基于双微麦克风阵列与 Wide ResNet 网络的语音命令词识别

祁潇潇 曾庆宁 赵学军*
(桂林电子科技大学信息与通信学院 广西 桂林 541004)

摘 要 为了提高噪声环境下语音识别的稳健性[1],提出宽残差深度神经网络的语音识别算法。该算法结合双微麦克风阵列系统、语音数据集为双微麦克风数据集,使用功率归一化倒谱系数作为特征参数输入到残差网络中进行训练。实验表明,与 ResNet15 模型、ResNet18 模型相比,只有三个残差模块的宽残差网络在噪声环境下语音命令词的识别和内外部说话人检测任务中具有较高的准确度,均达到了 95% 以上。

关键词 语音识别 宽残差神经网络 功率归一化倒谱系数 双微麦克风阵列

中图分类号 TP391.42 **文献标志码** A DOI:10.3969/j.issn.1000-386x.2024.05.020

SPEECH COMMAND WORD RECOGNITION BASED ON DUAL MICRO MICROPHONE ARRAY AND WIDE RESNET

Qi Xiaoxiao Zeng Qingning Zhao Xuejun*
(*School of Information and Communication*, *Guilin University of Electronic Technology*, *Guilin 541004*, *Guangxi*, *China*)

Abstract In order to improve the robustness of speech recognition in noise environment, a speech recognition algorithm based on wide residual deep neural network is proposed. The algorithm combined the dual micro microphone array system, and the voice data set was the dual micro microphone data set. The power normalized cepstrum coefficient was used as the characteristic parameter to input into the residual network for training. Experimental results show that, compared with the Resnet15 model and Resnet18 model, the wide ResNet with only three residual modules has higher accuracy in the recognition of speech command words and the internal and external speaker detection task under noise environment, both reaching more than 95%.

Keywords Speech recognition Wide ResNet Power normalized cepstrum coefficient Dual micro microphone array

0 引 言

近几年人工智能发展迅速,语音识别技术也得到大幅度提升,智能语音产品也随处可见,如智能语音助手、智能音箱、车载导航等语音控制产品[2]。提高语音识别技术的性能一直是研究人员的目标。语音命令词识别系统是大多数智能语音交互的第一步,也是开启语音设备的常用功能。当识别系统检测到唤醒短语,那么该设备将会被唤醒,从而与服务器连接,通过服务器端的自动语音识别系统应答。一个语音命令识别系统的评价指标主要有唤醒率、虚警率、实时率和功耗水平等。因此在提高检测精度的同时减少误报率,降低功耗并且优化系统减少响应时间是研究人员关注的重点,也是提升用户对于智能语音产品体验的一个着眼点[3]。

传统的命令词识别算法主要是三种:动态时间规整(Dynamic Time Warping,DTW)算法[4]、隐马尔可夫模型(Hidden Markov Model,HMM)[5]、基于非参数模型的矢量量化(VQ)[6]的方法。DTW 算法用于模板匹配[7],HMM 算法是对关键字和背景进行建模[8],基于 VQ 算法的语音识别的思路是在训练过程中首先提取特征,然后用语音特征训练码本,识别过程中,提取待识别语音特征与码本匹配[9]。

近年来,神经网络理论不断丰富,计算机算力不断

收稿日期:2021 - 01 - 20。国家自然科学基金项目(61961009);广西自然科学基金重点项目(2016GXNSFDA380018);广西无线宽带通信与信号处理重点实验室基金项目(GXKL06200107)。祁潇潇,硕士,主研领域:语音信号处理。曾庆宁,教授。赵学军,高工。

提升,大量开源的语音数据集给高精度语音识别提供了便利。目前,众多的专家学者已经提出了大量基于神经网络的语音识别算法。在文献[10]中,训练一个深度神经网络来直接预测关键字或关键字的子字单元,然后通过后置处理方法产生最终的置信度。关键词识别结果相对于基于隐马尔可夫模型的系统性能提升39%。文献[11]提出用一维卷积门控神经网络与 CTC 算法相结合的语音识别模型。该模型与基线模型相比,在性能上有了明显提升,字错误率降低了 3.3% 以上。

在众多专家学者的研究下,神经网络在语音识别中的应用日趋成熟。一般来说,一个语音命令词系统不是针对特定人设计的,也就是说任何用户都可以触发,但是当人们希望自己发出的命令不受外界命令干扰时,这就要求命令识别系统能够对内部和外部的说话人进行区分。由此,本文提出一种结合双微麦克风阵列与宽残差神经网络的语音命令词识别方法,在宽残差神经网络中构建内外部说话人检测和命令词识别的双系统,实验表明,该算法可以减少外部说话人的干扰,同时具有较高的命令词识别准确率。

1　残差神经网络

1.1　深度残差神经网络

如今,基于神经网络的语音识别算法层出不穷,但出于对语音命令词系统诸多限制的考虑,选取卷积神经网络(CNN)加以研究,因为 CNN 可以从经过少量预处理,甚至原始数据中学习到抽象的、本质的和高阶的特征[12],模型复杂度低,权值参数较少。

卷积神经网络发展到今天,网络的层数越来越多[13]。层数越多表明网络提取到的信息就越多,训练精度就越高。但是网络的深度并不是越深越好,网络的层数增加到一定的数量,训练精度反而降低,而且网络的参数也难以优化,这就是网络的退化现象[14]。为了解决这个问题,深度残差神经网络应运而生。其特点是简单高效,能够解决网络加深后网络性能退化问题[15]。

深度残差神经网络的设计思路是,在一个网络中,如果前 n 层已经达到最优,那么继续增加第 $n+1$ 层可能会引起网络退化。因此,将第 $n+1$ 层设计成恒等映射,这样即使增加网络层数也不会引起网络退化现象。

1.2　深度残差神经网络中的恒等映射

深度残差神经网络由一个个残差模块堆叠而成,每个残差模块的一般形式可以表示为:

$$y_l = h(x_l) + F(x_l, W_l) \tag{1}$$

$$x_{l+1} = f(y_l) \tag{2}$$

式中:x_l 与 x_{l+1} 为残差模块的输入与输出;F 表示残差模块;$h(x_l) = x_l$ 为一个恒等映射;f 表示 ReLU 激活函数。图 1 表示基本残差模块。

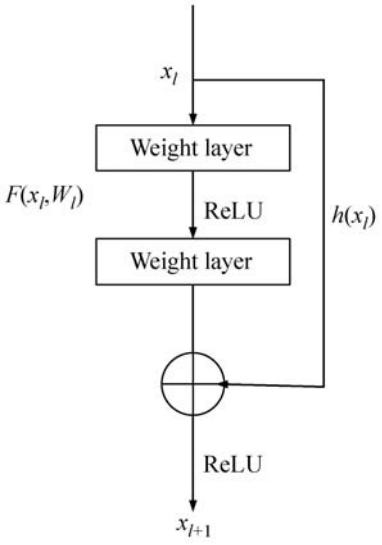

图 1　残差模块结构

为了简化推导公式过程,令 f 也为一个恒等映射:$x \equiv y$,则将式(2)和式(1)联立,便可得到:

$$x_{l+1} = x_l + F(x_l, W_l) \tag{3}$$

残差网络学习的目标是残差函数 $F(x_l, W_l) = x_{l+1} - x_l$,如果 $F(x_l, W_l) = 0$,那么就近似一个恒等映射,并且不改变网络的参数数量和网络的计算复杂度,那么根据式(3)进行递归,得到:

$$x_{l+2} = x_{l+1} + F(x_{l+1}, W_{l+1}) = x_l + F(x_l, W_l) + F(x_{l+1}, W_{l+1}) \tag{4}$$

经过逐层归纳可得:

$$x_L = x_l + \sum_{i=l}^{L-1} F(x_i, W_i) \tag{5}$$

式(5)表示深层的网络可以由浅层网络和残差函数构成,并且表现出了良好的反向传播特性。令损耗函数记为,反向传播的链式法则数学表达式如下:

$$\frac{\partial \varepsilon}{\partial x_l} = \frac{\partial \varepsilon}{\partial x_L} \times \frac{\partial x_L}{\partial x_l} = \frac{\partial \varepsilon}{\partial x_L} \times \left(1 + \frac{\partial}{\partial x_L} \sum_{i=l}^{L-1} F(x_i, W_i)\right) \tag{6}$$

从式(6)可以看出梯度可以拆分为两项,其中一项可以直接传播信息并且没有触及任何权重层。一般来说,在样本较少情况下,$\frac{\partial}{\partial x_l} \sum_{i=l}^{L-1} F(x_i, W_i) \equiv -1$ 的概率几乎为 0,这表明只要权重存在,梯度就存在,梯度信息能够根据反向传播算法传回较浅的网络层。

1.3　Wide ResNet(WRN)模型结构

按照标准的残差网络结构,WRN[16]的残差模块增加了宽度。对于卷积层来说,宽度指的是输出维度;对于一个网络来说,宽度指的是所有层数的总体输出维度。图 2 为 WRN 残差模块的基本结构。图中的 k 为

宽度系数,拓宽了卷积核。整体网络就是由图中结构一个一个堆叠而成。

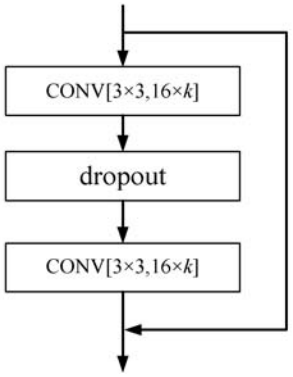

图 2 WRN 模块结构

WRN 模型的输入是语音的特征参数,第一层是卷积层,然后经过若干残差模块,最后通过两个全连接层。第一个全连接层使用 Softmax 函数,对数据集进行分类从而实现语音命令词识别;第二个全连接层使用 sigmod 函数,对数据集做二分类,区别内外部说话人。这种多任务网络设计不仅可以提高复杂环境中的识别精度,还可以避免接收到多余的命令使得设备误触发。WRN 网络如图 3 所示。

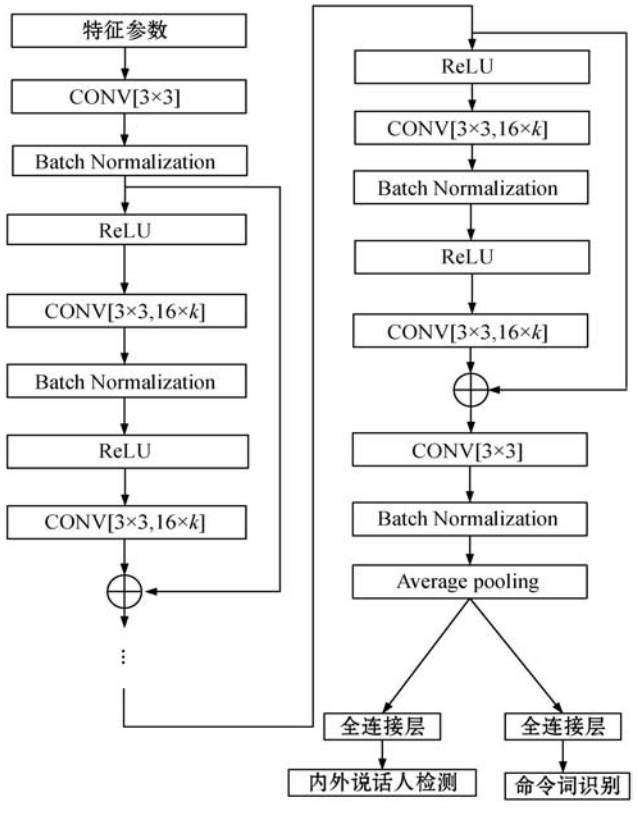

图 3 WRN 整体结构

在文献[16]中,作者将不同深度的和不同宽度的 WRN 模型应用到 CIFAR 图像数据集中,结果显示 16 层的残差神经网络的训练效果要比 1 000 层的卷积网络要好,在 imageNet 数据集中,具有 50 层的 WRN 模型要比 ResNet152 模型训练效果好。

2 双微麦克风阵列

2.1 双微麦克风阵列介绍

基于麦克风阵列的语音处理技术已经相当成熟。普通的麦克风阵列通常具有较大的尺寸,不适用于某些对麦克风尺寸有要求的设备,如双耳型数字助听器、人工耳蜗、机器人听觉系统等,因此提出“双微麦克风阵列”。双微麦克风阵列指的是两个具有一定距离的微型麦克风阵列,具有微型麦克风阵列与双麦克风的结构。双微麦克风阵列不仅可以满足小型化、微型化的要求,同时在多通道语音识别方面也打下良好的基础[17]。

2.2 数据集的产生

谷歌公开的语音命令词数据集是一类简单语音命令数据集,数据集包括 35 个简单的单词,每个语音的时长为 1 秒,共有 105 829 个语音片段,所以该数据集非常适合应用于语音命令词识别研究。

一般来说,多通道语音增强的效果要比单通道语音增强的效果好,由此展开合理推测:在神经网络训练中,多通道语音数据的训练效果通常要强于单通道语音数据的训练效果。为了产生多通道数据,设计如图 4 所示的场景:在人的四周摆放 12 个扬声器,这些扬声器组成圆形的阵列,圆的直径为 3 米,在人的双耳位置各放两个前后麦克风组成的双微阵列,浅色代表前置麦克风,黑色代表后置麦克风。随机选择说话人,通过阵列得到模拟用户声音,其余的说话人用于模拟外部说话人得到外部声音数据集,采样频率为 44.1 kHz,采集完成后经过脉冲响应过滤。采用哈希名称对数据集进行分割,确保数据集没有重叠。训练集总计有 34 187 组、验证集有 5 434 组、测试集有 11 005 组双微麦克风阵列数据。该模型识别的命令有 10 个,分别是“yes”“no”“up”“down”“left”“right”“on”“go”“stop”“off”。除此之外的单词作为未知单词,共计 11 类。

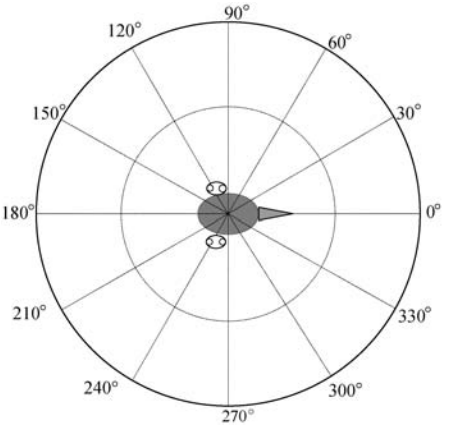

图 4 麦克风阵列

3　语音特征提取

梅尔频率倒谱系数(PNCC)是语音识别系统中应用最多的特征提取算法,但是 MFCC 算法的精度受到噪声环境影响,在低信噪比环境中的识别准确率不能满足实际应用需求。

PNCC 方法的特征参数提取方法是通过改进 MFCC 方法得到的,能够增强抗噪稳健性。PNCC 相比于 MFCC 的优势在于:(1) 将 MFCC 中的三角滤波器组替换为 Gammatone 滤波器组[18],因为 Gammatone 滤波器组的感知特性更加接近人耳;(2) 增加时间窗跨度,获得更好的性能;(3) PNCC 采用幂律非线性与幂律归一化相结合的方法[19],更加符合人耳听觉神经的压缩感知[20]。图 5 为 PNCC 结构图。

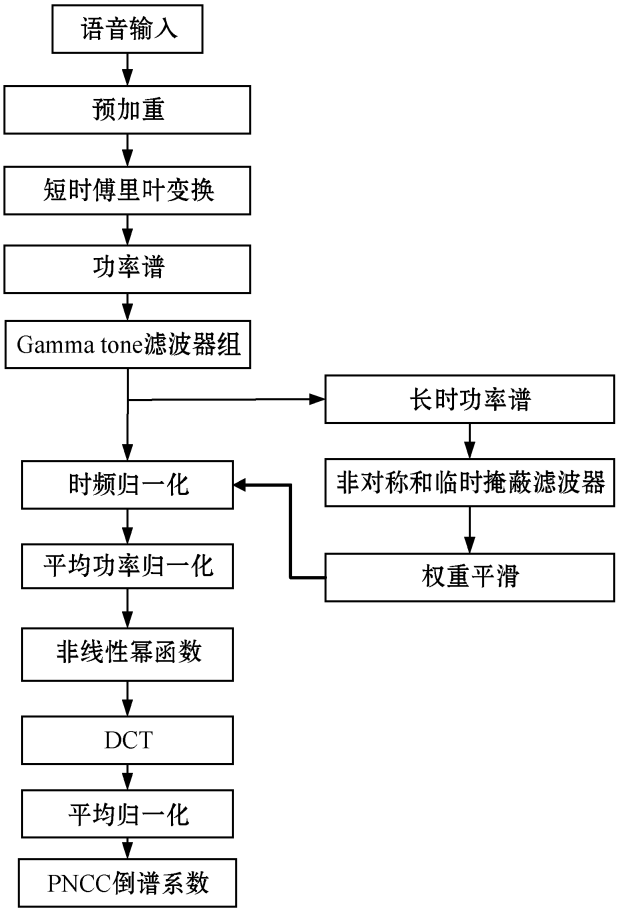

图 5　PNCC 结构图

4　实验结果与分析

4.1　实验结果

将本文模型与文献[21]提到的深度残差网络 ResNet15 模型以及经典的 ResNet18 模型作对比。

文献[21]中原始的 ResNet15 使用单侧双麦克风训练方法作为基础模型,该模型只包含命令词检测识别任务。文献[22]为命令词识别与内外部说话人检测双系统,模型均采用内外部说话人混合数据集进行训练。其中,表 1 为命令词识别准确率,表 2 为内外部说话人检测结果,在考虑整体命令词检测精度时加入了外部扬声器干扰,以检测系统在嘈杂环境中的稳健性。

表 1　不同算法下语音命令词的识别精度(%)

模型	算法	数据类型	特征参数	命令词识别精度	
				内部数据	整体数据
ResNet15	文献[21]	基线	MFCC	94.24 ±0.39	71.87 ±0.30
	文献[22]	前置麦克风	MFCC	94.28 ±0.37	89.48 ±0.74
		后置麦	MFCC	94.48 ±0.25	89.29 ±0.55
		单侧双	MFCC	94.59 ±0.32	94.86 ±0.39
WRN	本文	单侧双	PNCC	97.86 ±0.77	95.08 ±0.52
		双微阵列	PNCC	98.76 ±0.37	96.13 ±0.47
ResNet18	对比	双微阵列	PNCC	94.53 ±0.22	95.69 ±0.30

表 2　内外部说话人检测准确率(%)

模型	数据类型	特征参数	命令词识别精度		
			内部数据	外部数据	整体数据
ResNet15 文献[22]	基线	MFCC	—	—	—
	前置	MFCC	97.49 ±1.02	80.38 ±5.23	93.02 ±0.76
	后置	MFCC	97.28 ±1.08	79.03 ±5.06	92.51 ±0.68
	单侧双	MFCC	99.60 ±0.22	96.22 ±1.61	98.72 ±0.29
WRN	单侧双	PNCC	99.40 ±0.70	96.83 ±0.45	98.18 ±0.12
	双微阵列	PNCC	99.66 ±0.25	99.32 ±0.38	98.50 ±0.57
ResNet18	双微阵列	PNCC	99.50 ±0.10	99.40 ±0.28	98.84 ±0.63
说明:表 1 和表 2 中前置、后置、单侧双指的是前置麦克风、后置麦克风、单侧双麦克风					

4.2　实验分析

表 1 和表 2 列举了 ResNet15、WRN、ResNet18 在不同阵列、不同特征参数的单任务、多任务模式中的命令词识别、内外说话人区分的结果。

(1) 文献[21]中 ResNet15 的模型中有 6 个残差模块,每个残差模块中有两个卷积层。残差模块内部的结构为卷积层、激活函数、批量归一化。ResNet15 整体结构为第一层为卷积层,后面接 6 个残差模块,再接卷积层和全连接层,这样就形成了有 15 个权重层的

ResNet15 网络。

(2) ResNet18 的模型结构与 ResNet15 的结构相似,但是残差模块有 8 个,ResNet18 网络的残差模块内部结构与 ResNet15 的残差模块内部结构不同,内部结构为卷积层、批量归一化、激活函数。

(3) 本文的 WRN 网络的残差模块内部结构为卷积批量归一化、激活函数、卷积层。宽度指的是卷积层的宽度,在实验中,ResNet15 和 ResNet18 中,卷积核的数量均为 45,在 WRN 中设置了一个宽度系数 k,可以拓宽卷积核的数量。本文的 WRN 网络中应用了 3 个残差模块,深度为 7,宽度系数 k 设置为 2。

(4) 文献[21]中的 ResNet15 模型为基线,该系统为单任务系统,没有加入内外说话人检测模块,因此在加入干扰之后整体检测效果较差。

(5) 在文献[22]中,特征参数采用 MFCC 算法,前置麦克风、后置麦克风、单侧双麦克风系统中加入了内外部说话人检测模块。在内部说话人检测准确率上,4 种系统能力相当;在整体数据集上相比较于基线系统有了 18% 左右的性能提升。使用双麦克风系统的检测效果要好于单麦克风系统。

(6) 本文采用的特征提取算法为 PNCC 算法,该算法相较于 MFCC 算法具有一定的抗噪稳健性,从实验结果看,在噪声环境下采用 PNCC 算法的效果要好于 MFCC 算法。

(7) WRN 算法减少了网络层数,使用了 3 个残差模块,增加了网络的宽度,在命令词识别精度上要高于 ResNet15 和 ResNet18 模型。

5 结 语

本文对比了深度残差神经网络深度和宽度对语音识别性能的影响情况。增加网络的宽度虽然会增加计算量,但是计算机性能越来越强大,我们可以忽略增加的计算量,同时减少网络的层数也能抵消增加的宽度带来的大计算量问题。系统采用双微麦克风阵列数据集以及 PNCC 算法有效提高了命令词识别精度。未来将继续对模型进行优化,探究网络层数与宽度的最佳比例,控制模型的计算量。

参 考 文 献

[1] 唐步天. 音频信息隐藏关键技术研究及识别技术的信息安全应用[D]. 合肥:中国科学技术大学,2008.

[2] 李庆龙. 基于深度学习的在线波达方向估计方法研究[D]. 呼和浩特:内蒙古大学,2018.

[3] 赵晓群,张扬. 语音关键词识别系统声学模型构建综述[J]. 燕山大学学报,2017,41(6):471 - 481.

[4] 温玉华. 基于 DTW 算法的英语发音错误自动校正系统设计[J]. 现代电子技术,2020,43(10):124 - 126.

[5] 沈云云. 基于连续隐马尔可夫模型的语音识别抗噪问题研究[D]. 哈尔滨:哈尔滨工业大学,2020.

[6] 李春晓. 基于语音识别的莫尔斯报文系统设计与实现[D]. 哈尔滨:哈尔滨工程大学,2006.

[7] 林波,吕明. 基于 DTW 改进算法的孤立词识别系统的仿真与分析[J]. 信息技术,2006(4):56 - 59.

[8] 李云霞,李治柱,吴亚栋. 基于 HMM 的关键词识别系统[J]. 计算机工程,2004(7):130 - 132.

[9] 陈松. 基于 VQ 的室内说话人识别及 FPGA 实现研究[D]. 淮南:安徽理工大学,2019.

[10] He K M, Zhang X Y, Ren S Q, et al. Identity mappings in deep residual networks[C]//European Conference on Computer Vision,2016:630 - 645.

[11] Chen G, Parada C, And Heigold G. Small-footprint keyword spotting using deep neural networks[C]//IEEE International Conference on Acoustics, Speech and Signal Processing, 2014:4087 - 4091.

[12] 史忠植. 突破通过机器进行学习的极限[J]. 科学通报,2016,61(33):3548 - 3556.

[13] 徐英男. 面向人工神经网络计算核的加速优化及自动生成技术研究[D]. 长沙:国防科学技术大学,2017.

[14] 杨德举,马良荔,谭琳珊,等. 基于门控卷积网络与 CTC 的端到端语音识别[J]. 计算机工程与设计,2020,41(9):2650 - 2654.

[15] 刘虹,袁三男. 基于多尺度残差深度卷积神经网络的语音识别[J]. 计算机应用与软件,2020,37(11):275 - 279.

[16] Huynh H T, Nguyen H. Joint age estimation and gender classification of Asian faces using wide ResNet[J]. SN Computer Science,2020,1(5):284.

[17] 曾庆宁,肖强,王瑶,等. 一种双微阵列语音增强方法[J]. 电子与信息学报,2018,40(5):1187 - 1194.

[18] 楚博策. 鲁棒性语音特征提取研究[D]. 北京:北京邮电大学,2016.

[19] 钟顺明,况鹏,庄豪爽,等. 基于 PNCC 与基频的鲁棒电话语音性别检测方案[J]. 华南师范大学学报(自然科学版),2019,51(6):118 - 122.

[20] 张子涛. 基于小波和 PNCC 特征参数的语音识别技术研究[D]. 重庆:重庆大学,2018.

[21] Tang R, Lin J. Deep residual learning for small-footprint keyword spotting[C]//IEEE International Conference on Acoustics, Speech and Signal Processing,2018:5484 - 5488.

[22] López-Espejo I, Tan Z H, Jensen J. Keyword spotting for hearing assistive devices robust to external speakers[EB]. arXiv:1906.09417,2019.