基于 Tensorflow 的声纹识别模型

王晨晔 201900810349

1. 摘要

声纹识别技术是近年来发展起来的一项重要技术。本文阐述了声纹识别产生原因，技术基础及发展现状。并使用公开数据集基于 Tensorflow 训练了 ResNet50 分类模型实现声纹识别，在 Google 的 Colab 平台训练，并取得了 90%以上的准确率。

2. 介绍

随着信息技术的发展，高效，快速的认证个人信息是目前面临的一个问题。在古代，政府为了确认通缉犯往往会采用画影图形的方式确认身份，而如今，各种各样的账户，密码，构成了现在人们的生活：手机作为控制终端有手机密码，支付宝，微信作为支付平台有支付密码，微博豆瓣等社交平台有登陆密码。这些密码都代表了认证操作者的相关信息。但在科技和互联网高速发展的今天，多平台，多维度的复杂密码易被遗忘和攻击、验证码易被截取，相关一系列的安全隐患所带来的事故时有发生。因此，传统认证方式将逐渐成为历史， 而生物特征认证方式已经登场。[1]

生物特征认证方式可分为两类：生理特征和行为特征[2]。常见的基于生理特征的认证方式有：指纹认证，人脸识别，签名认证，语音识别，甚至是 DNA 认证。生物特征具有普遍性，唯一性，稳定性，不易复制性等。与其他生物特征相比，作为行为特征的声纹具有以下特点：[3]

(1)蕴含声纹特征的语音获取方便、自然,在采集过程中涉及到的用户个人隐私信息较少，因此使用者更易接受；

(2)语音采集装置成本低廉,使用简单,一个麦克风即可,在使用通讯设备(如电话、手机)时更无需额外的录音设备；

(3)配合语音识别技术,可使声纹口令动态变化而无需担心密码遗忘、丢失和窃取问题，防止录音假冒，因此也尤为适合远程身份认证.

基于此，本文主要关注语音识别中的声纹识别技术。本文首先介绍声纹识别基础概念，声纹识别发展现状，再介绍一种基于 Tensorflow 的声纹识别技术实现以及改进，最后检验模型优劣，并应用实例。

2.1 声纹识别基础概念及发展现状

声纹识别指的是对个人声音的特征进行测定，是一种生物识别技术。一般来说，由于不同人在讲话时使用的发声器官如舌头，牙齿，口腔等在尺寸和形态上有所不同，年龄，语言习惯发生频率等不同，导致每个发言人的声纹特征都是独一无二的，为了将一个人的身份和声音相匹配，需要事先录下发言人的声音与身份信息，然后就可以根据预录制的声纹特点做身份验证。与传统的密码和密钥相比，声纹因为其独一无二的特点，识别时可以实现 1:1 匹配，具有极高的安全特性，因此现代往往将这项技术应用在数据安全和财产安全保护中。

目前这项技术应用十分广泛，例如，它用于个人智能设备的语音认证，例如电话，车辆和笔记本电脑，华为公司的小艺语音助手就具有声纹识别功

能。它保证了银行交易和远程付款的交易安全性。它已被广泛应用于判断犯罪嫌疑人是否有罪[4]，或监视和自动身份标记[5]。在基于音频的信息检索方面，它是广播新闻，会议录制和电话呼叫的重要信息。

下表是机器之心网站整理的声纹识别发展历史及参考文献，我选取了部分与本文相关的概念：[6]

年份	事件	相关论文/Reference
1945	提出了“声纹(Voiveprint)”的概念	L.G Kersta. Voiceprint ndentification[J]. Nature. 1962, 196: 1253-1257.
1970	利用语谱图(Spectrogram)对说话人的身份进行判断	Bolt, R. H., Cooper, F. S., David Jr, E.E., Denes, P. B., Pickett, J. M., & tevens, K. N. (1970). Speaker dentification by speech spectrograms: a scientists' view of its reliability for legal purposes. *The Journal of the coustical Society of America*, *47*(2B), 597-612.
1980	S.B.Davis 和 Hermansky 对人耳的听觉特性的分析和研究，并针对性地提出了 Mel 频谱的梅尔倒谱系数 (Mel Frequency Cepstral Coefficients，MFCC)	Davis.S. and Mermelstein. P. Comparison of Parametric Representations for onosyllabic Word Recognition in ontinuously Spoken Sentences[D]. IEEE ransactions on Signal Processing. 1980, 28(4): 357-366.
1994	隐含马尔科夫模型(Hidden Markov Model， HMM)作为概率模型的代表被应用于说话人确认领域，此方法通过对状态的 转移进行描述而在说话人确认领域得到广泛的应用	Matsui, T., & Furui, S. (1994). Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMM's. *IEEE Transactions on speech and audio processing*, *2*(3), 456-459.
1995	Reynolds 提出混合高斯模型 (Gaussian Mixture Model，GMM)来对语音特征的分布建模	D. Reynolds. Speaker Identification and Verification Using Gaussian Mixture Speaker Models[J]. Speech Communication. 1995, 17: 91-108.
2000	其后，Reynolds 针对 GMM 方法在建模过程中对数据量需求大的缺陷，提出先利用通用背景模型 (Universal Background Model，UBM)对所有说话人的语音段特征建模	Reynolds. D.A. Speaker Verification Using Adapted Gaussian Mixture Models[J]. Digital signal processing, 2000,7 (I) :19-41

年份	事件	相关论文/Reference
2007	Kenny 等人提出将人脸识别领域中的概率线性鉴别分 析(Probabilistic Linear Discriminant Analysis，PLDA)应用于说话人识别领域。	Prince S. J. and Elder J. H., Probabilistic Linear Discriminant Analysis for Inferences about Identity[C]. IEEE Intentional Conference Computer Vision, 2007. 1–8.
2011	2011 年提出一种 i-vector 的长度规一化技术(length normalization，LN)，主要针对 i-vector 的分布与 PLDA 中的高斯假设不匹配的问题	Daniel G. R. and Carol Y. Espy-Wilson, "Analysis of i-vector Length Normalization in Speaker Recognition Systems," in INTERSPEECH 2011, Florence, Italy, August, 2011: 3283–3291.
2014	深度神经网络应用到说话人识别领域	Lei, Y., Scheffer, N., Ferrer, L., & McLaren, M. (2014, May). A novel scheme for speaker recognition using a phonetically-aware deep neural network. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*(pp. 1695-1699). IEEE.

总的来说，自从深度学习首次被应用到语音识别任务[7]，相比于传统的高斯混合模型－隐马尔科夫模型(Gaussian mixture model-hidden Markov model,GMM-HMM)语音识别系统获得了超过 20%的相对性能提升。此后，基于深度神经网络(Deep neural networks,DNN)的声学模型逐渐替代了 GMM 成为语音识别声学建模的主流模型，并极大地促进了语音识别技术的发展，突破了某些实际应用场景下对语音识别性能要求的瓶颈，使语音识别技术走向真正实用化。[8]

3. 模型

3.1 数据

本模型中用到的数据集为冲浪科技（www.surfing.ai）提供的免费中文普通话语料库(T-CMDS-20170001_1, Free ST Chinese Mandarin Corpus)，包含 855 个说话者的话语，同时有男声和女声，102600 个话

语，大约 100 余小时。内容以平时的网上语音聊天和智能语音控制语句为主。10 万余条语音文件， 855 个不同说话者，适合多种场景下使用。

3.2 梅尔频谱

为了使用 Tensorflow 训练模型数据，我们需要将难以计算的语音数据转化为可计算的数据。大多数的音频信号的频率成分随着时间变化（非周期性信号），我们通过对信号的多个窗口部分执行 FFT 来计算多个频谱，FFT 是在信号的重叠窗口部分上计算的，这就是频谱图，当信号在不同频率下随时间变化时，这是一种直观地表示信号响度或幅度的方法。研究表明，人类不会感知线性范围的频率。我们在检测低频差异方面要胜于高频。 例如，我们可以轻松分辨出 500 Hz 和 1000 Hz 之间的差异，但是即使之间的距离相同，我们也很难分辨出 10,000 Hz 和 10,500 Hz 之间的差异。

1937 年，Stevens，Volkmann 和 Newmann 提出了一个音高单位，以使相等的音高距离听起来与听众相等，被称为梅尔音阶。我们对频率执行数学运算，以将其转换为 mel 标度。mel 谱图是频率转换为 mel 标度的谱图。使用 python 的 librosa 音频处理库它只需要几行代码就可以实现。[9]

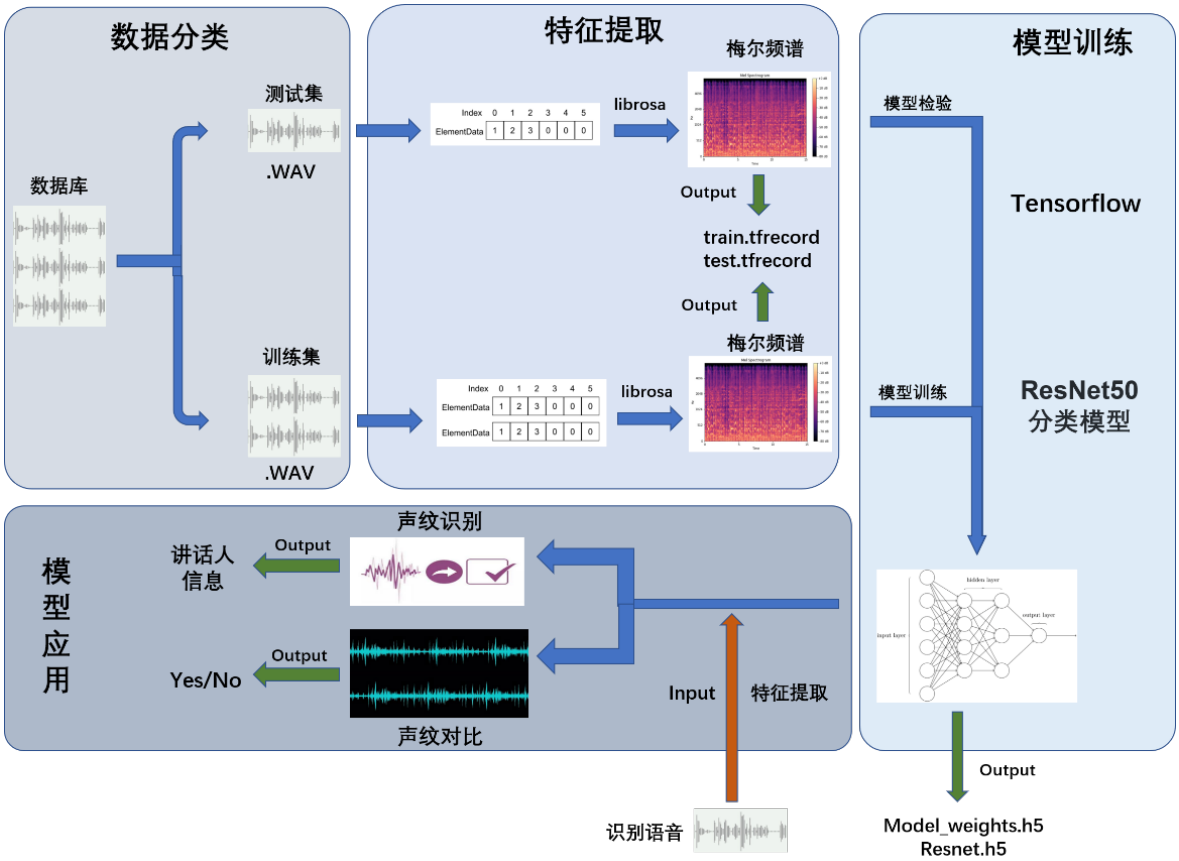

图 1 基于 Tensorflow 的声纹识别流程

3.3 模型

常见的基于深度学习的声纹识别模型主要分成下述四个步骤：数据分类，特征提取，模型训练，最后是模型应用。

如图 1 所示，首先将数据库中的数据按照一定比例分为测试集和训练集（在本模型中按照 1：60 分类，图 1 数据分类），语音数据小而多，我们为了加快训练速度将音频文件生成 TFRecord。之后使用 python 的 librosa 库得到音频的梅尔频谱，（图 1 特征提取），输出 train.tfrecord 和 test.tfrecord，为方便之后的训练，此时将音频数据的梅尔频谱转换为一维 list。接下来搭建一个 ResNet50 分类模型（图 1 模型训练），ResNet 通过残差学习解决了深度网络的退化问题，可以训练出更深的网络，这个模型在多个领域尤其是图

像识别中表现出极好的表现，甚至刷新了 CNN 模型在 ImageNet 上的历史。由于本模型训练需要较高的硬件条件，我先后使用本机和云服务器训练模型，但均因为内存和速度放弃，后选用 Google 的 Colab 的 GPU 节点训练，速度得到明显提升，并最终训练出模型，每训练 200 个 batch 执行一次测试和保存模型，包括预测模型和网络权重（Model_weights.h5,Resnet.h5），直到准确率达到 90%以上（图 2）。最后实现模型应用（图 1 模型应用），我们输入两个语音，通过预测函数获取他们的特征数据，使用这个特征数据可以求他们的对角余弦值，得到的结果可以作为他们相识度。(图 3) 在上面的声纹对比的基础上，我们可以实现声纹识别。同样是使用上面声纹对比的数据加载函数和预测函数，通过这两个同样获取语音的特征数据。首先加载语音库中的语音数据，这些音频就是相当于已经注册的用户，他们注册的语音数据会存放在这里，如果有用户需要通过声纹登录，就需要拿到用户的语音和语音库中的语音进行声纹对比，如果对比成功，那就相当于登录成功并且获取用户注册时的信息数据。完成识别的主要在 recognition()函数中，这个函数就是将输入的语音和语音库中的语音一一对比。

图 2 使用 Colab 训练模型，准确率超过 90%

图 3 声纹对比结果

4. 结果与展望

在模型训练时记录了每次训练迭代的准确率，如图四，训练过程是反复中提升的，最终准确率达到 0.9 以上，计算过程漫长，即使使用 Google 的 Colab 计算资源，训练时间仍然持续了大于 7 小时，使用 1 核 2G 的云服务器运行三天才完成训练过程 50%，最终因内存不够被迫中止，如图 5。经过随机选取数据库中的语音进行声纹对比，本模型可基本满足声纹识别需求。其实，本校具有超算资源，我也尝试过在本校超算平台上训练模型，但由于时间限制，python 并行计算没有弄清楚，加之学校超算资源为付费节点，权衡之下，使用 Google Colab 的免费 GPU 资源是更好的选择。

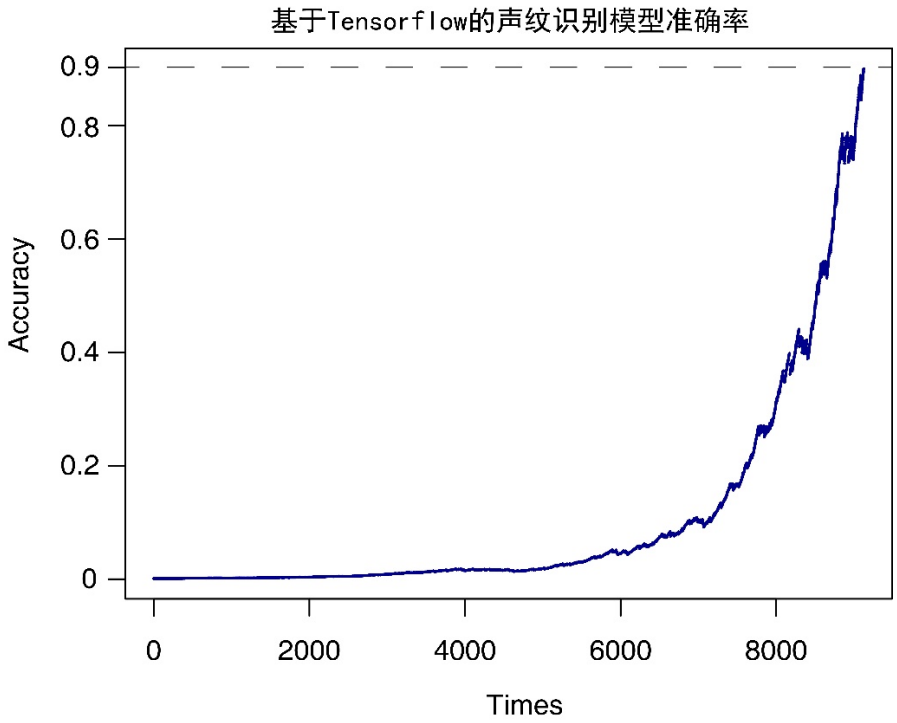

图 4 训练迭代准确率

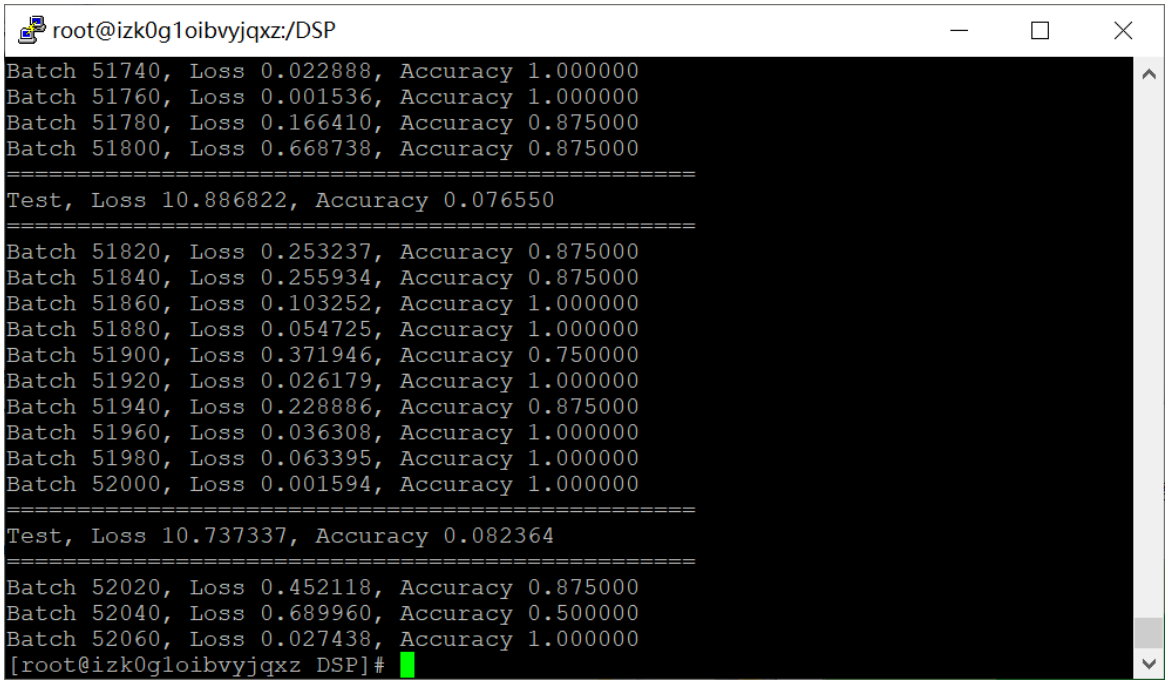

图 5 使用 1 核 2G 服务器训练模型

本模型使用单 GPU 训练模型，由于单 GPU 的计算能力有限，很难处理海量数据（本模型使用数据包含 855 人），所以探究如何进行多 CPU 或者 GPU 并行计算是深度学习模型的一个需要突破的点。关于这方面的研究首先是分数据的策略，在文献[10]中提出将训练数据分成很多小份，然后每份在一个单独的 GPU 上进行运算，将得到的梯度求平均去更新模型。这种方法受限于不同机器间的频繁交互会导致通信代价很高，从而没法带来很大的训练速度提升。文献[11]提出将原始数据平均分成 N 份，然后每份数据利用一台机器单独训练一个子网络，每次迭代后将这些子网络求平均得到一个总模型，再分到各个机器上进行训练。这种方式可以有效避免机器之间的通讯代价，但是会导致较大的性能损失。由于机器之间的通讯代价是并行计算的一个瓶颈，文献[12]提出异步随机梯度下降可以有效地掩蔽通讯代价，利用包含数千个 CPU 的集群来进行 DNN 的并行训练。而文献[13]将这种方法扩展到了 GPU 上，利用多 GPU 进行并行化训练，节约了设备成本。

5.参考文献

1. 郑方，李蓝天，张慧，艾斯卡尔·肉孜: **声纹识别技术及其应用现状 %J 信息安全研究**. 2016, **2**(01):44-57.
2. Biometrics,wiki [https://en.wikipedia.org/wiki/Biometrics]
3. 张陈昊，郑方，王琳琳: **基于多音素类模型的文本无关短语音说话人识别 %J 清华大学学报(自然科学版)**. 2013, **53**(06):813-817.
4. Campbell JP, Shen W, Campbell WM, Schwartz R, Bonastre JF, Matrouf D: **Forensic speaker recognition**. *IEEE Signal Processing Magazine* 2009, **26**(2):95-103.
5. Togneri R, Pullella D: **An Overview of Speaker Identification: Accuracy and Robustness Issues**. *IEEE Circuits Syst Mag* 2011, **11**(2):23-61.
6. **机器之心** [https://www.jiqizhixin.com/graph/technologies/5b4fcd3d-d8d2-451f-9793-c2aac880bb6b]
7. Mohamed A-r, Sainath TN, Dahl G, Ramabhadran B, Hinton GE, Picheny MA, Ieee: **DEEP BELIEF NETWORKS USING DISCRIMINATIVE FEATURES FOR PHONE RECOGNITION**. In: *2011 Ieee International Conference on Acoustics, Speech, and Signal Processing.* 2011: 5060-5063.
8. 戴礼荣，张仕良，黄智颖: **基于深度学习的语音识别技术现状与展望 %J 数据采集与处理**. 2017, **32**(02):221-231.
9. understaning the mel spectrogram [https://medium.com/analytics-vidhya/understanding-the-mel-spectrogram-fca2afa2ce53]
10. Vesely K, Burget L, Grezl F: **Parallel Training of Neural Networks for Speech Recognition**. In: *Text, Speech and Dialogue.* Edited by Sojka P, Horak A, Kopecek I, Pala K, vol. 6231; 2010: 439-446.
11. Park J, Diehl F, Gales MJF, Tomalin M, Woodland PC, Isca-Inst Speech Commun A: **Efficient Generation and Use of MLP Features for Arabic Speech Recognition**; 2009.
12. Le QV, Ieee: **BUILDING HIGH-LEVEL FEATURES USING LARGE SCALE UNSUPERVISED LEARNING**. In: *2013 Ieee International Conference on Acoustics, Speech and Signal Processing.* 2013: 8595-8598.
13. Zhang S, Zhang C, You Z, Zheng R, Xu B, Ieee: **ASYNCHRONOUS STOCHASTIC GRADIENT DESCENT FOR DNN TRAINING**. In: *2013 Ieee International Conference on Acoustics, Speech and Signal Processing.* 2013: 6660-6663.