Computer Science and Application 计算机科学与应用, 2021, 11(9), 2342-2349
Published Online September 2021 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2021.119239

Hans汉斯

基于卷积神经网络的声纹检测签到系统

梁景泉，周子程，刘秀燕*

青岛理工大学信息与控制工程学院，山东 青岛
Email: ljingquan@outlook.com, *liuxiuyan_ok@163.com

收稿日期：2021年8月14日；录用日期：2021年9月10日；发布日期：2021年9月17日

摘　要

针对高校传统考勤方式，如人工点名、手写签到等方式存在他人代替、耗时且效率低下等问题，基于深度学习强大的建模能力，本项目提出基于改进神经网络模型的智能化课堂语音签到系统，采用卷积神经网络(Convolutional Neural Network, CNN)进行语音模型的训练，自动提取语音深层次的声纹特征并识别，实验结果表明该系统能有效提高点名效率并能够制止代签等行为，具有有效提高教师的课堂教学效率的重要意义。

关键词

卷积神经网络，声纹识别，签到系统

Voiceprint Detection Sign in System Based on Convolutional Neural Network

Jingquan Liang, Zicheng Zhou, Xiuyan Liu*

College of Information and Control Engineering, Qingdao University of Technology, Qingdao Shandong
Email: ljingquan@outlook.com, *liuxiuyan_ok@163.com

Received: Aug. 14th, 2021; accepted: Sep. 10th, 2021; published: Sep. 17th, 2021

Abstract

Aiming at the problems of substitution, time-consuming and low efficiency of traditional attendance methods in colleges and universities, such as manual roll call and handwritten check-in, based on the powerful modeling ability of deep learning, this project proposes an intelligent classroom voice check-in system based on improved neural network model, which uses convolu-

*通讯作者。

文章引用: 梁景泉, 周子程, 刘秀燕. 基于卷积神经网络的声纹检测签到系统[J]. 计算机科学与应用, 2021, 11(9): 2342-2349. DOI: 10.12677/csa.2021.119239

tional neural network (CNN) to train the voice model. The experimental results show that the system can effectively improve the roll call efficiency and stop the signing behavior, which is of great significance to improve the efficiency of teachers' classroom management.

Keywords

Convolutional Neural Network, Voiceprint Recognition, Sign-In System

1. 引言

随着互联网技术的飞速发展，广大高等院校引入了智能化办公系统提高工作效率[1]，同时也将大量的课堂点名系统投入使用，如使用二维码进行身份识别的签到系统[2]，通过在微信群发言来识别身份的签到系统等[3]。这些系统的出现大大提高了学校学生们的出勤率，保证了学校的教学质量。但是通过对现有系统的进一步研究，我们能够发现这些签到系统在进行身份识别的时候仍然使用着比较传统的方式，这些传统的方式虽然提高了效率，但是都具有一定的局限性，他们仅仅只能通过账号或扫码识别身份，这就导致如果账号被集中到一个学生手上或者二维码图片被发送，就能完成大量的签到作弊行为。虽然有的系统引入 GPS 定位来进行签到时的实时定位[4]，但是这仍然无法完全制止签到作弊的行为，而这种签到作弊行为对学校教学产生了很大影响。因此，广大高校现在需要一种更加完备的方法来完成课堂考勤，以便进一步提高教学质量。随着深度神经网络等人工智能技术的快速发展，深度学习算法作为新一代人工智能技术的核心算法，为数据挖掘、模式识别、计算机视觉、语音识别、自然语言处理等领域带来了颠覆性变化[5]。

传统的考勤方式，如学生自行刷卡签到，教师人工点名等方式存在他人代替、卡片遗失和卡片被盗用等风险，且因为缺乏准确的模型，无法处理数量级较大的数据集，而深度学习的兴起提供了强大的建模能力，为进行语音识别奠定了技术基础，提高了初始数据的利用效率。早期由于硬件与软件条件的限制，国内对语音识别的研究开始得比较晚，但是引入技术之后发展的速度特别快。文献[6]针对语音识别率低以及噪声影响等问题，设计了结合经验模态分解和 RBF 的语音识别模型，使得实验结果受噪音的影响达到更低，最终的仿真实验也验证了与其他算法相比该算法进一步提高了语音识别率。文献[7]提出一种基于深度神经网络的麦克风阵列降噪算法，这种算法能够有效提高真实噪声环境下的语音识别率。文献[8]提出基于深度置信网络隐马尔可夫混合模型(DBN-HMM)无监督语音签到系统，但实验结果并没有获得识别率在 95%以上理想的 DBN-HMM 语音识别模型，如何在更复杂的环境干扰下，尽可能使用最少的语音数据学习训练仍然存在一定挑战。文献[9]提出基于高斯混合模型(GMM)课堂语音签到器，但声纹的不固定性影响了识别效率，如何以一种较为准确的方法识别语音仍是一个较难攻克的课题。国外语音识别的研究始于 1952 年，世界上第一个语音识别系统是 Audrey 系统[10]，该系统可以识别 10 个英文字母，自此开辟了语音识别研究之路。1980 年后，研究员们有了更大的目标，研究方向逐渐偏向更大词汇量的语音识别系统上。同时随着计算机的发展和神经网络技术的出现，识别算法的技术开始过渡到统计模型，隐马尔科夫模型(HMM)更是成了当时的主流技术[8] [11]。文献[12]提出基于时域的无监督单通道语音源分离方法，通过将语音特定信息与经验模式分解相结合，可产生语

音混合的单通道分离的优越结果。文献[13]提出基于卷积神经网络(CNN)模型的语音情感识别。文献[14]提出一种有监督的机器学习方法来解决谷歌错误识别语音的情况，采用支持向量机(SVM)和卷积神经网络(CNN)模型的校正系统，结果显示使用支持向量机和 CNN 机器学习算法显著提高了语音命令的识别能力，使得越南语语音命令识别错误率由 35.06%降低到 7.08%，但是国内外的语音识别系统在低信噪比的环境下都会面临识别性急剧下降的问题。

本文通过结合签到系统与声纹识别技术，首先通过 MFCC 提取训练音频特征信号，归一化处理后，输入卷积神经网络进行语音模型的训练，实际音频经相同方法处理后输入卷积神经网络进行声纹识别，与声纹库中的数据进行比对，输出最大匹配结果，最终进行签到检验。

2. 卷积神经网络

卷积神经网络(Convolutional Neural Network, CNN)是深度学习中十分常用的一种神经网络模型[15]，作为深度学习中三大神经网络模型之一，与深层神经网络以及循环神经网络并列。在其诞生之初，卷积神经网络常被用来处理图像相关问题[16]，因此它的输入层数据也常常与图像有关，而各个节点不同类别的置信度则是由输出层输出。LeNet-5 (图 1 所示)是最早的基本卷积神经网络[17]，包含 CNN 的基本单元：输入层(INPUT)、卷积层(C1, C3, C5)、池化层(S2, S4)、全连接层(F6)以及径向基层(OUTPUT)。输入层(Input Layer)是模型的第一层用于接收从模型输入的数据。卷积层(Convolutions Layer)又称特征提取层是本网络中的关键模块，主要用于分块处理网络中的数据，提取抽象特征。池化层(Pooling Layer)又称下采样层，一般用于压缩图片信息，在两个卷积层之间进行降维，来达到减少模型参数量的目的。全连接层(Fully Connected Layers)通常在整个神经网络的尾部担任分类任务，接收上层处理的数据信息，最终给出结果。径向基层(Radial Basis Layer)又称输出层，主要用于计算非线性激活函数和输出结果。

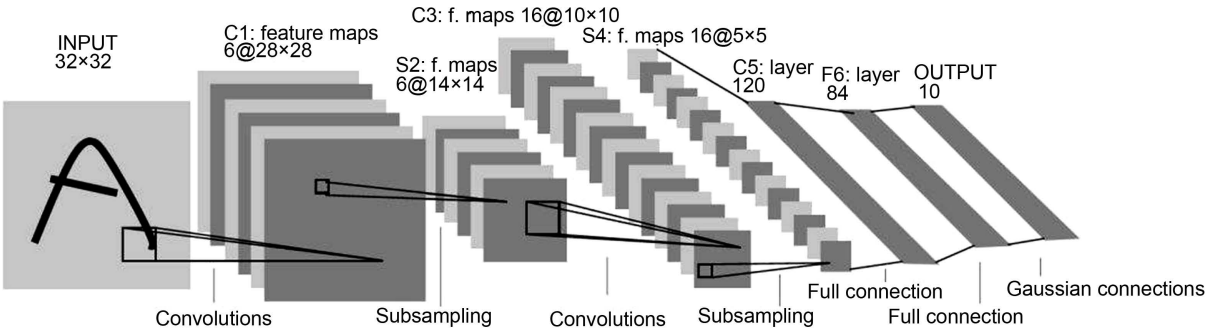

Figure 1. LeNet-5 network structure
图 1. LeNet-5 网络结构

3. 声纹识别技术

声纹是指能唯一识别某人或某物的声音特征。声纹识别流程如图 2 所示：

步骤 1：训练音频中说话人的声纹注册即进行有效音频的提取；

步骤 2：提取出声纹特征信息建立声纹库；

步骤 3：当实际进行声纹检测时，进行相同步骤提取出说话人的声纹特征信息；

步骤 4：将提取到的信息与数据库中的信息进行特征比对，得到结果。

时至今日，声纹识别技术的应用已经十分广泛，如公安领域[18]、车载语音交互[19]、家居场景[20]等都能见到该技术的出现。

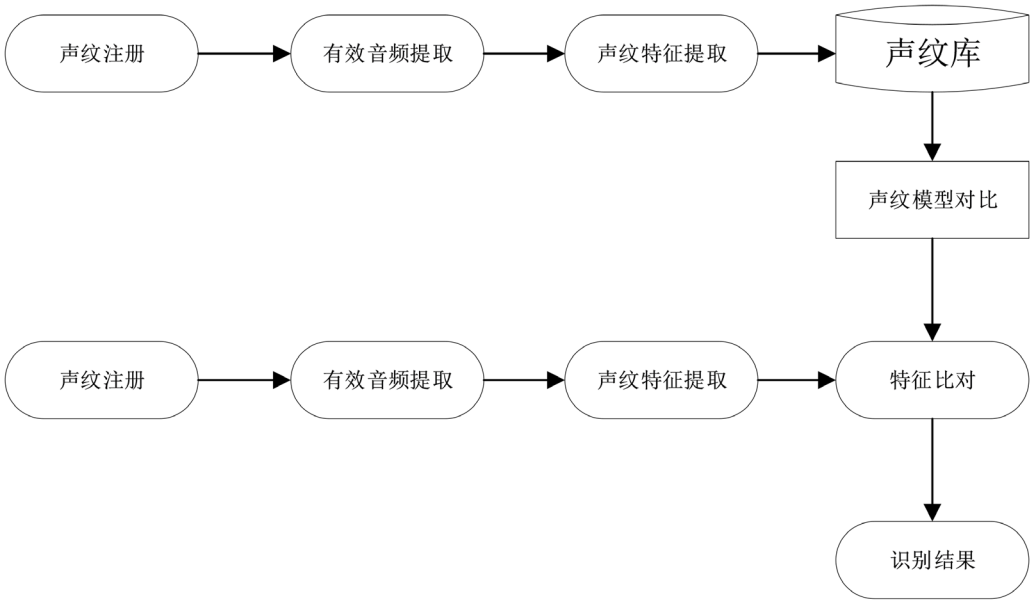

Figure 2. Flow chart of voiceprint recognition technology
图 2. 声纹识别技术流程图

随着深度学习的不断发展，越来越多的深度学习理论与方法被应用到声纹识别领域当中，其中一个非常具有代表性的算法就是 CNN。卷积神经网络因为具自学习和自组织能力、很强的复杂分类边界区分能力以及对不完全或不准确信息的抗干扰性，在训练过程中能不断调整自身的参数权重和具有拓扑结构的优点，在声纹识别当中被频繁的使用[21]。

4. 改进的卷积神经网络模型

深度学习模型因具有较强的非线性描述能力，更好的泛化能力，更小的系统参数变化敏感性等优势。因此，本文提出基于深度学习技术提取语音声纹特征，将已提取出的不同人的语音信号特征归一化后的数据输入神经网络，网络在学习过程中使用反向误差传递算法对深度神经网络进行训练，使得该网络具有识别预测能力。

网络结构如图 3 所示，包含两层卷积层，特征图每个单元计算如式(1)所示，两层池化层(见式(2))，两层全连接层(见式(3))，为增加网络的非线性分割能力，激活函数选择 softmax 函数(见式(4))，对全连接层得到的多个输出，映射到(0,1)区间内，做归一化，使得所有元素的和累加起来等于 1，可以直接当作概率对待，选取概率最大的分类作为识别出的说话人。

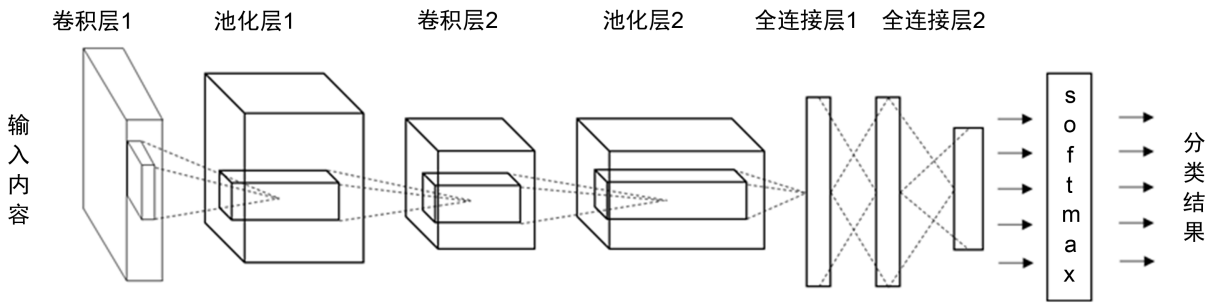

Figure 3. Improved network model
图 3. 改进网络模型

$$q_{j,m} = \sigma\left(\sum_i^I \sum_{n=1}^F O_{i,n+m-1}\omega_{i,j,n} + \omega_{n,j}\right) \tag{1}$$

其中 $O_{i,m}$ 表示第 i 个输入层特征图 O_i 的第 m 个窗口，$q_{j,m}$ 表示第 m 个卷积层特征图的第 j 个窗口，$\omega_{i,j,n}$ 是指第 n 个元素的权重，$\omega_{i,j}$ 是输入层特征图与卷积层特征图的连接，F 指卷积核数量。

$$p_{i,m} = \max_{n=1,\cdots,G} q_{i,(m-1)\times s+n} \tag{2}$$

其中 $p_{i,m}$ 是第 m 个池化层特征图的第 i 个窗口，G 是池化窗口的大小，s 的大小决定了相邻池化窗口的重叠。

$$z_i = \omega_j \times x + b \tag{3}$$

其中 ω_j 是第 j 个全连接层特征图的权重，x 是全连接层的输入，b 是偏置项。

$$S(y_i) = \frac{e^{y_i}}{\sum_j e^{y_j}} \tag{4}$$

其中 y_i 表示第 i 个神经元的输出，y_j 表示第 j 个神经元的输出，j 为神经元的总个数。

5. 基于改进 LeNet5 的语音签到系统

签到过程流程图如图 4 所示：

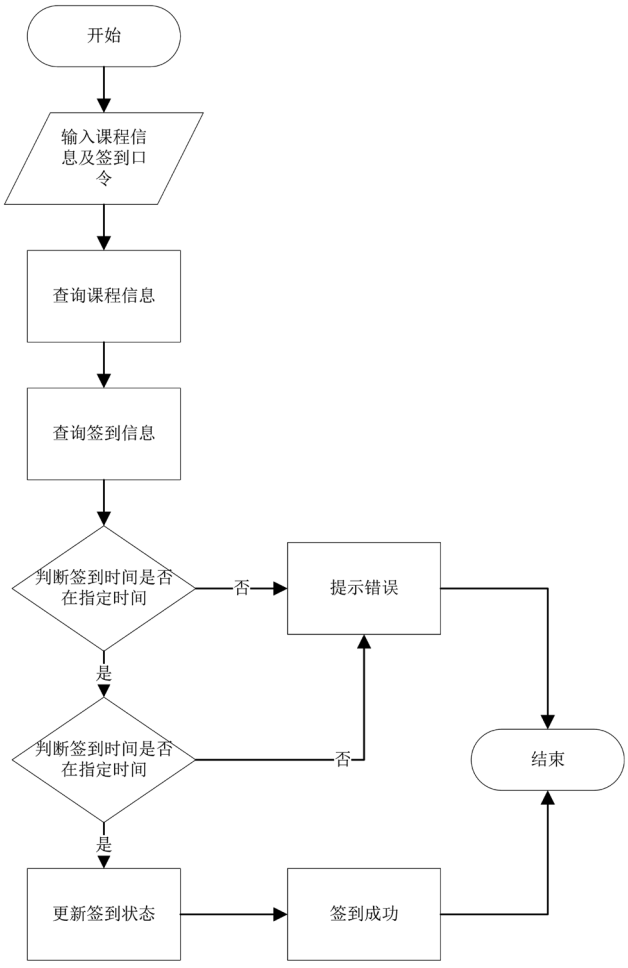

Figure 4. Check in flow chart
图 4. 签到流程图

步骤 1：学生在相关课程界面进行口令签到；

步骤 2：查找签到课程信息；

步骤 3：查找签到场次信息；

步骤 4：检查是否处于合法签到时间；

步骤 5：若不在签到时间，提示错误，若在签到时间，判断签到声纹是否成功匹配；

步骤 6：若签到不成功，提示错误，若成功匹配，更新数据库状态为签到成功。

学生进入课程签到页面之后，不仅可以进行当次签到，还可以查看之前的历史记录，避免意外错过签到。

在实现声纹识别时第一步需要构建系统的声纹数据库，因签到口令较短且不复杂，可以在安静环境下自行使用麦克风等输入设备录制。录制完成后，进行相关数据处理操作，包括降噪和标注标签分类等。完成准备工作之后，就可以开始进行训练。获取到反映签到人声纹区别的时域频域特征，通过模式识别方法进行匹配，最终确定签到者的身份信息。

语音识别流程如图 5 所示：

步骤 1：提取音频信号的特征向量数据，通过调用 MFCC 获取处理语音信号，存储获取到的 13 维特征向量数据；

步骤 2：在完成提取之后，将提取出的特征向量进行归一化处理然后输入 CNN 模型(本文使用 Conv1D)；

步骤 3：利用不同尺度卷积核对不同人的声纹特征进行提取并融合，进行不断的迭代训练；

步骤 4：当训练完成时，我们可以观察到损失函数数值趋近于稳定；

步骤 5：根据 CNN 网络输出的每名学生的特征向量，与数据库中的声纹特征进行匹配，实现语音签到。

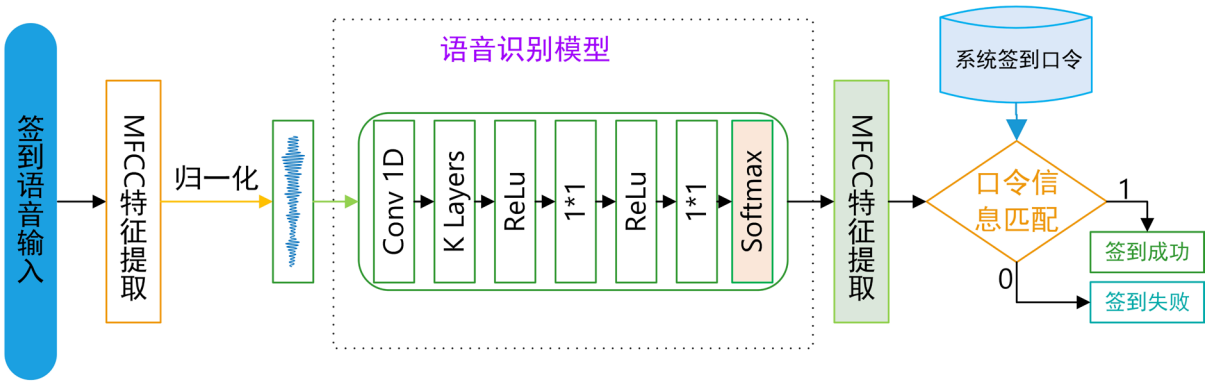

Figure 5. Speech recognition CNN network model structure
图 5. 语音识别 CNN 网络模型结构

将训练好的模型作为声纹数据库保存，以便实际应用时与实时输入语音进行比对，完成系统功能，并最终将识别返回到签到系统，判断该学生签到是否成功，以进行相应系统考勤操作。

6. 实验结果

采用的数据为使用麦克风自行实际录制的音频，数据及其预处理情况如表 1 所示。

在本次实验中除了改进 LeNet-5 网络之外，也使用了 SVM 方法和原始 LeNet-5 网络进行实际测试，经训练集训练之后，各个网络对测试集的识别准确率如图 6 所示，可以看出改进 LeNet-5 网络的识别率相较于原始 LeNet-5 与 SVM 方法有了明显提升。改进 LeNet-5 对两组数据集的识别准确率并无较大差异，这说明数据集中的噪音对其识别过程影响较小，该方法具有较强的抗干扰能力。

Table 1. Experimental data and pretreatment
表 1. 实验数据及预处理情况

数据集标记	数据集组成	特点	训练集:测试集
DatasetA	30 人共 300 段	安静环境，无噪音	8:2
DatasetB	20 人共 200 段	少量噪音存在	8:2

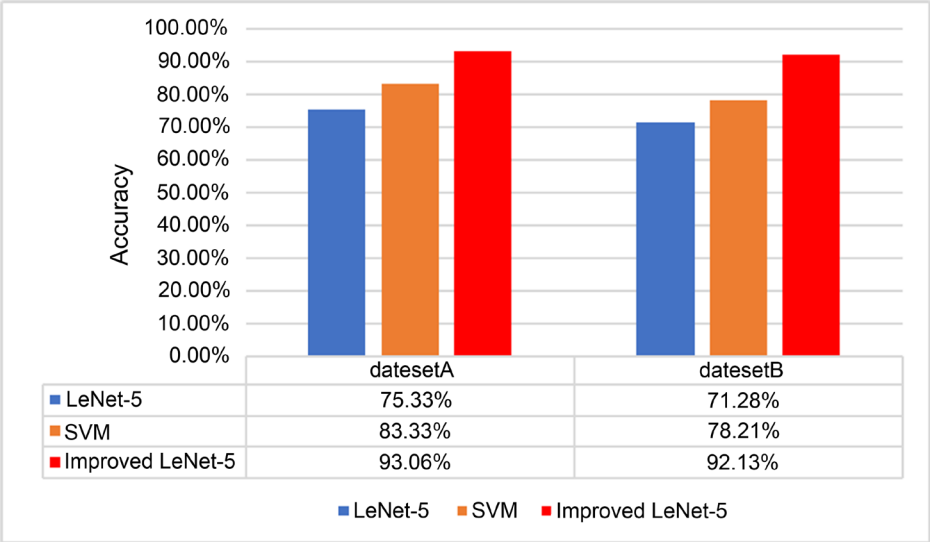

Figure 6. Speech recognition CNN network model structure
图 6. 语音识别 CNN 网络模型结构

7. 结束语

针对传统高校点名方式存在大量代签的问题，本文提出结合传统签到系统与声纹识别方法，利用卷积神经网络提取音频声纹特征信号进行音频库的训练与实际签到中的识别，使得噪音对识别结果的影响尽可能减小，成功避免了其他方法造成的代签等问题，对提高教学效率有着深远的意义。

然而，因为人工采集实际情况限制，用于训练的音频数量仍存在不足，在今后的研究当中，会进一步获取更多的更大数量级的音频，以达到更加优秀的识别准确度。

基金项目

山东省大学生创新创业训练计划项目(S202010429018，S202010429208)，国家自然科学基金项目(62001262)，山东省自然科学基金项目(ZR2020QF008)。

参考文献

[1] 胡晶晶, 周斌. 高校行政办公自动化系统建设与发展的研究[J]. 数码世界, 2020(12): 235-236.

[2] 王润泽, 王亮, 刘涛, 栗斌. 考虑实时路况反馈的动态路径规划算法研究[J]. 测绘科学, 2020, 45(7): 163-169.

[3] 林志伟, 王庆九, 马超虹, 谢礼礼, 吴森洋. 基于 itchat 的微信群签到系统开发[J]. 实验室研究与探索, 2020, 39(1): 108-115.

[4] 王健杰, 汪鹏程, 范祥林, 郁书好. 基于 Android 的企业 GPS 签到 APP 的设计与实现[J]. 计算机科学与应用, 2019, 9(7): 1352-1357. https://doi.org/10.12677/CSA.2019.97152

[5] 张丹. 深度学习神经网络在语音识别中的应用探讨[J]. 电子世界, 2021(6): 67-68.

[6] 施巍巍. 经验模态分解方法及其在语音识别算法中的研究[D]: [硕士学位论文]. 杭州: 浙江理工大学, 2014.

[7] Zhang, X., Wang, Z. and Wang, D. (2017) A Speech Enhancement Algorithm by Iterating Single- and Multi-Microphone Processing and Its Application to Robust ASR. 2017 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), New Orleans, LA, 5-9 March 2017, 276-280. https://doi.org/10.1109/ICASSP.2017.7952161

[8] 赵从健, 雷菊阳, 李明明. 基于无监督学习的语音签到系统[J]. 软件, 2019, 40(12): 183-187.

[9] 朱泰, 管震臻, 李新战, 等. 基于语音识别的课堂签到器[J]. 科技资讯, 2018, 16(10): 32-33.

[10] 范长青. 小词汇量非特定人连续语音识别系统的研究[D]: [硕士学位论文]. 沈阳: 沈阳理工大学, 2008.

[11] Rabiner, L.R. (1978) On Creating Reference Templates for Speaker Independent Recognition of Isolated Words. *IEEE Transactions on Acoustics Speech and Signal Processing*, **26**, 34-42. https://doi.org/10.1109/TASSP.1978.1163037

[12] Prasanna Kumar, M.K. and Kumaraswamy, R. (2017) Single-Channel Speech Separation Using Empirical Mode Decomposition and Multi Pitch Information with Estimation of Number of Speakers. *International Journal of Speech Technology*, **20**, 109-125. https://doi.org/10.1007/s10772-016-9392-y

[13] Issa, D., Demirci, M.F. and Yazici, A. (2020) Speech Emotion Recognition with Deep Convolutional Neural Networks. *Biomedical Signal Processing and Control*, **59**, Article ID: 101894. https://doi.org/10.1016/j.bspc.2020.101894

[14] Nguyen, Q.H. and Cao, T.D. (2020) A Novel Method for Recognizing Vietnamese Voice Commands on Smartphones with Support Vector Machine and Convolutional Neural Networks. *Wireless Communications and Mobile Computing*, **2020**, Article ID: 2312908. https://doi.org/10.1155/2020/2312908

[15] Tang, J., Deng, C. and Huang, G.B. (2015) Extreme Learning Machine for Multilayer Perceptron. *IEEE Transactions on Neural Networks and Learning Systems*, **27**, 809-821. https://doi.org/10.1109/TNNLS.2015.2424995

[16] Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012) Imagenet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, **25**, 1097-1105.

[17] LeCun, Y., Bottou, L., Bengio, Y., *et al.* (1998) Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, **86**, 2278-2324. https://doi.org/10.1109/5.726791

[18] 刘晓晨, 潘孝勤, 曹金璇, 芦天亮. 声纹识别和语音识别技术在公安领域的应用[J]. 网络安全技术与应用, 2021(4): 153-155.

[19] 李育贤, 李玓. 声纹识别技术在车载语音交互中的应用前景[J]. 汽车工业研究, 2021(1): 30-32.

[20] 张金鑫. 面向家居场景的声纹识别关键技术研究[D]: [硕士学位论文]. 南京: 南京邮电大学, 2020.

[21] 白莹, 赵振东, 戚银城, 王斌, 郭建勇. 基于小波神经网络的与文本无关说话人识别方法研究[J]. 电子与信息学报, 2006, 28(6): 1036-1039.