语级特征和 CNN-Cosine 模型。可以看到，所提 CNN 模型整体都是处于 3 个基线模型的下侧。换言之，所提方法的 FRR 和 FAR 都低于其他方法，这体现了所提模型的优势。

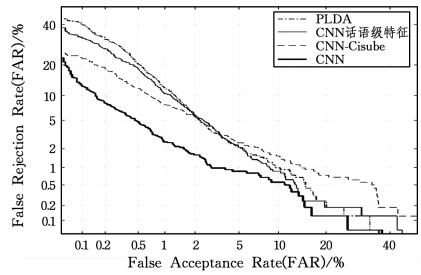

图 4 不同方法的 DET 图

Fig. 4 DET diagram of different methods

最后，表 2 列出了各个模型在 MASC 语音数据库的实验结果。其中，PLDA、CNN 话语级特征和 CNN-Cosine 两阶段模型为基线模型，最后两行 CNN 为所提模型。根据模型预测标签计算的不同，CNN 模型又分为最多预测类别和预测值平均，由于 EER 值是根据模型最终的预测得分来计算的，因此两种计算方法所得到的值是一致的。表中 ACC 值这一列，CNN 模型都优于基线模型，分别达到了 96.36%和96.67%，两者都高于基线模型的 95.74%、95.12%和 96.05%。同时可以发现，采用预测值平均的计算方法得到的模型准确率更高，即这一计算方法更为合理。在 EER 值方面，CNN 模型取得了 1.73%的识别表现，该值相较于 PLDA 的 3.28%性能优化了 47%，相比话语级特征 CNN 的 3.40%性能优化了 49%，相比两阶段 CNN-Cosine 的 3.35%性能优化了 48%，因此所提方法的性能得到了极大提升。

表 2 各个模型的准确率和等错误率

Table 2 Accuracy and equal error rate of each model

(单位：%)

MODEL	ACC	EER
PLDA	95.74	3.28
CNN(话语级特征)	95.12	3.40
CNN-Cosine(两阶段)	96.05	3.35
CNN(最多预测类别)	96.36	**1.73**
CNN(预测值平均)	**96.67**	**1.73**

结束语 本文针对现有的特征训练加分类得分的两阶段识别方法过于复杂的情况，提出了端到端的 CNN 识别模型，该方法简化了传统识别方法的步骤，将特征训练和模型识别一体化。同时，神经网络强大的数据拟合能力也提升了识别效果。另外，本文对针对不同计算方法得到预测标签的 CNN 模型进行了讨论，发现基于平均值预测的方式是更加合理的。所提方法在 MASC 数据集上的结果也验证了其优越的性能。但是，本文所提方法仍存在些许不足，即没有考虑语音数据的时序性问题。未来，我们将尝试使用不同的网络模型，如长短

时记忆网络，并通过加深、优化网络结构来解决上述问题，进一步提升识别的性能。

参 考 文 献

[1] HANSEN J H L, HASAN T. Speaker Recognition by Machines and Humans: A tutorial review [J]. IEEE Signal Processing Magazine, 2015, 32(6): 74-99.

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

[3] 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.

[4] 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.

[5] 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.

[6] MACHLICA, LUKÁ Š, ZAJÍC, et al. An Efficient Implementation of Probabilistic Linear Discriminant Analysis [C]// IEEE International Conference on Acoustics. IEEE, 2013.

[7] DEHAK N, KENNY P J, DEHAK R, et al. Front-End Factor Analysis for Speaker Verification [J]. IEEE Transactions on Audio, Speech and Language Processing, 2011, 19(4): 788-798.

[8] WANG H L, QI X L, WU G S. Research Progress of Object Detection Technology Baseon Convolutional Neural Network in Deep Learning[J]. Computer Science, 2018, 45(9): 11-19.

[9] ZHU J Y, PARK T, ISOLA P, et al. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks [C]// International Conference on Computer Vision (ICCV). 2017: 2242-2251.

[10] LIU J, JIN Z Q. Facial Expression Transfer Method Based on Deep Learning[J]. Computer Science, 2019, 46(S1): 250-253.

[11] JI R F, CAI X Y, BO X. An End-to-End Text-Independent Speaker Identification System on ShortUtterances. [C]// Annual Conference of the International Speech Communication Association(INTERSPEECH). 2018: 3628-3632.

[12] LUKIC Y, VOGT C, DURR O, et al. Speaker identification and clustering using convolutional neural networks[C]// 2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP). IEEE, 2016.

[13] LI N , TUO D Y, SU D, et al. Deep Discriminative Embeddings for Duration Robust Speaker Verification[C] // Conference of the International Speech Communication Association. 2018.

[14] TORFI A, DAWSON J, NASRABADI N M. Text-Independent Speaker Verification Using 3D Convolutional Neural Networks [C]// IEEE International Conference on Multimedia and Expo. 2018: 1-6.

[15] NAGRANI A, CHUNG J S, ZISSERMAN A. VoxCeleb: a large-scale speaker identification dataset[C]// Conference of the International Speech Communication Association. 2017.

[16] HRŮZ M, ZAJÍC Z. Convolutional Neural Network for speaker change detection in telephone speaker diarization system[C]// IEEE International Conference on Acoustics. IEEE, 2017.

[17] VARIANI E, LEI X, MCDERMOTT E, et al. Deep neural networks for small footprint text-dependent speaker verification [C]// 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2014.

[18] YU-HSIN C, MORENO I L, TARA N S, et al. Locally-Connected and Convolutional Neural Networks for Small Footprint Speaker Recognition [C] // Conference of the International Speech Communication Association. 2015.

[19] HINTON G E. Rectified Linear Units Improve Restricted Boltzmann Machines Vinod Nair [C]// International Conference on International Conference on Machine Learning. Omnipress, 2010.

[20] YUAN L, YAN M Q, NAN X C, et al. Deep feature for text-dependent speaker verification[J]. Speech Communication, 2015, 73: 1-13.

[21] WU T, YANG Y, WU Z, et al. MASC: A Speech Corpus in Mandarin for Emotion Analysis and Affective Speaker Recognition [C]// Speaker & Language Recognition Workshop. IEEE, 2006.

[22] YANG Y C, WU Z H, WU T, et al. Mandarin Affective Speech LDC2007S09. [EB/OL]. https://catalog. ldc. upenn. edu/LDC2007S09.

[23] IOFFE S, SZEGEDY C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift [C]// 2015 International Conference on Machine Learning.

[24] SRIVASTAVA N, HINTON G, KRIZHEVSKY A, et al. Dropout: A Simple Way to Prevent Neural Networks from Overfitting [J]. Journal of Machine Learning Research, 2014, 15(1): 1929-1958.

[25] KINGMA D P, BA J. Adam: A Method for Stochastic Optimization[C]// 2015 International Conference on Learning Representations(Poster). 2015.

[26] SADJADI S O, SLANEY M, HECK A L. MSR Identity Toolbox v1. 0: A MATLAB Toolbox for Speaker Recognition Research [EB/OL]. https://www. microsoft. com/en-us/research/publication/msr-identity-toolbox-v1-0-a-matlab-toolbox-for-speaker-recognition-research-2.

HUA Ming, born in 1995, postgraduate, is a member of China Computer Federation. His main research interests include speaker recognition and deep learning.

LI Dong-dong, born in 1981, Ph.D, associate professor. Her main research interests include speech processing and affective computing.