DOI:10.19651/j.cnki.emt.1802145

基于深度神经网络和多元损失的说话人识别

关 健 王 敏
(河海大学 计算机与信息学院 南京 211100)

摘 要:生物特征识别技术相对于传统密码等方式具有更高的可靠性,而作为生物特征识别技术的重要研究方向之一的声纹识别方法,研究更精确的声纹识别方法具有更高的研究意义。随着深度学习的发展,深度学习应用在声纹识别技术上成为在声纹识别领域研究的重点。提出一种基于深度神经网络和 beyond triplet loss 相结合的说话人识别方法,模型通过梅尔频率倒谱系数(MFCC)提取 MFCC 声学特征,对 MFCC 声学特征提取说话人声纹特征,然后进行多元损失的模型训练。实验结果表明,DNN-BTL 算法在说话人识别领域比高斯混合-隐马尔可夫模型(GMM-HMM)具有更好的识别效果。

关键词:多元损失;深度学习;深度神经网络(DNN);说话人识别;声纹识别
中图分类号:TP391;TN919.81 **文献标识码**:A **国家标准学科分类代码**:520.2040

Speaker verification based on deep learning and beyond triplet loss

Guan Jian Wang Min
(College of Computer and Information, Hohai University, Nanjing 211100, China)

Abstract: Biometric recognition technology has higher reliability than traditional cryptography. As one of the important research directions of biometrics, voiceprint recognition method has more research significance to study more accurate voiceprint recognition methods. With the development of deep learning, the application of deep learning in voiceprint recognition technology has become the focus of research in voiceprint recognition field. In this paper, a speaker recognition method based on deep neural network and beyond triplet loss is proposed. The model extracts the acoustic characteristics of MFCC through Mel-frequency cepstral coefficients, and extracts the voiceprint characteristics of the speaker from the MFCC acoustic characteristics, and then carries out the beyond triplet loss model training. Experimental results show that DNN-BTL algorithm has better recognition effect in speaker recognition field than Gussian mixture model-hidden Markov model.

Keywords: beyond triplet loss; deep learning; speaker verification; voiceprint recognition

0 引 言

说话人识别是生物特征识别方面的主要研究方向之一,未来可以广泛的应用在商业、教育、安全等领域,而说话人识别研究的重点就是识别的准确性和可靠性。在实际商用方面,说话人识别受环境、噪音、例如感冒等声音变化等因素影响较大,从而降低识别准确性。因此,更具鲁棒性和抗干扰性的说话人识别技术成为如今说话人识别领域研究的重点。

随着计算机技术的高速发展,人们对与电子计算机之间的交互能力要求越来越高,尤其是在生物特征识别方面。而作为生物特征识别的重要组成部分的声纹识别技术正在逐渐走向成熟并且商用。2014 年,微软发布全球第一款人工智能助理“小娜”,随后发布中文版本。苹果于 2015 年将人工智能应用在其产品 iphone 的语音识别助手“siri”上。而在国内方面,科大讯飞于 2015 年发布与各大银行合作的说话人验证+人脸识别进行银行密码认证服务,其后发布声纹识别在移动端的 SDK。腾讯推出采用微信语音识别技术和腾讯云相结合的智能服务系统——人工智能“小微”。

现有的说话人识别方法主要可以分为以下 3 类:

1)模板匹配方法。主要过程是对每个说话人的训练语句提取特征形成特征矢量,然后对优化后的特征矢量求去一个特征矢量集合来表征,在识别时同样提取特征矢量,然后跟特征矢量集合所有模版进行匹配,其中匹配主要通过距离算法实现。常用模板匹配方法为动态时间规整(dynamic

收稿日期:2018-09-27

time warping，DTW）和矢量量化（vector quantization，VQ）方法。该方法在一定程度上提高了说话人识别率，在小训练集时建模简单实时性好。但受外界干扰因素大，且若两个说话人的声音相似时效果不好。

2）概率模型方法。主要过程为对每个说话人提取特征矢量，然后基于统计为其建立数学模型。在识别阶段，将测试语音的特征矢量与训练数学模型进行匹配，从概率方面计算测试特征矢量与模型的相似度。常用概率模型主要为高斯混合-隐马尔可夫模型（GMM-HMM）和 GMM-UBM 模型。该方法在解决大量词汇、连续语音等方面表现良好，但该方法在前期需要大量训练，且为每个说话人建立模型占用大量存储资源。

3）深度学习方法。主要过程是对每个说话人提取声学特征，采用深度学习方法声学特征进行训练，然后基于奖惩函数进行评分匹配。常用的深度学习方法为 i-vector 特征与深度学习（CNN 或 DNN）相结合方法以及利用 LSTM 网络建模。该方法的容错性、抗干扰性和灵活性较高，对于处理不确定信息能力较强，但该方法训练时间较长，切网络规模随说话人数呈线性或指数增长，可能大到难以训练。

基于 MM-HMM、深度学习方法的缺点，本文提出了深度学习与 beyond triplet loss 相结合方法，实验证明，该方法取得了比 GMM-HMM 和深度学习方法更好的识别效果，识别速度快，且系统的容错性更高。

1　模型建立

传统的 GMM-HMM 模型作为一种基于统计学习的特征识别方法，其主要方法为建立声学模型和语言模型，然后基于 HMM 的最大后验估计来识别。现阶段的深度学习方法主要是通过神经网络对大量数据集进行训练，自动提取训练数据中的特征。在传统 GMM-HMM 模型对声学特征一般采用推荐或者评分算法进行分类，而在深度学习方法中，会自动学习数据中的声纹特征信息，然后进行分类识别，显著提高了声纹识别的识别效果[1-3]。

1.1　系统流程

如图 1 所示，一个完整的说话人识别系统由噪声抑制、特征提取、声学建模和评分匹配组成，在特征提取阶段，将输入的语音由声波信号提取语音的声学特征，如音节、音素等。到声学建模阶段，模型作为识别语音声学特征基元的模板，确定声学模型然后经过解码器的处理输出相应的识别结果。

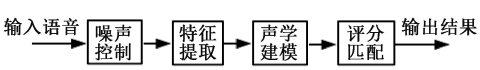

图 1　说话人识别过程

现有的说话人识别模型一般采用 GMM-HMM 模型，其主要过程如下：对输入语音进行噪音抑制处理，提取语音的 Mel 频率倒谱系数（Mel-frequency copstral coefficients，MFCC）特征序列，然后采用 HMM 统计学习模型对输出序列进行基于概率的评分，得到的评分结果过滤某个设定的阈值，超过此阈值的最高评分为输出判别结果[4-5]。具体过程如图 1 所示。

1.2　特征提取

MFCC 的整个提取过程如图 2 所示[6]。

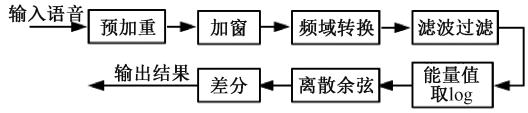

图 2　MFCC 提取过程

其中周期持续时间设置为 10～25 ms，移码为 10 ms，每步骤具体过程如下。

1）预加重

在声学发声过程中，声带和嘴唇会对声纹造成影响，因此需要突显高频的共振峰，来补偿受到声带和嘴唇影响的高频部分语音信号。即在频域上乘以某与频率正相关的系数。预加重公式如式（1）所示。

$$S'_n = S_n - k \times S_{n-1} \tag{1}$$

2）加窗

预加重之后需要对语音信号进行加窗处理，即每次处理数据仅处理窗中数据。采用汉明窗对信号进行加窗处理，可以平滑语音信号。相对于用矩阵直接截断加窗会造成频率泄露，加汉明窗的幅频特性是旁瓣衰减较大。汉明窗公式[7]如下：

$$S'_n = \left\{0.54 - 0.46 \times \cos\left(\frac{2\Pi(n-1)}{N-1}\right)\right\} \times S_n \tag{2}$$

3）频域转换

然后将时域信号转化到频域进行后续的频率分析，具体实现如下[8]。

幅度谱：

$$S_i(k) = \sum_{n=1}^{N} S_i(n) \mathrm{e}^{-\mathrm{j}2\Pi kn \mid N} \quad 1 \leqslant k \leqslant K \tag{3}$$

功率谱：

$$P_i(k) = \frac{1}{N} \mid S_i(k)^2 \mid \tag{4}$$

4）使用 Mel 刻度滤波器组过滤

使用 Mel 刻度滤波器组过滤，因为在频域转换后的频域信号中有多余部分，因此采用 Mel 刻度滤波器组对频域的幅值进行删减。具体实现过程为，对于快速傅里叶变换（FFT）得到的幅度谱，分别跟每一个滤波器进行频率相乘累加，得到的值即为该帧数据在该滤波器对应频段的能量值[9-10]。

5）能量值取 log

由于人耳对声音的感知并非线性，因此需要对过滤后的频段的能量值进行取 log 处理，便于之后对取 log 的线性能量值进行倒谱分析。

6)离散余弦变换

即对上述获得的能量值做离散余弦(discrete cosine transform,DCT)变换,其目的是获取频率谱的低频信息,相当于做反傅里叶变换之后用低通滤波器获取低频信号。采用 DCT 的优点是,由于相邻滤波器之间有重叠,因此获取的低频信号之间具有关联性。并且 DCT 可以对数据进行压缩和提取特征参数。计算公式为:

$$c_i = \sqrt{\frac{2}{N}} \sum_{j=1}^{N} m_j \cos\left(\frac{\Pi_i}{N}(j-5)\right) \tag{5}$$

7)差分

由于之前的语音为采用 10～25 ms 周期为 1 帧的分帧提取,因此每一帧仅反应了本帧语音的特征,而语音之间是有时序上的连贯性,因此为了体现这种时序上的连贯性,一般采用一阶或者二阶差分用以增加特征维度前后帧信息的维度。计算公式为:

$$d_t = \frac{\sum_{\theta=1}^{\phi}(c_{t+\theta} - c_{t-\theta})}{2\sum_{\theta=1}^{\phi}\theta^2} \tag{6}$$

1.3　基于统计的建模

传统 GMM 和 UBM 的训练过程如图 3 所示[11]。

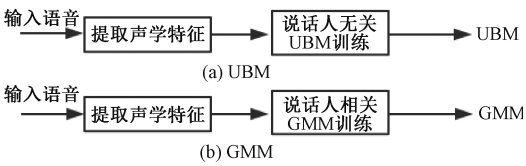

图 3　GMM 和 UBM 的训练过程

给定输入的语音序列,提取其特征向量以及说话人的模型参数,经过 HMM 计算,在第 i 次迭代后验估计后,输出其评分值,当返回评分值大于某设定阈值时,即返回识别输入语音为某说话人。否则,即评分低于某阈值即返回否[10-12]。

2　基于 DNN-多元损失的建模

在现有的语音识别领域采用的深度学习方法多数为循环神经网络(recurrent neural network,RNN),其主要原因是 RNN 引入定向循环,更强调关联性以及时序性,因此 RNN 模型(例如 LSTM 模型)在处理即时翻译等方面具有明显优势。但在说话人识别领域的效果却差强人意。而深度神经网络(DNN)在很多方面却表现出明显优势,例如图像识别、声纹识别、步态识别等方面,因此本文采用 DNN 模型用作说话人识别,并基于 DNN 与 beyond triplet loss 相结合提出一种新的说话人识别方法。

2.1　算法结构

DNN-BTL 由两个模型组成,主要过程为首先从提取的声学特征中用 DNN 模型提取说话人身份特征,然后利用余弦相似度训练多元损失,在 embedding 时最大化同类说话人的余弦相似度并最小化异类说话人的余弦相似度,然后进行识别确认。

2.2　DNN 模型

DNN 按照功能划分为输入层、隐藏层、输出层。且在每一层之间的链接都是全连接层,即第 n 层的任意一个神经元和第 $n+1$ 层任意一个神经元都相连接。因此对小规模方面来说,DNN 与感知机一样,即由一个线性关系 $z = w_i + b_i$ 和一个激活函数 $\partial(z)$ 组成。模型如图 4[12-13]所示。

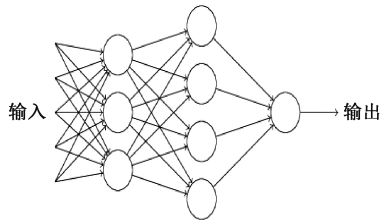

图 4　DNN 模型

具体提取 DNN-ivector 算法如下:

1)输入帧序列特征;

2)对于每一帧输入特征序列,经过 DNN 训练,然后按照后验概率估计分配到对应发声单元;

3)每个发声单元对分配到单元内的声学特征做统计处理;

4)将每个发声单元的统计量形成包含特征信息的特征矢量。

2.3　beyond triplet loss

三元损失(triplet loss)由 anchor、positive、negative 3 个元组组成,训练过程如图 5 所示。具体为从训练集中随机选取样本(anchor),再随机选取一个与 anchor 属同类的样本(positive)以及与 anchor 不属于同类的样本(negative),然后通过训练使得 anchor 与 positive 之间的距离尽可能小,anchor 与 negative 之间的距离尽可能大,即如图 6 中要求 $B_1B_3 < B_1A_3$。具体 loss 的计算公式如下[14-15]:

$$L_{trp} = \sum_{i,j,k}^{N}[\| f(x_i) - f(x_j) \|_2^2 - \| f(x_i) - f(x_k) \|_2^2 + a_{trp}]_+ \tag{7}$$

式中:x_i、x_j 属于同一类;x_k 属于另一类;a_{trp} 表示预设的 margin;$f(x_i)$ 表示归一化的高度嵌入特征。

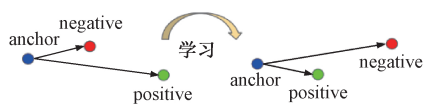

图 5　triplet loss 模型

由于 triplet loss 强调在训练时 anchor 与 positive 之间的距离尽可能小,因此在测试集上的泛化效果比较一般,其主要原因是类间距比较大,因此在 CVPR2017 上提出的 beyond triplet loss 成为解决该问题的方案之一。beyond

triplet loss 相对 triplet loss 的改进主要为减少类内方差，增大类间方差，即 loss 不仅要求 $B_1B_3 < B_1A_3$ 同时要求 $C_1C_2 < B_1A_3$，为了增加异类方差，本文采用自动最大阈值采样策略，以使损失函数中边界阈值自适应得到，如图 6 所示。

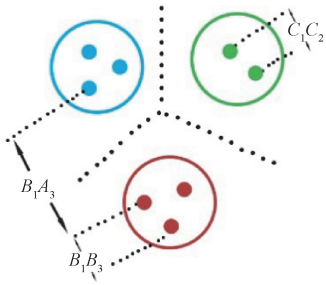

图 6　beyond triplet loss 方差

相对 triplet loss 直接采用欧氏距离作为相似性度量，本文采用 $g(x_i,x_j)$ 作为新的相似度度量。

$$L_{quad} = \sum_{i,j,k}^{N}[g(x_i,x_j)^2 - g(x_i,x_k)^2 + a_1]_+ + \sum_{i,j,k}^{N}[g(x_i,x_j)^2 - g(x_i,x_j)^2 + a_2]_+ \tag{8}$$

其中 $s_i = s_j, s_l \neq s_k, s_i \neq s_l \neq, s_i \neq s_k, g(x_i,x_j)$ 相似度评价；a_1 为类间距；a_2 为类内距，因为类间距的重要性大于类间距，因此规定 $a_1 > a_2$。具体为了确定 a 的值，防止 margin 过大无法收敛，或者 margin 过小会造成收敛慢，因此采用自适应 margin 方法。

$$a = w(u_n - u_p) = w(\frac{1}{N_n}\sum_{i,k}^{N} g(x_i,x_k)^2 - \frac{1}{N_n}\sum_{i,j}^{N} g(x_i,x_j)^2) \tag{9}$$

其中 $s_i = s_j, s_i \neq s_k$，margin 为同类距离均值与异类距离均值之差。w 为权重，具体为对应 a_1 时 $w=1$，对应 a_2 时 $w=0.5$。

3　实验结果

本文实验 1 采用了北京希尔贝壳科技有限公司开源的 aishell 中文语料库，约 178 h 包含 400 位不同口音区域说话人，在实验环境下使用高保真麦克风录制，无噪声环境，语料的采样率 16 kHz，16 bit，单声道。实验 2 采用牛津大学的 voxceleb 语料库，约 351 h 不同国家的 1 251 位说话人，该预料全部采自 YouTube 纯英文语料，且语音带有一定真实噪声，包含环境突发噪声、背景人声、笑声、回声、室内噪音、录音设备噪音等，因此该预料下用于实际使用方面识别率更有代表性。采样率 16 kHz，16 bit，单声道，PCM-WAV 音频格式。

首先对语料进行预处理，将语料进行时域到频域上的转换，然后将转换完成的频谱图作为 DNN-BTL 模型的输入。实验 1 的识别率为 99.3%；实验 2 的识别率为 92.1%。

如表 1、2 所示，DNN-BTL 模型较传统提取 i-vector 方法或者普通深度学习方法识别率明显提升。

表 1　aishell 语料库识别率　(%)

ACC	识别率
i-vector/SVM	82.5
i-vector/PLDA	87.1
CNN	91.7
DNN-BTL	99.3

表 2　voxceleb 语料库识别率　(%)

ACC	识别率
i-vector/SVM	74.6
i-vector/PLDA	81.3
CNN	86.4
DNN-BTL	92.1

4　结　　论

本文并且提出了一种相对 SVM 或者 PLDA 用作说话人识别方面更好的 DNN-BTL 模型，通过实验数据对比，基于本文 DNN-BTL 模型的说话人识别方法在识别率和鲁棒性方面较传统基于 i-vector 方法以及 CNN 方法都有明显提高。由于深度学习的发展，采用深度学习方法用作说话人识别领域具有更深的研究价值，相信在未来很长一段时间，深度学习用作说话人识别方面的识别率和鲁棒性依然会有些持续提高，因此本文在此方面依然有改进空间。

参考文献

[1] KINNUNEN T, LI H. An overview of text-independent speaker recognition: From features to supervectors[J]. Speech Commu-nication, 2010, 52(1): 12-40.

[2] NAKAGAWA S, WANG L, OHTSUKA S. Speaker identification and verification by combining MFCC and phase information[J]. IEEE Transactions on Audio Speech & Language Processing, 2012, 20(4): 1085-1095.

[3] 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.

[4] KANAGASUNDARAM A, DEAN D, VOGT R, et al. Weighted LDA techniques for i-vector based speaker verification[C] Proceedings of IEEE international conference on acoustics, speech, and signal processing, Japan: IEEE, 2012: 4781-4794.

[5] HINTON G, DENG L, DONG Y, et al. Deep neural

networks for a coustic modeling in speech recognition: The shared views of four research groups[J]. IEEE Signal Processing Magazine, 2012,29(6):82-97.

[6] DAHL G E, YU D, DENG L, et al. Context-dependent pre-trained deep neural networks for large-vocabulary speech rec ognition[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2012, 20(1):30-42.

[7] 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.

[8] JIANG Y, LEE K A, WANG L. PLDA in the i-supervector space for text-independent speaker verification[J]. EURASIP Journal on Audio, Speech, and Music Processing, 2014(1):1-13.

[9] LARCHER A, BOUSQUET P, KONG A L, et al. i-Vectors in the context of phonetically-constrained short utterances for speaker verification [C]. ICASSP, IEEE, 2012: 4773-4776.

[10] TAN T, QIAN Y, YU D, et al. Speaker-aware training of LSTM-RNNS for acoustic modeling [C]. 2016 IEEE International Conference on Acoustics, Speech and Signal Processing, 2016: 5280-5284.

[11] 杨瑞田,周萍,杨青.TEO能量与Mel倒谱混合参数应用于说话人识别[J].计算机仿真,2017,34(8):215-219,264.

[12] 周春晖,卢荣,潘姿蓉.说话人识别特征参数MFCC的提取与分析[J].电子技术与软件工程,2016(22):90.

[13] 李慧慧.基于深度学习的短语音说话人识别研究[D].郑州:郑州大学,2016.

[14] 酆勇,熊庆宇,石为人,等.一种基于受限玻尔兹曼机的说话人特征提取算法[J].仪器仪表学报,2016,37(2):256-262.

[15] 郑方,李蓝天,张慧,等.声纹识别技术及其应用现状[J].信息安全研究,2016,2(1):44-57.

作者简介

关健,硕士,主要研究方向为计算机视觉。

E-mail:guanjiankuku@hotmail.com

王敏,博士、副教授,主要研究方向为计算机视觉。