doi:10.11835/j.issn.1000-582X.2023.01.012

基于知识蒸馏与 ResNet 的声纹识别

荣玉军[1],方昳凡[2],田 鹏[2],程家伟[2]
(1.中移(杭州)信息技术有限公司,杭州 310000;2.重庆邮电大学 自动化学院,重庆 400065)

摘要:针对声纹识别领域中存在信道失配与对短语音或噪声条件下声纹特征获取不完全的问题,提出一种将传统方法与深度学习相结合,以 I-Vector 模型作为教师模型对学生模型 ResNet 进行知识蒸馏。构建基于度量学习的 ResNet 网络,引入注意力统计池化层,捕获并强调声纹特征的重要信息,提高声纹特征的可区分性。设计联合训练损失函数,将均方根误差(MSE,mean square error)与基于度量学习的损失相结合,降低计算复杂度,增强模型学习能力。最后,利用训练完成的模型进行声纹识别测试,并与多种深度学习方法下的声纹识别模型比较,等错误率(EER,equal error rate)至少降低了 8%,等错误率达到了 3.229%,表明该模型能够更有效地进行声纹识别。

关键词:深度学习;知识蒸馏;声纹识别;说话人识别
中图分类号:TP751 **文献标志码**:A **文章编号**:1000-582X(2023)01-113-012

Voiceprint recognition based on knowledge distillation and ResNet

RONG Yujun[1], *FANG Yifan*[2], *TIAN Peng*[2], *CHENG Jiawei*[2]
(1.China Mobile Hangzhou Information Technology Co. Ltd., Hangzhou 310000, P. R. China;
2.Chongqing University Posts & Telecommunication, College Automation, Chongqing 400065, P. R. China)

Abstract: Aiming at the problem of channel mismatch in the field of voiceprint recognition and incomplete acquisition of voiceprint features under short speech or noise conditions, a method that combines traditional methods with deep learning is proposed, and the ResNet model is used as the student model to perform knowledge distillation on the I-Vector model as the teacher model. We construct a ResNet network based on metric learning, introduce an attentive statistics pooling layer, capture and emphasize the important information of voiceprint features, and improve the distinguishability of voiceprint features. The mean square error (MSE) is combined with the loss based on metric learning to reduce computational complexity and enhance model learning capabilities. Finally, the trained model is used for voiceprint recognition test, and compared with the voiceprint recognition model under a variety of deep learning methods. It's found that the equal error rate (EER) is reduced by at least 8%, and the equal error rate has reached 3.229%, indicating that the model can perform speaker verification more effectively.

Keywords: deep learning; knowledge distillation; voiceprint recognition; speaker verification

收稿日期:2021-07-12 **网络出版日期**:2022-11-08
基金项目:教育部-中国移动科研基金资助项目(MCM20180404);国家自然科学基金(52272388)。
Supported by Ministry of Education China Mobile Research Fund (MCM20180404), and the National Natural Science Foundation of China (52272388).
作者简介:荣玉军(1976—),男,高级工程师,主要从事人工智能、数字家庭、物联网方向研究,(E-mail) rongyujun@cmhi.chinamobile.com。
通信作者:方昳凡,女,硕士研究生,主要从事深度学习、声纹识别及说话人分离方向研究,(E-mail)625450568@qq.com。

随着物联网、智能设备、语音助手、智能家居和类人机器人技术的发展，以及人们对安全的日益重视，生物识别技术的应用越来越多，包括脸部、视网膜、声音和虹膜等识别技术[1]。其中声纹识别因易于实现，使用成本低而被用户广泛接受。声音是一种生物行为特征，它传递一个人特征相关的信息，比如说话人的种族、年龄、性别和感觉。说话人识别是指根据人的声音识别人的身份[2]。研究表明，声音因其独特的特征可以用来区分不同人的身份[3]，除了虹膜、指纹和人脸外，语音提供了更高级别的安全性，是一种更加有效的生物识别技术。

说话人识别可分为说话人确认和说话人辨认2个任务，说话人确认是实现智能交互的关键技术，可广泛应用于金融支付、刑事侦查、国防等领域[4]。说话人确认是一对一的认证，其中一个说话者的声音与一个特定的特征匹配，可以分为文本依赖型和文本独立型[5]。与文本相关的说话人确认系统要求从固定的或提示的文本短语产生语音，利用说话人语音的尺度不变性、特征不变性和文本相关不变性等特性，对说话人语音进行识别[6]，而与文本无关的说话人确认系统操作的是无约束语音，是一个更具有挑战性与实用性的问题。

Reynolds[7]等人在2000年提出了高斯混合模型–通用背景模型(GMM-UBM)，它能够利用大量的数据进行计算，但是具有较高的计算复杂度，不能很好地区分多个说话人，表示较为复杂。Kenny[8]等人在2007年提出联合因子分析法(JFA)，它将GMM-UBM得到的均值超矢量所在的空间进一步建模，可以补偿信道变化带来的影响，但因其模型复杂度过高而不便应用。为了对JFA进行简化，提出了利用身份向量I-Vector[9]的方法，因为它不对信道进行单独建模，所以要通过线性判别分析(LDA, linear discriminant analysis)，概率线性判别分析(PLDA, probabilistic linear discriminant analysis)等方法对其进行信道补偿。

当前，深度学习方法广泛应用于语音识别[10]、计算机视觉领域，并逐渐应用于说话人识别等其他领域，均取得了显著成效。2014年，Google的d-vector使用神经网络隐层输出替代I-Vector，虽然实验效果不如I-Vector，但证明了神经网络方法的有效性。循环神经网络(RNN, recurrent neural network)在语音识别方面效果良好，在处理变长序列方面具有明显优势，现在也被应用在说话人识别任务中。2017年，Snyder[11]等人使用时延神经网络提取帧级特征，语句级特征则从统计池化层聚合而来，利用PLDA进行后端打分，处理短语音的效果优于I-Vector，在它基础上加入离线数据增强后效果整体超过了I-Vector，成为新的基准模型。目前，有方法将图像领域的卷积神经网络用于说话人识别语音信号的预处理以提高说话人识别率[12]，VGG结构网络[13]，深度残差网络[14]等卷积神经网络结构也被用于处理说话人识别任务。尽管深度学习的应用使得说话人识别技术有了巨大进步，但目前还存在以下问题：1)对于短语2 s的短时语音识别性能差；2)缺乏对信道多变性的补偿能力；3)对于噪声条件适应性不足，鲁棒性差。

笔者提出一种采用知识蒸馏技术，将传统的I-Vector方法与深度学习相结合的方法，进行与文本无关的说话人确认。设计所使用的ResNet网络模型结构，并对模型进行训练。比较不同打分后端下基准模型和增加数据增强后的实验结果，以及采用知识蒸馏的I-Vector模型与ResNet网络相结合后的实验结果，并对实验结果进行了讨论。

1 基于知识蒸馏与ResNet的声纹识别

笔者设计的声纹识别模型方法分为4步，如图1所示。1)对输入语音进行预处理；2)对输入语音提取I-Vector，采用训练好的I-Vector模型作为教师模型，ResNet为学生模型；3)将I-Vector模型与ResNet模型联合训练；4)利用PLDA方法或者余弦方法进行打分。

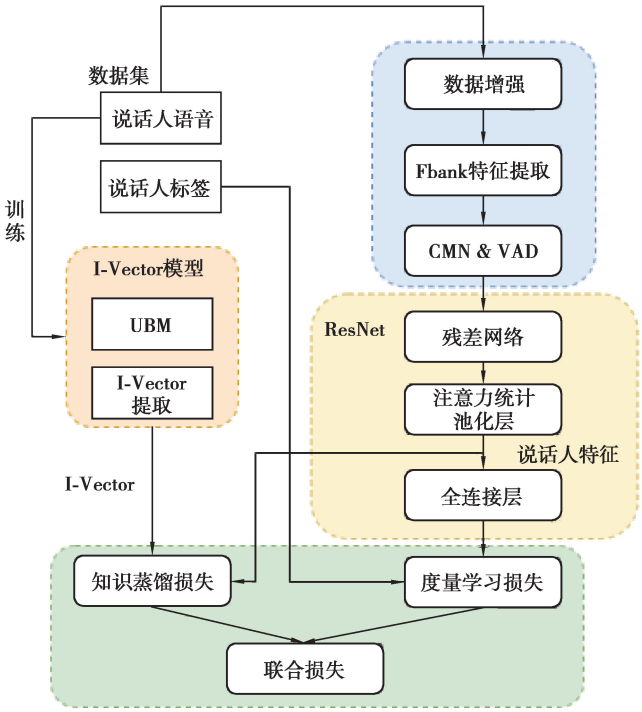

图 1　基于知识蒸馏与 ResNet 的声纹识别模型框架图

Fig. 1　Frame diagram of voiceprint recognition model based on knowledge distillation and ResNet

1.1　语音预处理

原始语音的预处理主要进行的步骤为：

1）提取语音数据的 Fbank（Filter Bank）特征。

2）对语音数据进行增强，包括使用噪声数据集与原始数据集叠加合频谱增强方法。

1.1.1　特征提取

Fbank 是频域特征，能更好反映语音信号的特性，由于使用了梅尔频率分布的三角滤波器组，能够模拟人耳的听觉响应特点。Fbank 特征的提取步骤为：

1）首先使用一阶高通滤波器应用于原始信号，进行信号预加重，达到提高信号的信噪比与平衡频谱的目的，

$$y(n)=s(n)-\alpha s(n-1), \tag{1}$$

式中：$y(n)$是预加重后的信号；$s(n)$为原始语音信号；α 是预加重系数，其典型值为 0.95 或 0.97。

2）将预加重后的语音分割成多个短时帧，每一帧之间具有部分重叠。接着利用 Hamming 窗为每一个短时帧进行加窗操作，防止离散傅里叶变化过程中产生频谱泄露。

$$w(n)=0.54-0.46\cos\left(\frac{2\pi n}{N-1}\right), \tag{2}$$

式中：$w(n)$即为窗函数；N 为该帧的采样点数量。

3）对加窗后的语音信号进行离散傅里叶变换

$$S(k)=\sum_{n=0}^{N-1}s(n)\exp\left\{-2\pi i\frac{nk}{N}\right\},k=0,1,\cdots,N-1, \tag{3}$$

其中：$s(n)$为一个短时帧语音信号；$S(k)$是其频率响应。根据频率响应计算功率频谱 $P(k)$

$$P(k)=\frac{|S(k)|^2}{N},\ k=0,1,\cdots,N-1, \tag{4}$$

4）最后使用梅尔频率均匀分布的三角滤波器对功率谱图进行滤波，得到了 Fbank 特征，该特征模拟了非

线性的人耳的听觉响应。可以通过下列计算公式进行频率 f 和梅尔频率 m 之间进行转换

$$m = 2\ 595 \log_{10}\left(1+\frac{f}{700}\right)\ , \tag{5}$$

$$f = 700(10^{m/2\ 595}-1)。 \tag{6}$$

通过上述过程即可得到作为网络输入的 Fbank 特征。

5)Fbank 特征是频域特征，而通道噪声是卷积性的，其在频域为加性噪声，因此利用倒谱均值归一化方法(CMVN，cepstral mean and variance normalization)抑制该噪声。之后利用语音活动检测(VAD，voice activity detection)基于语音帧能量移除语音数据在中的静音段。

1.1.2 数据增强

在深度学习中，数据集大小决定了模型能够学习到内容的丰富性，对模型的泛化性与鲁棒性起着关键性作用。在研究中将采用 2 种数据增强方式：1)离线数据增强；2)在线数据增强。

离线数据增强方法在 X-Vector 模型中[18]最先使用，该方法采用 2 个额外的噪声数据集 MUSAN 和 simulated RIRs[15]，其中 MUSAN 数据集由大约 109 h 的音频组成，包括 3 种类型：1)语音数据，包括可公开获得的听证会、辩论等录音；2)音乐数据，包括爵士、说唱等不同风格音乐；3)噪声数据，包括汽车声、雷声等噪声。simulated RIRs 数据集包括具有各种房间配置的模拟房间脉冲响应。离线数据增强的具体方式为，随机采用下列的一种方式进行增强：

1)语音叠加：从 MUSAN 数据集的语音数据中随机选取 3—7 条语音，将其叠加后，再以 13～20 dB 的信噪比与原始语音相加。

2)音乐叠加：从 MUSAN 数据集的音乐数据中随机选取一条音频，将其变换至原始语音长度，再以 5～15 dB 的信噪比与原始语音相加。

3)噪声叠加：从 MUSAN 数据集的噪声数据中随机选取一条音频，以 1 s 为间隔，0～15 dB 的信噪比与原始语音相加。

4)混响叠加：从 simulated RIRs 数据集中随机选取一条音频，与原始信号进行卷积。

在线语音增强方法是一种直接作用于频谱图上的方法，可以在网络接收输入后直接计算。本研究中主要采用在线语音增强的其中 2 种方式：频率掩膜与时间掩膜方式。如图 2 所示，以下设置原始语音特征的频谱特征为$\mathbf{S}\in R^{F\times T}$：

1)频率掩膜：特征的频率维度是 F，设置频率掩膜区间长度为 f，f 为可调参数。然后从区间$[0,F-f]$中任意选取掩膜区间的开始位置 f_0，最后对语音特征 $\mathbf{S}$ 的$[f_0,f_0+f]$区间进行掩膜操作，即将其区间内的值设置为 0，该操作可重复多次。

2)时间掩膜：特征的时间维度是 T，设置频率掩膜区间长度为 t，t 为可调参数。然后从区间$[0,T-t]$中任意选取掩膜区间的开始位置 t_0，最后对语音特征 $\mathbf{S}$ 的$[t_0,t_0+t]$区间进行掩膜操作，即将其区间内的值设置为 0，该操作可重复多次。

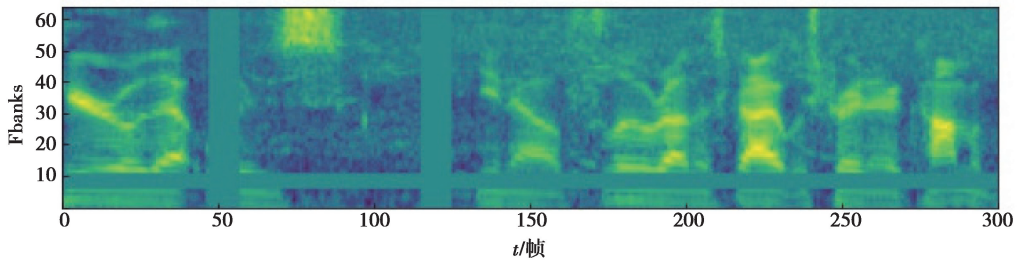

图 2 频谱增强后的 Fbank 特征

Fig 2 Fbank features after spectrum enhancement

1.2 I-Vector 提取

假设一帧语音特征的大小为 F，即 Fbank 特则维度为 F，将数据集中第 i 条语音的特征表示为$\boldsymbol{O}_i=$

$(\boldsymbol{o}_{i1}\cdots\boldsymbol{o}_{i2}\cdots\boldsymbol{o}_{iT_i})\in \mathrm{R}^{F\times T_i}$，式中：$T_i$ 表示该语音的帧数；$\boldsymbol{o}_{it}$ 表示第 t 帧的特征向量，$t=T_1,T_2,\cdots,T_i$。在 I-Vector框架中，假设每一帧的特征向量$\boldsymbol{o}_{it}$都由各自的高斯混合模型(GMM，gaussian mixture model)生成，同一条语音中的不同语音帧特征由同一模型独立分布生成，而每条语音所对应的高斯混合模型都由通用背景模型(UBM，universal background model)进行均值超向量平移操作得到

$$\boldsymbol{\mu}_i=\boldsymbol{\mu}^{(b)}+\boldsymbol{T}\boldsymbol{w}_i, \tag{7}$$

式中：$\boldsymbol{\mu}_i,\boldsymbol{\mu}^{(b)}\in \mathrm{R}^{CF}$，$\boldsymbol{\mu}_i$ 为第 i 条语音对应的 GMM 的均值超向量，$\boldsymbol{\mu}^{(b)}$ 为 UBM 的均值超向量；C 为 GMM 模型分量的数目；$\boldsymbol{T}\in \mathrm{R}^{\mathrm{CF}\times \mathrm{D}}$为 I-Vector 提取器，定义了总变化矩阵；$\boldsymbol{w}_i\in \boldsymbol{R}^{\mathrm{D}}$ 为总变化空间内服从标准高斯分布的隐变量，对其进行 MAP 估计，即可得到 I-Vector。

因此，I-Vector 是一种概率分布模型，通常使用期望极大(EM，expectation-maximization)算法进行训练。接着使用 UBM 计算零阶统计量 N_{ic}和一阶统计量$\tilde{\boldsymbol{f}}_{ic}$

$$N_{ic}=\sum_{t=1}^{T_i}\gamma_c(\boldsymbol{o}_{it}),\ c=1,2,\cdots,C, \tag{8}$$

$$\tilde{\boldsymbol{f}}_{ic}=\sum_{t=1}^{T_i}\gamma_c(\boldsymbol{o}_{it})(\boldsymbol{o}_{it}-\boldsymbol{\mu}_c^{(b)}),\ c=1,2,\cdots,C, \tag{9}$$

式中，$\gamma_c(\boldsymbol{o}_{it})$为第 c 个高斯分量的后验概率

$$\gamma_c(\boldsymbol{o}_{it})=\frac{\lambda_c^{(b)}N(\boldsymbol{o}_{it}\mid\boldsymbol{\mu}_c^{(b)},\sum_c^{(b)})}{\sum_{j=1}^{C}\lambda_j^{(b)}N(\boldsymbol{o}_{it}\mid\boldsymbol{\mu}_j^{(b)},\sum_j^{(b)})},\ c=1,2,\cdots,C, \tag{10}$$

接着可通过以下计算得到 I-Vector

$$x_i=\boldsymbol{L}_i^{-1}\sum_{c=1}^{C}\boldsymbol{T}_c^{\mathrm{T}}(\sum\nolimits_c^{(b)})^{-1}\tilde{\boldsymbol{f}}_{ic}, \tag{11}$$

式中：$\boldsymbol{L}_i$ 为

$$\boldsymbol{L}_i=\boldsymbol{I}+\sum_{c=1}^{C}N_{ic}\boldsymbol{T}_c^{\mathrm{T}}(\sum\nolimits_c^{(b)})^{-1}\boldsymbol{T}_c。 \tag{12}$$

$\sum_c^{(b)}\in \boldsymbol{R}^{F\times F}$是 UBM 第 c 个分量的协方差矩阵；$\boldsymbol{T}_c^{\mathrm{T}}$ 是 $\boldsymbol{T}=[\boldsymbol{T}_1^{\mathrm{T}},\boldsymbol{T}_2^{\mathrm{T}},\cdots,\boldsymbol{T}_C^{\mathrm{T}}]^{\mathrm{T}}$ 的第 c 个分量。

估计总变化矩阵 $\boldsymbol{T}$ 的过程即为 I-Vector 模型的训练过程，使用 EM 算法训练模型，其中 E 步骤为

$$\langle\boldsymbol{w}_i\mid\boldsymbol{X}_i\rangle=\boldsymbol{u}_i, \tag{13}$$

$$\langle\boldsymbol{w}_i\boldsymbol{w}_i^{\mathrm{T}}\mid\boldsymbol{X}_i\rangle=\boldsymbol{L}_i^{-1}+\langle\boldsymbol{w}_i\mid\boldsymbol{X}_i\rangle\langle\boldsymbol{w}_i^{\mathrm{T}}\mid\boldsymbol{X}_i\rangle, \tag{14}$$

M 步骤为

$$\boldsymbol{T}_c=[\sum_{i=1}^{N}\tilde{\boldsymbol{f}}_{ic}\langle\boldsymbol{w}_i\mid\boldsymbol{X}_i\rangle^{\mathrm{T}}][\sum_{i=1}^{N}N_{ic}\langle\boldsymbol{w}_i\boldsymbol{w}_i^{\mathrm{T}}\mid\boldsymbol{X}_i\rangle]^{-1},\ c=1,2,\cdots,C, \tag{15}$$

式中，N 为数据集中包含的语音总数。利用 I-Vector 模型解决短时语音与信道失配问题，并学习到信息“教”给基于 ResNet 的声纹识别模型。

1.3　基于 ResNet 的声纹识别设计

笔者设计的基于 ResNet 声纹识别模型结构如表 1 所示。

表 1　ResNet 的具体结构

Table 1　The specific structure of ResNet

网络层	结构	输出大小
输入层	无	$(1,T,64)$
卷积层	$[3\times3,64,2]$	$(64,\frac{T}{2},32)$

续表 1

网络层	结构	输出大小
残差块 1	$\begin{bmatrix}3\times3,64,1\\3\times3,64,1\end{bmatrix}\times3$	$(64,\frac{T}{2},32)$
残差块 2	$\begin{bmatrix}3\times3,128,2\\3\times3,128,1\end{bmatrix}\times1$	$(128,\frac{T}{4},16)$
	$\begin{bmatrix}3\times3,128,1\\3\times3,128,1\end{bmatrix}\times2$	$(128,\frac{T}{4},16)$
残差块 3	$\begin{bmatrix}3\times3,256,2\\3\times3,256,1\end{bmatrix}\times1$	$(256,\frac{T}{8},8)$
	$\begin{bmatrix}3\times3,256,1\\3\times3,256,1\end{bmatrix}\times2$	$(256,\frac{T}{8},8)$
统计池化层	无	(4 096)
全连接层 1	[4 096,400]	(400)
批量归一化层	无	(400)
全连接层 2	[400,1 211]	(1 211)

在表 1 中,[·,·,·]表示卷积的卷积核的大小、通道数、卷积步长。输出大小中 T 为输入特征的帧数。在每一个残差块后都接有批量归一化层与 ReLU 激活函数。

从表 1 的结构中,网络输入是大小为$(1,T,64)$的张量,Fbank 特征维度为 64,网络结构中的卷积层、残差块 2、残差块 3 对输入采取了通道数翻倍、频率维度减半、时间维度减半的操作。全连接层 1 的输出即为提取的声纹特征,全连接层 2 为分类层,仅仅在训练过程中使用,1211 为训练数据集包含的人数。

统计池化层(statistics pooling)是声纹识别模型中所特有的结构,用来处理语音输入序列变长问题。卷积层也可以接收不定大小的输入,但对于不同大小的输入,其输出大小也会不同,但在声纹识别任务中,需要将不同大小的特征映射至固定维度大小的声纹特征。残差块将维度为 F_0 的 Fbank 特征变为形状为 $\boldsymbol{X}\in R^{C\times F\times T}$ 的多通道特征,其中:C 为通道数;F 和 T 对应于网络原始输入 $\boldsymbol{O}\in R^{1\times F_0\times T_0}$ 中的特征维度 F_0 和 T_0,$F<F_0$,$T<T_0$。该特征为局部特征,即帧级特征,且该特征是相邻帧共享。统计池化层的输入为卷积块的输出,研究采用如图 3 所示结构的注意力统计池化[16]将局部特征转化为全局特征,即语句级特征。

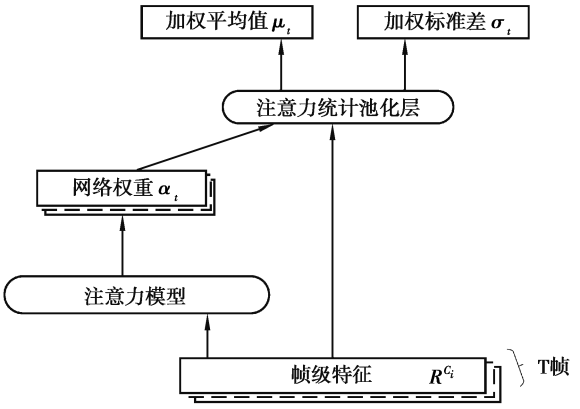

图 3 注意力统计池化层结构

Fig. 3 Attention statistics pooling layer structure

首先为每个帧级特征计算标量分数 score_t

$$\text{score}_t = \boldsymbol{v}^{\mathrm{T}} f(\boldsymbol{W}\boldsymbol{R}^{C_t} + \boldsymbol{b}) + k, \tag{16}$$

式中：$f()$表示非线性激活函数；$\boldsymbol{R}^{C_t}$为第 t 个通道的特征；$\boldsymbol{v}$，$\boldsymbol{W}$，$\boldsymbol{b}$，$\boldsymbol{k}$ 为要学习的参数，接着利用 softmax 函数在所有的帧上做归一化操作，得到归一化分数 α_t：

$$\alpha_t = \frac{e^{s_t}}{\sum_{\tau=1}^{L} e^{s_\tau}}, \tag{17}$$

将分数 α_t 作为每一个通道特征的权重分数，然后计算加权平均向量 $\tilde{\boldsymbol{\mu}}$

$$\tilde{\boldsymbol{\mu}} = \sum_{t=1}^{L} \alpha_t R^{C_t}, \tag{18}$$

通过这种方式，利用加权平均提取到的语句级特征将关注到信息量更丰富的帧，接着使用加权标准差将统计池化与注意力机制相结合

$$\tilde{\boldsymbol{\sigma}} = \sqrt{\sum_{t=1}^{L} \alpha_t R^{C_t} \odot \boldsymbol{R}^{C_t} - \tilde{\boldsymbol{\mu}} \odot \tilde{\boldsymbol{\mu}}}, \tag{19}$$

将加权平均值与加权标准差拼接后作为全连接层的输入，提高了语句级声纹特征的可区分性。将 ResNet 提取到的声纹特征记为 Embedding。

1.4　基于知识蒸馏的联合训练

1.4.1　后端打分

使用 2 种后端打分策略：PLDA 打分后端与余弦打分后端。

1）PLDA 打分计算过程：记 2 条语音的 I-Vector 分别为$\boldsymbol{u}_1$，$\boldsymbol{u}_2$，使用对数似然比进行打分

$$\text{score}(\boldsymbol{u}_1, \boldsymbol{u}_2) = \ln \frac{p(\boldsymbol{u}_1, \boldsymbol{u}_2 \mid H_1)}{p(\boldsymbol{u}_1 \mid H_0) p(\boldsymbol{u}_2 \mid H_0)} =$$

$$\ln N\left(\begin{bmatrix}\boldsymbol{u}_1 \\ \boldsymbol{u}_2\end{bmatrix} \middle| \begin{bmatrix}\boldsymbol{m} \\ \boldsymbol{m}\end{bmatrix}, \begin{bmatrix}\sum + \boldsymbol{V}\boldsymbol{V}^{\mathrm{T}} & \boldsymbol{V}\boldsymbol{V}^{\mathrm{T}} \\ \boldsymbol{V}\boldsymbol{V}^{\mathrm{T}} & \sum + \boldsymbol{V}\boldsymbol{V}^{\mathrm{T}}\end{bmatrix}\right) -$$

$$\ln N\left(\boldsymbol{u}_1 \mid \boldsymbol{m}, \sum + \boldsymbol{V}\boldsymbol{V}^{\mathrm{T}}\right) - \ln N\left(u_2 \mid m, \sum + \boldsymbol{V}\boldsymbol{V}^{\mathrm{T}}\right), \tag{20}$$

式中：$p(\boldsymbol{u}_1, \boldsymbol{u}_2 \mid H_1)$为 2 条语音来自同一说话人的似然函数，$p(\boldsymbol{u}_1 \mid H_0)$与 $p(\boldsymbol{u}_2 \mid H_0)$分别为$\boldsymbol{u}_1$ 和$\boldsymbol{u}_2$ 来自不同说话人的似然函数。

2）余弦打分后端计算过程：记 2 条语音的 Embedding 分别为$\boldsymbol{x}_1$，$\boldsymbol{x}_2$，使用 2 向量的余弦距离计算得分

$$\text{score} = \frac{\boldsymbol{x}_1^{\mathrm{T}} \boldsymbol{x}_2}{\| \boldsymbol{x}_1 \|_2 \| \boldsymbol{x}_2 \|_2}, \ (-1 \leqslant \text{score} \leqslant 1)。 \tag{21}$$

1.4.2　训练损失函数

研究使用的第一种损失函数是蒸馏损失，它是 I-Vector 与经 ResNet 提取到的 Embedding 之间的均方误差（MSE，mean squared error），将一个批次中的第 i 个样本的 I-Vector 记为$\boldsymbol{u}_i \in \boldsymbol{R}^D$，ResNet 提取到的 Embedding 记为$\boldsymbol{x}_i \in \boldsymbol{R}^D$，两者之间的损失由以下公式计算

$$L_d = \frac{1}{B} \sum_{i=1}^{B} \| \boldsymbol{x}_i - \boldsymbol{u}_i \|_2^2, \tag{22}$$

式中 B 表示一个批次的大小。通过优化这一损失，可以使 ResNet 提取到的 Embedding 向 I-Vector 学习，由于 I-Vector 服从高斯分布，与 PLDA 中的假设相符，所以使用这种损失可以提高以 PLDA 为打分后端时的声纹识别模型性能。

第二种损失为度量学习损失，采用的为 additive margin softmax（AM-Softmax）损失，它的计算过程为

$$L_m = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{e^{\alpha(s_j^{(i)} - m)}}{e^{\alpha(s_j^{(i)} - m)} + \sum_{k=1, k \neq j}^{N} e^{\alpha s_k^{(i)}}}, \tag{23}$$

式中：$s_j^{(i)}$ 是在当前批次中标签为 j 的第 i 个采样的相似度分数；α 为缩放因子用来使训练过程更加平滑；m 为附加距离，通过优化该损失函数能够使相同说话人的声纹特征之间的相似性增加，不同说话人的声纹特征

之间的相似性降低。该损失函数根据相似度评分计算，当使用余弦打分后端时，可以提高声纹识别模型性能。

最终，笔者提出一种将2种损失相结合的基于知识蒸馏的联合损失函数

$$L_{\text{combine}}=\gamma L_m+(1-\gamma)L_d, \tag{24}$$

式中，γ 为超参数，$0\leqslant\gamma\leqslant1$，可以控制2种损失之间的比例。

当单独使用蒸馏损失MSE时，相当于使用无监督训练，I-Vector直接与Embedding计算损失；单独使用度量学习损失AM-Softmax时，仅使用ResNet模型训练，未有监督训练，未利用I-Vector进行知识蒸馏；使用联合损失时，既使用了蒸馏损失MSE，也使用了度量学习损失AM-Softmax进行训练。3种损失都用来更新ResNet网络的权值。

2 实验与结果分析

笔者所提出的模型使用VoxCeleb1[17]公开数据集进行训练，该数据通过一套基于计算机视觉技术开发的全自动程序从开源视频网站中捕捉而得到，完全属于自然环境下的真实场景，说话人范围广泛，场景多样。其中包括一个验证集和一个测试集，分别用于模型的训练和测试，数据集的数量统计如表2所示，此外，从测试集中随机抽取了37 720对语句用于模型的验证。在拥有了原数据后，使用离线增强和在线增强方法对原数据进行数据增强，对比分析数据强化对实验结果的影响。数据增强策略中的离线增强方式增加的样本数量为100 000个，在训练之前加入数据集；另一种是在线增强方式，频率掩膜参数设置为10，重复次数为1次，时间掩膜参数为15，重复次数为2次，即参数 F，N_f，T，N_t 分别被设置为10，1，15和2。

表2 模型数据集统计

Table 2 The statistics of the data set of the model

数据集	验证集	测试集
说话人数量	1211	40
语句数量	148642	4874

2.1 基准模型

基准模型用于和本实验中所设计的模型进行性能比较，从而证明本实验中的模型有效性。实验使用的基准模型包括I-Vector、X-Vector，2种模型均使用Kaldi框架实现。I-Vector模型的语音特征使用24维的梅尔频率倒谱系数(MFCC，mel frequency cepstral coefficents)，经过了二阶差分处理、基于滑动窗口的CMN和VAD处理后为72维，所用的UBM模型具有2048个高斯分量，所得的I-Vector维度为400。X-Vector模型的语音特征使用30维MFCC，并经过了基于滑动窗口的CMN和VAD，网络结构为5层的TDNN，说话人特征为512维，其使用了离线增强方式的数据增强方法，增强的样本数量为100 000个，并在训练之前加入数据集，网络训练所用的损失函数为交叉熵损失。2种模型均采用PLDA打分后端，并在PLDA前，使用LDA将离线增强后的数据维度降至200维并进行了 $L2$ 归一化。

2.2 模型训练

本次实验使用基于PyTorch的深度学习框架构建了所需要的ResNet模型，使用单个NVIDIA Tesla P100显卡训练30个迭代。使用Kaldi框架提取64维的Fbank作为输入特征，并经过了基于滑动窗口的CMN和VAD。Cosine和PLDA打分被用于模型结果评估，其中均采用等错误率(EER，equal error rate)和最小检测代价功能(minDCF，minimum detection cost function)来衡量模型的性能，等错误率是指当决策阈值变化时，错误接受率(FAR，false acceptance rate)与错误拒绝率(FRR，false rejection rate)相等时FAR或FRR的值，检测代价是说话人识别中常用的一种性能评定方法，定义式为 $\text{DCF}=C_{fr}\times\text{FRR}\times P_{\text{target}}+C_{fa}\times\text{FAR}\times(1-\text{P}_{\text{target}})$，其中 C_{fr} 和 C_{fa} 为错误拒绝和错误接受的惩罚权重，取 $C_{fr}=C_{fa}=1$，P_{target} 为目标说话人在总人群中的比例，最小检测代价即阈值变化时，检测代价的最小值。2种指标的值越小，表明模型性能越

强。在 AM-Softmax 损失函数中，缩放因子 α 和附加距离 m 分别设置为 30 和 0.2。在训练中，输入被截断或填充为 3 s 的长度，以形成大小为 128 的小批量数据。使用初始学习率为 0.001 的 Adam 优化器，将验证集数据用于检验训练效果，当验证集上的结果没有得到改善时，将学习率衰减到之前的 1/2。

2.3　主要实验结果分析

如表 3 所示。其中 DCF(0.01)表示检测代价函数中的 p-target 参数为 0.01，基准模型都使用了交叉熵损失函数。

表 3　基准模型实验结果

Table 3　The experiment results of baseline model

模型名称	Cosine			PLDA		
	EER(%)	DCF(0.01)	DCF(0.001)	EER(%)	DCF(0.01)	DCF(0.001)
I-Vector 基准模型	13.88	0.676 7	0.797 9	5.519	0.478 6	0.669 8
X-Vector 基准模型	10.16	0.727 3	0.845 5	5.344	0.468 8	0.615 5
ResNet 基准模型	6.957	0.578 9	0.707 7	5.698	0.512 6	0.696 7
ResNet 基准＋在线增强	6.890	0.575 6	0.701 8	5.311	0.503 3	0.654 4
ResNet 基准＋离线增强	6.24	0.516 6	0.642 6	5.110	0.450 1	0.567 5

ResNet 基准模型的性能具有较好的稳定性，在余弦打分后端和 PLDA 打分后端下的结果相差不大，在余弦打分下的实验结果相较 X-Vector 和 I-Vector 模型有很大的提升。ResNet 基准模型采用在线增强后，PLDA 打分下的实验结果优于 ResNet 基准模型，而余弦打分下的实验结果变化不大。ResNet 基准模型采用离线增强后，在 2 种打分方式下，所有性能指标相较于 ResNet 基准模型都有了很大的提升，并且同样优于采用在线增强方式的 ResNet 基准模型。因此对于同样的数据处理流程，离线增强方式更为复杂，但能获得更好的结果，在线增强方式计算速度快，仍然值得尝试。

采用知识蒸馏技术的 ResNet 模型使用了 MSE 损失函数，基于度量学习的 ResNet 模型损失函数使用 AM-Softmax，从表 4 可以看出，针对知识蒸馏技术优化的损失方法，ResNet 模型结果明显优于 I-Vector 基准模型，因为 ResNet 是从 I-Vector 中提取和学习部分相关参数且得到了更好的结果，这证明了 ResNet 方法和蒸馏损失方法结合的有效性。采用基于度量学习的损失函数 AM-Softmax 得到的模型结果优于 I-Vector 基准模型和采用知识蒸馏技术得到的模型结果。因此，考虑采用联合训练的方式来提升实验效果。

表 4　三种模型的对比实验

Table 4　Comparative experiment of the three models

模型名称	损失函数	Cosine			PLDA		
		EER(%)	DCF(0.01)	DCF(0.001)	EER(%)	DCF(0.01)	DCF(0.001)
I-Vector 基准模型		13.88	0.675 6	0.789 9	5.525	0.480 2	0.673 2
ResNet 知识蒸馏	MSE	10.44	0.683 9	0.776 6	4.788	0.452 4	0.609 5
ResNet 知识蒸馏＋数据增强	MSE	10.26	0.688 9	0.791 1	4.716	0.452 2	0.583 4
ResNet 度量学习	AM-Softmax	5.410	0.476 8	0.689 0	4.512	0.480 3	0.633 4
ResNet 度量学习＋数据增强	AM-Softmax	5.086	0.470 6	0.699 8	4.055	0.438 6	0.598 9

在联合训练中，使用的模型都是 ResNet 模型，采用在线增强进行数据增强，从上表可看出，γ 分别取 0.2、0.1、0.05，γ 控制这 2 个损失之间的比例，通过减小 γ 来强调蒸馏损失。结合表 4 和表 5 可以看出，针对损失函数，使用 MSE 损失和 AM-Softmax 损失联合训练的方法能够很大程度的提升模型的结果，实验结果还表明，AM-Softmax 损失有助于提高模型在 Cosine 打分下的性能，而 MSE 损失有助于提高模型在 PLDA 打分下的性能。

表 5 联合训练实验结果

Table 5 The experiment results of different loss function combination

损失函数	数据增强	Cosine			PLDA		
		EER(%)	DCF(0.01)	DCF(0.001)	EER(%)	DCF(0.01)	DCF(0.001)
AM-Softmax+MSE(0.2)		5.138	0.484 0	0.646 4	4.369	0.420 3	0.587 5
AM-Softmax+MSE(0.2)	是	5.419	0.504 4	0.648 2	3.884	0.414 2	0.486 2
AM-Softmax+MSE(0.1)		4.767	0.445 8	0.678 1	3.723	0.403 5	0.540 9
AM-Softmax+MSE(0.1)	是	4.629	0.445 6	0.553 2	3.635	0.392 5	0.530 0
AM-Softmax+MSE(0.05)		4.786	0.469 1	0.616 2	3.521	0.349 9	0.439 8
AM-Softmax+MSE(0.05)	是	4.467	0.388 8	0.515 8	3.229	0.337 3	0.438 2

2.4 联合训练与模型集成的对比

模型集成是指通过分数融合的方式，集成采用 MSE 损失函数和 AM-Softmax 损失函数的 2 种模型，2 种模型使用离线增强，将它们测试集的打分结果进行加权平均，然后再计算 EER 等性能指标。

从表 6 中可以看出，在余弦打分后端与 PLDA 打分后端下，模型集成的性能均略低于联合训练方式，2 种训练方式实验结果相差不大，但相比于模型集成需要训练多个模型进行集成，而联合训练只需要一个模型，节约了计算资源，更加高效。

表 6 模型集成的实验结果

Table 6 The experimental results of model integration

打分后端	联合损失	模型集成
余弦打分	4.396	4.533
PLDA	3.229	3.426

2.5 与其他方法的对比

结合以上实验可以得出性能最好的是采用数据增强和联合损失的网络结构，表 7 展示了和其他论文中同样使用 VoxCeleb1 数据集的说话人识别方法的实验结果的比较。

表 7 与其它方法对比的实验结果

Table 7 The experimental results compared with other methods

方法	训练集	EER(%)
VGG[13]	VoxCeleb1	7.8
1D-CNN[18]	VoxCeleb1	5.9
SincNet+LIM[19]	VoxCeleb1	5.8
Deep length normalization[20]	VoxCeleb1	5.01

续表 7

方法	训练集	EER(%)
RawNet[21]	VoxCeleb1	4.0
TDNN＋Attention[22]	VoxCeleb1	3.83
Res2Net-50-sim[23]	VoxCeleb1	3.484
ResNet＋联合损失(本文)	VoxCeleb1	3.229

从表 7 可以看出，提出的方法与其他方法对比，EER 最低降低了 8%，达到了 3.229%，性能均优于表中提到的其他方法。

3　结　语

提出了一种基于知识蒸馏与 ResNet 的声纹识别方法。将传统无监督声纹识别方法与基于深度学习的声纹识别方法相结合，用蒸馏损失 MSE 约束 ResNet 声纹特征和 I-Vector 的差异，提高了声纹识别的准确率。此外，研究进一步采用了 2 种不同数据增强方式对数据集进行扩充，增强了模型对噪声环境的适应性，提高了系统的鲁棒性，验证了 2 种增强方式在声纹识别任务中的有效性。设计的 ResNet 模型包括了注意力统计池化，结合知识蒸馏损失与度量学习损失设计了新的联合训练损失，相较于模型集成的方法，在 2 种打分后端下，联合训练方法的 EER 均低于模型集成方法。构建了端到端的声纹识别模型，与大多数基于深度学习的方法相比，能够将 EER 进一步降低为 3.229%。

参考文献：

[1] 郑方，Askar R，王仁宇，等. 生物特征识别技术综述[J]. 信息安全研究，2016，2(01)：12-26.
Zheng F，Askar R，Wang R Y，et al. Review of biometric recognition technology [J]. Information Security Research，2016，2(01)：12-26.(in Chinese)

[2] Hanifa R M，Isa K，Mohamad S. A review on speaker recognition：Technology and challenges[J]. Computers & Electrical Engineering，2021，90(4)：107005.

[3] 孙冬梅，裘正定. 生物特征识别技术综述[J]. 电子学报，2001(S1)：1744-1748.
Sun D M，Qiu Z D. Review of biometric recognition technology[J]. Acta Electronica Sinica，2001(S1)：1744-1748.(in Chinese)

[4] Zhang C，Kazuhito K，Hansen J. Text-independent speaker verification based on triplet convolutional neural network embeddings[J]. IEEE/ACM Transactions on Audio，Speech，and Language Processing，2018，26(9)：1633-1644.

[5] Hansen J，Hasan T. Speaker recognition by machines and humans：a tutorial review[J]. IEEE Signal Processing Magazine，2015，32(6)：74-99.

[6] 谭萍，邢玉娟. 噪声环境下文本相关说话人识别方法改进[J]. 西安工程大学学报，2016，30(005)：639-644.
Tan P，Xing Y J. Improvement of text-related speaker recognition method in noise environment[J]. Journal of Xi'an Polytechnic University，2016，30(005)：639-644.(in Chinese)

[7] 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.

[8] Kenny P，Boulianne G，Ouellet P，et al. Joint factor analysis versus eigenchannels in speaker recognition[J]. IEEE Transactions on Audio，Speech and Language Processing，2007，15(4)：1435-1447.

[9] 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.

[10] Hinton G，Deng Li，Yu Dong，et al. Deep neural networks for acoustic modeling in speech recognition：The shared views of four research groups[J]. IEEE Signal processing magazine，2012，29(6)：82-97.

[11] Snyder D，Garcia-Romero D，Povey D，et al. Deep neural network embeddings for text-independent speaker verification

[C]// International Speech Communication Association. Proc. Interspeech 2017, August 20-24, 2017. Stockholm, Sweden, France: ISCA, 2017: 999-1003.

[12] 胡青, 刘本永. 基于卷积神经网络的说话人识别算法[J]. 计算机应用, 2016, 36(A01): 79-81.
Hu Q, Liu B Y. Speaker recognition algorithm based on convolutional neural network[J]. Journal of Computer Applications, 2016, 36(A01): 79-81.(in Chinese)

[13] Nagraniy, A, Chung J S, Zisserman A, et al. VoxCeleb: A large-scale speaker identification dataset[J]. Proceedings of the Annual Conference of the International Speech Communication Association, 2017: 2616 - 2620.

[14] 郭玥秀, 杨伟, 刘琦, 等. 残差网络研究综述[J]. 计算机应用研究, 2020, 37(5): 1292-1297.
Guo Y X, Yang W, Liu Q, et al. Research on residual networks[J]. Application Research of Computers, 2020, 37(5): 1292-1297.(in Chinese)

[15] Ko T, Peddinti V, Povey D, et al. A study on data augmentation of reverberant speech for robust speech recognition[C]// 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), March 5-9, 2017, New Orleans, LA. New Jersey: IEEE, 2017: 5220-5224.

[16] Okabe K, Koshinaka T, Shinoda K. Attentive statistics pooling for deep speaker embedding[J]. Proc. Interspeech 2018, 2018: 2252-2256.

[17] Nagrani A, Chung J S, Xie W, et al. Voxceleb: large-scale speaker verification in the wild[J]. Computer Speech & Language, 2020, 60: 101027.

[18] Shon S, Tang H, Glass J. Frame-level speaker embeddings for text-independent speaker recognition and analysis of End-to-End model[C]// 2018 IEEE Spoken Language Technology Workshop (SLT), December 18-21, 2018, Athens, Greece. New Jersey: IEEE, 2018: 1007-1013.

[19] Ravanelli M, Bengio Y. Learning speaker representations with mutual information[C]// International Speech Communication Association. Proc. Interspeech 2019, September 15-19, 2019. Graz, Austria, France: ISCA, 2019: 1153-1157.

[20] Cai W C, Chen J K, Li M. Analysis of length normalization in end-to-end speaker verification system[C]// International Speech Communication Association. Proc. Interspeech 2018, September 2-6, 2018. Hyderabad, India, France: ISCA, 2018: 3618-3622.

[21] Jung J, Heo H S, Kim J, et al. RawNet: advanced end-to-end deep neural network using raw waveforms for text-independent speaker verification[C]// International Speech Communication Association. Proc. Interspeech 2019, September 15-19, 2019. Graz, Austria, France: ISCA, 2019: 1268-1272.

[22] Zhu Y K, Ko T, Snyder D, et al. Self-attentive speaker embeddings for text-independent speaker verification[C]// international speech communication association. September 2-6, 2018. Hyderabad, India, France: ISCA, 2018: 3573-3577.

[23] 陈志高, 李鹏, 肖润秋, 等. 文本无关说话人识别的一种多尺度特征提取方法[J]. 电子与信息学报, 2021, 43(11): 3266-3271.
Chen Z G, Li P, Xiao R Q, et al. A multi-scale feature extraction method for text-independent speaker recognition[J]. Journal of Electronics and Information Technology, 2021, 43(11):3266-3271.(in Chinese)

(编辑 侯 湘)