一种结合帧级特征预测的多任务学习声纹确认方法

李　晋
（科大讯飞股份有限公司，安徽 合肥 230088）

摘要：目前主流的声纹确认算法通常采用有监督、区分性的训练方式得到神经网络模型，如卷积神经网络（Convolutional Neural Network，CNN）、长短时记忆网络（Long-Short-Term Memory Network，LSTM）等，再利用该神经网络模型提取语音中包含个性化信息的声纹模型向量，从而进行相同人或不同人的声纹相似性比对。文章提出一种结合帧级特征预测的多任务学习训练方式，额外增加对帧级特征进行预测的神经网络模型分支，通过联合训练达到提高声纹确认算法性能的目的。在基准 VoxCeleb 三个测试集合上开展的实验结果表明，本文提出的方法可以有效提升声纹确认算法性能。

关键词：声纹确认；多任务学习；帧级特征预测
中图分类号：TN912.34　　**文献标识码：**A　　**文章编号：**2096-9759（2023）06-0001-04

A Voiceprint Verification Method Using Multi task Learning Combined with Frame-Level Feature Prediction

LI Jin
(IFLYTEK Co., Ltd, AnhuiHefei 230088)

Abstract: The current mainstream voiceprint verification algorithms usually use a supervised and discriminative training method to optimize a neural network, such as Convolutional Neural Network (CNN), Long-Short-Term Memory Network (LSTM), etc., and then to obtain voiceprint model vectors of utterances containing personalized information through this optimized neural network, so as to compare the voiceprint similarity of utterances from the same person or different persons. This paper proposes a multi-task learning training method combined with frame-level feature prediction. By adding an additional neural network model branch for frame-level feature prediction, the performance of the voiceprint verification algorithm can be improved through joint training. The experimental results carried out on three testsets of benchmark VoxCeleb show that the method proposed in this paper can effectively improve the performance of the voiceprint verification algorithm.

Keywords: voiceprint verification; multi-task learning; frame-level feature prediction

0 引言

声纹确认是生物认证领域中的关键技术之一，利用语音信号直接进行说话人的身份认证，不仅具有无需记忆、判决简单的特点，而且可以在用户不知情的情况下进行认证，具有较高的用户接受度，其广泛应用在国家安全、金融领域、智能家居等场景。在其中某些应用领域中，由于身份认证准确度要求较高，建立精准、鲁棒的声纹模型，对准确进行身份认证起到关键作用。Najim Dehak 等人提出的 i-vector 系统[1]，利用大量的语料，训练得到覆盖各种环境和信道的全变量空间（Total Variability）。利用此空间，将一段语音映射成维度固定统一的声纹模型向量（i-vector），从而根据声纹模型向量的相似性来判断两段语音是否来自于同一个说话人。该方法已成为测试语音时长较长（3 分钟以上）情况下声纹确认的主要算法。但该算法在语音时长较短的情况下，由于统计量计算不充分，导致建立的声纹模型不够稳定，认证准确度下降严重。

近年来，深度学习方法在众多研究领域取得令人瞩目的成就，通过组合分析低层特征，形成抽象的高层属性描述，以发现数据的结构性特征表示。一部分研究学者采用深度神经网络（Deep Neural Networks, DNN）方法[2]，对特征进行扩帧处理后，通过预测声纹的身份标签来提取声纹模型向量。这种方法虽然充分考虑到特征间的连贯性，采用多帧特征共同预测声纹的身份标签，但该种方法仍然局限于时域特性，无法建立时域和频域之间的关联关系，最终声纹确认准确度不高。深度卷积神经网络（Convolutional Neural Networks, CNN）[2]是近年发展起来并引起广泛重视的一种高效学习方法。首先对局部特征进行感知，然后在更高层将局部的信息综合起来，并利用多层卷积网络堆叠，提取时域和频域间交织的声纹信息。相对于单纯的全变量空间的因子分析方法，卷积神经网络可以对时域和频域进行联合分析，深刻挖掘语音时频谱图中的声纹信息，获得更加细致的声纹特征表达，从而建立精准的声纹模型。

深度残差神经网络（Deep Residual Network, ResNet）[3]是目前声纹确认领域中比较先进的一种网络结构，它在一定程度上缓解了普通卷积神经网络随着网络深度增加而出现的性能退化问题，即随着网络深度增加，梯度反向传播至前面的网络层时，冗长的连乘运算会使得梯度变得非常小，造成梯度消失的现象。深度残差神经网络在普通卷积神经网络的基础上增加恒等映射，使残差结构学习输出和输入之间的差值，保证了梯度回传过程中的完整性，降低了网络深度增加带来的训练难度。压缩和激励网络模块（Squeeze-and-Excitation Network, SENet）[4]是一种应用广泛的基于注意力机制的神经网络优化算法。它使用一个规模较小的子网络，一般采用全局平均池化（Global Average Pooling）、下采样（Down Sampling）、激活（Activation）、上采样（Up Sampling）、激活等操作，计算卷积神经网络中每个通道的特征图权重，用计算得到的权重对各个特征图进行加权求和，得到重新标定后的特征图。压缩和激励网络模块充分利用各个通道的特征图之间的相关性，促进网络对特征提取和挑选等性能的不断提升。压缩和激励网络

收稿日期：2023-03-17
基金项目：国家重点研发计划资助（2022YFF0608503）。
作者简介：李晋（1987-），男，安徽蒙城人，博士研究生，科大讯飞股份有限公司，工程师，主要研究方向：声纹识别、深度学习等。

模块可以和现有的多个神经网络结构相结合，将其嵌入到深度残差神经网络中，通过学习残差神经网络中每个通道的重要程度，且根据每个通道重要程度的大小对残差神经网络的特征图进行抑制或提升，有效提升残差神经网络的特征表达能力。

声学特征中声纹个性化信息抽取是决定声纹确认准确度高低的决定性因素，排除文本等信息干扰，可以有效提升声纹确认的准确度。最近两年，基于自监督学习（Self-Supervised Learning）的训练方法在语音识别、图像分类等领域大放异彩，其突出特点是通过自回归重构的方式，学习特征层面潜在的结构性信息。和人工设计的诸多特征相比，自监督学习方法能够更好得进行声学特征表达。对比预测编码（Contrastive predictive coding，CPC）[5]和 wav2vec2.0[6]是比较典型的两种自监督学习方法，其中对比预测编码因其简单的网络结构和训练数据需求较低等优势，受到更多研究者的青睐。

本文提出一种在带有压缩和激励网络模块的深度残差神经网络基础上，采用多任务学习的方式，增加基于自监督学习的精简对比预测编码辅助网络的声纹确认方法。最终实现优化残差神经网络浅层网络特征表达能力，提高声纹确认性能的目的。在声纹确认任务的基准测试集合 VoxCeleb[7]上开展的实验表明，相对基线方法，本文提出的声纹确认方法，准确度有一定的提升。

1 声纹确认的基线方法

由声学特征组成的语谱图可以看成是一种纹理图像，说话人的语音中蕴含的个性化声纹信息，通常体现在语谱图中时域和频域特征的组合变化上。利用卷积神经网络进行声纹确认，恰好可以充分利用卷积神经网络中的感受野，对语谱图的时频域同时进行卷积运算，获取最优的特征表达提取空间。因此本文采用带有压缩和激励神经模块的深度残差神经网络作为声纹确认的基线方法。

该基线方法主要包括两个重要组件：残差网络模块与压缩和激励网络模块，下面将对这两个组件进行具体介绍。

1.1 残差网络模块

从理论上讲，随着神经网络深度的不断增加，神经网络的特征表达能力应该逐渐增强。但是在论文[3]中的实验发现，网络深度的增加会导致退化问题，即随着网络深度的增加，在训练集和测试集上目标函数损失值不减反增。这是因为网络深度增加会引起梯度在回传过程中消失，导致无法对前面若干网络层的权重和偏差进行有效更新。采用具有“shortcut connections”的跳跃式结构后，深度残差网络可以越过中间若干层，直接将参数传递给后面层，实现对深度网络全部网络层的权重和偏差参数的全局更新。

残差网络模块如图 1 所示。

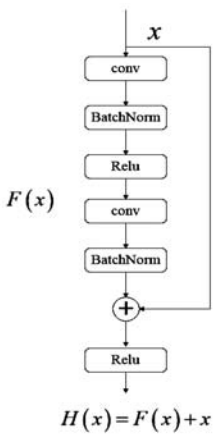

图 1 残差网络模块

在图 1 所示的残差网络模块中，x 是输入特征图，$H(x)$ 是经过网络映射后的输出特征图，$F(x)$ 表示输入特征图和输出特征图之间的残差项。残差神经网络训练的目标是要使 $F(x)=H(x)-x$ 不断趋向于 0。当 $F(x)=0$，有 $H(x)=x$，此时即为恒等映射。通过学习 $F(x)\to 0$，从而避免产生神经网络训练过程中出现梯度趋向于 0 的问题，使得深层神经网络可以正常进行权重和偏差的更新，实现深层神经网络的优化。

1.2 压缩和激励网络模块

压缩和激励网络模块用于学习卷积神经网络中不同通道特征图的权重，这种结构可以对通道之间的依赖关系进行建模，捕捉各个通道的空间相关性，从而提高神经网络的表达能力。该模块能够自动学习每个通道特征图之间的相对重要程度，强化与任务相关的特征，同时抑制与任务无关的特征，从而实现对各个通道的特征图进行重新标定的目的。具体由三部分构成：

（1）压缩过程：通过全局平均池化，将每个通道的特征图压缩成一个实数，这个实数表征各个通道上相应的全局分布，并且在靠近输入的浅层上也可以获得全局的感受野，突破普通卷积神经网络中浅层感受野较小的制约；

（2）激励过程：为每个特征通道生成归一化权重参数，其中归一化权重参数是通过 Relu 和 Sigmoid 激活函数显式地为特征通道间的相关性进行建模。其中为降低参数量和提高计算效率，先通过下采样降低通道个数，再通过上采样恢复为原始的通道个数；

（3）重标定过程：将激励过程后输出的归一化权重参数看成是经过特征选择后的每个特征通道的重要程度，通过点乘逐通道加权到原先的特征图上，在各个通道上完成对原始特征的重标定。

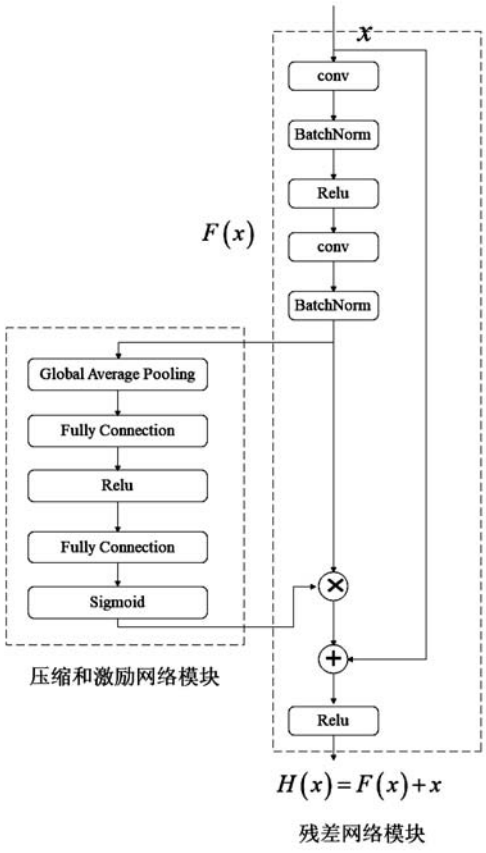

图 2 残差网络模块与压缩和激励网络模块结合结构

通过采用压缩和激励网络模块，给卷积提取的各个通道的特征图之间赋予相应的相对权重，提升了特征表达的质量，促使神经网络的加速收敛。

在实际应用中，通常将残差网络模块与压缩和激励网络模块结合使用，如图 2 所示。对残差网络模块中产生的残差项 $F(x)$ 进行各个通道间相关性权重系数调整，在局部的卷积操作模块中插入全局上下文信息，可以获得更加精准的残差项 $F(x)$ 的特征表达能力，进而提高网络输出特征图 $H(x)$ 的特征表达能力。通过压缩和激励网络模块，使得深度残差神经网络在浅层时也充分融入语音片段的全局属性，拓展上下文信息。相对标准深度残差神经网络，特征表达性能得到进一步提升。

2 多任务学习的声纹确认方法

2.1 精简对比预测编码

对比预测编码是一种自监督的特征提取模型，主要通过自回归（Autoregressive model，AR）的模型，对原始声学特征进行潜在深层特征的对比学习，获得对特征预测最有效的潜在深层特征的表达能力，实现对原始特征重建误差最小的目的。[5] 中提出先在语音波形层面进行非线性编码得到浅层特征，再采用自回归模型进行潜在深层特征提取。本文为了和现有基线方法进行多任务学习，采用和基线方法相同的声学特征作为对比预测编码网络的输入，加速神经网络的训练。本文采用的精简对比预测编码（Simple CPC，SCPC）的结构图如图 3 所示。

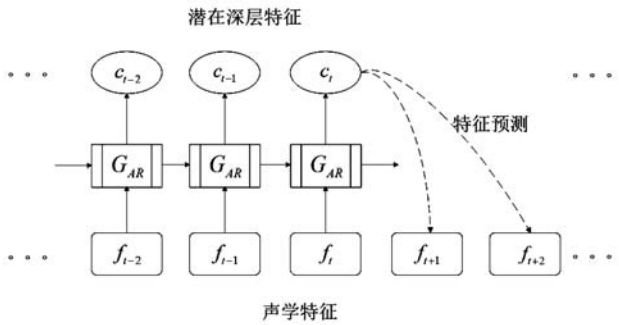

图 3 精简对比预测编码的结构图

在本文采用的精简对比预测编码的结构中，自回归模型选择长短时记忆网络（Long-Short-Term Memory Network，LSTM），因为长短时记忆网络可以通过调节语音声学特征 f_t 的长度，完成对上下文信息的充分挖掘和表达，得到潜在深层特征 c_t，即：

$$c_t = LSTM(f_{\le t}) \tag{1}$$

根据 c_t 对第 k 帧后的声学特征 f_{t+k} 进行线性 W_k 预测，得到预测的声学特征 f'_{t+k}：

$$f'_{t+k} = W_k \cdot c_t \tag{2}$$

最终使得预测的声学特征 f'_{t+k} 和真实的声学特征 f_{t+k} 尽可能相似，同时和其它帧的声学特征 f_j 差异尽可能大，其损失函数记表示为：

$$L_{SCPC} = -\mathrm{E}\left[\log \frac{g_k(f_{t+k}, c_t)}{\sum_{f_j \in F} g_k(f_j, c_t)}\right] \tag{3}$$

其中 (f_{t+k}, c_t) 是正样本对，(f_j, c_t) 是负样本对，$g_k(f_{t+k}, c_t)$ 是密度比函数，用于描述声学特征的预测值和真实值之间的相似关系：

$$g_k(f_{t+k}, c_t) = \exp\left(f_{t+k}^T W_k c_t\right) \tag{4}$$

精简对比预测编码的损失函数 L_{SCPC} 表明，正样本对 (f_{t+k}, c_t) 的相似度越高，网络的特征预测能力越强，此时可以提取出与任务相关性较高的潜在深层特征。

2.2 多任务学习的声纹确认方法

多任务学习[8]通过共享相关任务的特征表达，将多个相关任务联合在一起进行神经网络训练，可以提高神经网络模型的预测能力和泛化能力。多任务学习的先决条件是多个任务之间具有一定的相关性，即可以通过共享隐层的方式，促进多个任务同步学习，得到各个任务内部不易单独挖掘的隐含相关特性。

本文提出一种结合帧级特征预测的多任务学习声纹确认方法，以残差神经网络 ResNet-34[9] 为主要网络，并在此基础上增加压缩和激励网络模块，构成 SE-ResNet-34 主干网络。采用精简对比预测编码神经网络为辅助网络的网络结构，实现声纹确认准确度的提升。该多任务学习的声纹确认方法如图 4 所示，具体使用方法如下：

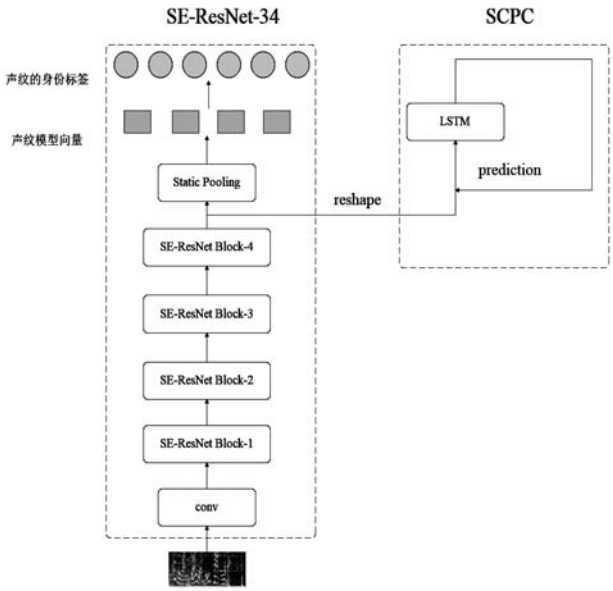

图 4 多任务学习的声纹确认方法

步骤 1：提取语音的声学特征，按照固定窗长切分形成若干个语谱图。对于不足固定窗长的部分，可以采用特征复制补齐的方式，填充到满足固定窗长的要求；

步骤 2：将语谱图分别经过卷积模块、带有压缩和激励的残差网络模块（SE-ResNet Block-1~4），获得深层特征表达 f_t；

步骤 3：f_t 经过统计池化（StaticPooling，SP）后，接入一层全连接层，得到用于表征声纹身份信息的声纹模型向量。再将此声纹模型向量通过一层全连接层，用于对声纹的身份标签进行预测，该预测值和真实值之间的差异用 AAM（Angular Additive Margin）[10]softmax 损失函数来表示，记作 L_{AAM}；

步骤 4：在每个训练的批（batch）中，都可以提取出语音对应的声纹模型向量。该批内的声纹模型向量中，既包含相同人的多个声纹模型向量，也包含不同人的多个声纹模型向量。将该批中的声纹模型向量按照是否属于同一人，采用批内困难三元组（batch hard triplet，BHT）[11] 损失函数计算批内的声纹模型向量之间的差异，记作 L_{BHT}；

步骤 5：f_t做一次尺度变换，将四维特征图中的通道维度和特征维度拼接起来，形成三维特征图，接入 LSTM 网络进行帧级特征预测，其帧级特征预测值和真实值之间的差异用前文公式（3）表示，记作 L_{SCPC}；

步骤 6：利用全局损失函数 $L_{OSS}=L_{AAM}+L_{BHT}+L_{SCPC}$ 对多任务学习的声纹确认模型进行网络参数更新，完成声纹确认模型的训练。

3 实验结果和分析

本文在声纹确认任务的基准测试集合 VoxCeleb 上开展相关实验进行效果对比验证。VoxCeleb 数据集合是牛津大学的研究人员在 YouTube 上通过自动裁剪和人工抽检后，构建的大规模互联网领域的开源声纹识别数据，可以进行声纹确认和声纹检索的研究工作。它由两个数据子集构成，其中一个是包含 1211 声纹个体的数据集合 VoxCeleb-1，另一个是包含 5994 声纹个体的数据集合 VoxCeleb-2。和其它研究人员的实验类似，本文也采用 VoxCeleb-2 作为训练数据，VoxCeleb-1 的三个子集（VoxCeleb1-O_cleaned、VoxCeleb1-E_cleaned、VoxCeleb1-H_cleaned）作为声纹确认的测试数据，测试指标用等错误率（Equal Error Rate, EER）来表示。该指标越小，说明声纹确认的性能越好，其准确度越高。

实验采用 FilterBank 声学特征构成语谱图，多任务学习声纹确认方法中的自回归模型是三层 LSTM，特征预测模型是一层线性层，负样本对的个数为 16。

采用多任务学习声纹确认方法的测试效果如表 1 所示。

表 1 多任务学习声纹确认方法的测试效果

实验方法	特征线性预测模型的个数	测试集合（EER，%）		
		VoxCeleb1-O_cleaned	VoxCeleb1-E_cleaned	VoxCeleb1-H_cleaned
多任务学习声纹确认方法	1	1.36	1.34	2.37
	4	1.23	1.26	2.25
	8	1.24	1.33	2.34
	12	1.34	1.37	2.35

从表 1 中可以看出，随着特征线性预测模型的个数增多（从 1 个提高至 4 个），在基准 VoxCeleb 的三个测试集合上，声纹确认的准确度均有一定提升；但继续增加特征线性预测模型的个数，声纹确认的准确度有所下降，这可能是由于多个线性预测模型导致特征预测差异较大时，无法准确对特征预测结果进行归并导致。

采用声纹确认的基线方法和多任务学习方法的测试结果对比如表 2 所示。

表 2 测试结果对比

实验方法	测试集合（EER, %）		
	VoxCeleb1-O_cleaned	VoxCeleb1-E_cleaned	VoxCeleb1-H_cleaned
ResNet-34[12]	1.67	1.81	3.23
基线方法 SE-ResNet-34	1.26	1.33	2.32
本文提出的多任务学习方法	1.23	1.26	2.25

从表 2 中可以看出，本文采用的声纹确认基线方法（SE-ResNet-34），同目前主流的残差神经网络 ResNet-34 相比，通过增加压缩和激励网络模块，声纹确认的效果有大幅提升。在 SE-ResNet-34 的基础上，采用本文提出的结合帧级特征预测，即精简对比预测编码 SCPC 结构后，对 SE-ResNet-34 的浅层特征提取模块（图 4 中的 SE-ResNet Block-1~4 结构）进行同步优化，声纹确认的效果又得到进一步提升，获得目前较优的声纹确认准确度。

参考文献：

[1] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. Front-end factor analysis for speaker verification[J]. IEEE Transactions on Audio,Speech,and Language Processing, 2011,19(4):788-798.

[2] 陈晨,韩纪庆, 陈德运等.文本无关说话人识别中句级特征提取方法研究综述[J]. 自动化学报, 2022, 48(3): 664-688.

[3] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition[C]. in IEEE 29th Conference on Computer Vision and Pattern Recognition, 2016:770-778.

[4] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks[C]. in IEEE 31th Conference on Computer Vision and Pattern Recognition, 2018:7132-7141.

[5] Oord A, Li Y, Vinyals O. Representation learning with cont rastive predictive coding[J]. arXiv preprint arXiv:1807.03748, 2018.

[6] Baevski A, Zhou Y, Mohamed A, et al. wav2vec 2.0: A framework for self-supervised learning of speech representations [J]. Advances in neural information processing systems, 2020, 33:12449-12460.

[7] A. Nagrani, J. Chung, and A. Zisserman. Voxceleb: a large-scale speaker identification dataset[C]. in 18th Annual Conference of the International Speech Communication Association, 2017:2616-2620.

[8] Ruder S. An overview of multi-task learning in deep neural networks[J]. arXiv preprint arXiv:1706.05098, 2017.

[9] J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2:deepspeaker recognition[C].in 19th AnnualConference of the InternationalSpeech Communication Association, 2018: 1086-1090.

[10] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface : Additive angular margin loss for deep face recognition[C]. in IEEE 32th Conference on Computer Vision and Pattern Recognition, 2019:4685-4694.

[11] A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification [J]. arXiv preprint arXiv: 1703.07737, 2017.

[12] Tong F, Zhao M, Zhou J, et al. ASV-Subtools: Open source toolkit for automatic speaker verification[C]. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 6184-6188.