基于 WGAN 的音频关键词识别研究

李全兵[1,2,3],文 钊[4*],田艳梅[4*],詹茂豪[1],余秦勇[2,3],杨 辉[2,3]
(1. 中国电子科技网络信息安全有限公司,四川 成都 610041;
2. 提升政府治理能力大数据应用技术国家工程实验室,贵州 贵阳 550022;
3. 中电科大数据研究院有限公司,贵州 贵阳 550022;
4. 电子科技大学 信息与软件工程学院,四川 成都 610054)

摘 要:基于语音识别的关键词识别方法增大了关键词识别工作量,降低了识别效率,还使得识别准确率受语音识别和文字查找办法影响,并对无文字语言不适用。针对此问题,提出将 Wasserstein 生成式对抗网络(WGAN)应用于语音关键词识别中,利用生成器输出的生成序列分析语音中有无关键词。为了获取语音中关键词的位置信息,该文为 WGAN 网络定义了一个定位损失函数,以此保证生成的掩码序列可以精确定位出关键词的位置。在四川话、普通话和粤语三门语言的数据集上进行实验,结果表明该技术可以识别无文字语言的关键词,相比于模板匹配方法其识别速度有显著提升。

关键词:语音识别;音频关键词识别;深度学习;Wasserstein 生成式对抗网络;关键词定位
中图分类号:TP183 **文献标识码**:A **文章编号**:1673-629X(2021)08-0026-07
doi:10.3969/j.issn.1673-629X.2021.08.005

Research on Audio Keywords Recognition Based on Wasserstein Generative Adversarial Network

LI Quan-bing[1,2,3], WEN Zhao[4*], TIAN Yan-mei[4*], ZHAN Mao-hao[1],
YU Qin-yong[2,3], YANG Hui[2,3]
(1. China Electronic Technology Cyber Security Co., Ltd., Chengdu 610041, China;
2. Big Data Application on Improving Government Governance Capabilities National Engineering Laboratory, Guiyang 550022, China;
3. CETC Big Data Research Institute Co., Ltd., Guiyang 550022, China;
4. School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China)

Abstract: The keyword recognition method based on speech recognition increases the workload of keyword recognition, reduces the recognition efficiency and makes the accuracy affected by speech recognition and text search methods, which is not applicable to language without words. To solve this problem, the Wasserstein generative adversarial network (WGAN) is applied to speech keyword recognition, and the generated sequence output by generator is used to analyze whether there are keywords in speech. In order to obtain the position information of the keywords in speech, we define a positioning loss function for the WGAN to ensure that the generated mask sequence can accurately locate the position of the keywords. Results on datasets of three languages, Sichuan dialect, Mandarin and Cantonese, show that the proposed method can recognize keywords in languages without characters, and the recognition speed is significantly improved compared with the template matching method.

Key words: speech recognition; audio spoken keyword detection; deep learning; Wasserstein generative adversarial network (WGAN); keyword targeting

0 引 言

随着互联网技术的发展,语音逐渐成为人们在日常共享信息和交流的主要方式,例如在 QQ、微信等社交网站上与朋友聊天,以前人们以文字、图片来传递信息,如今主要通过语音和视频来交流信息,这样既方便又快捷。语音关键词检测(spoken keyword detection,

收稿日期:2020-08-03 **修回日期**:2020-12-04
基金项目:四川省重大科技专项项目(2017GZDZX0002)
作者简介:李全兵(1978-),男,硕士,工程师,研究方向为云计算、大数据、网络空间安全、机器学习。

SKD)是指从连续语音流中识别或检测出一个或多个特定关键词。迄今为止,已有大量文献对语音关键词识别做了研究,总的来说语音关键词识别方法主要分为以下几种:

(1)基于模板匹配的关键词识别[1]。该方法的思想是通过比对模板特征与待识别语音特征的相似性来实现关键词识别。query-by-example(QBE)是模板匹配中主要的检测技术,利用滑动匹配的思想检测关键词。模板特征可用高斯混合模型、隐马尔可夫模型、人工神经网络、梅尔频率系数和线性预测系数等表示。Zhang Y 等人用高斯后验概率作为模板特征,并结合动态时间规整算法(dynamic time wrapping,DTW)的变体 segmental DTW 实现语音关键词的检出[2]。文献[3-4]结合 self-organizing map 和高斯后验概率特征,利用 sub-sequence DTW 实现关键词的检测。Dhananjay 等人[5]利用 CNN 强大的特征提取能力,将 CNN 引用到模板匹配算法中,使得准确率有了显著提升。由文献[6]可知,DTW 算法的时间和空间复杂度均为 $O(n^2)$,因此 n 的长度不能太长,也就是在连续语音中使用此方法,识别速度也相对较慢。

(2)基于大词汇量连续语音识别(large vocabulary continuous speech recognition,LVCSR)[7]的关键词识别方法。这种需要将语音信号解码成词序列或音素网格[8-9],然后在此基础上搜索关键词,搜索方法有两种:基于混淆网络(confusion network,CN)和基于状态转换器(finite state transducer,FST)。在文献[10]中,Chiu 等人提出将 FST 和 CN 组合在一起进行关键词检测,实验表明组合后的搜索方法优于任何单一的搜索策略[9]。针对集外词(out of vocabulary,OOV)[11]问题,Chen 等人利用 G2P(grapheme-to-phoneme)方法和代理关键词的概念有效地解决了 OOV 问题[12]。在文献[13]中,侯一民等人介绍了几种具有代表性的深度学习模型,并对其在语音识别中的应用进行了简单的说明。语音识别的声学模型除了可以使用 DNN-HMM 模型外[14],还可以使用其他的神经网络,比如循环神经网络(recurrent neural network,RNN)[15]和 DNN-RNN[16]。文献[17]中,作者将自编码器深度学习神经网络应用于语音识别中,实现了语音孤立词的识别。文献[18]中,针对低资源情形下,语音识别系统性能不佳的问题,提出了一种基于 i-vector 特征的 LSTM 递归神经网络语音识别系统。

(3)端到端的语音关键词识别。基于端到端的关键词检测系统通常包括三个部分:特征提取模块、神经网络模块和输出后验得分的计算模块[19-21]。在识别阶段,关键词检测系统首先提取语音特征,之后将特征送入训练好的神经网络模型中,输出各个关键词和非关键词的后验概率,最后对后验概率以一定的窗长进行平滑,平滑后的后验得分如果超过预先设定的阈值或者选取平滑后多个关键词中最大的后验得分,就认为识别出了某个关键词。这种基于端到端的关键词检测系统,主要应用于语音唤醒任务中,在连续语音中不适用。

启发于端到端的语音关键词识别,该文将生成式对抗网络应用于连续语音关键词识别中,特定为小众且无文字的语言设计一种基于音频的关键词识别方法,并为 GAN 设计一个定位关键词的目标函数,以此追踪定位出关键词的具体位置。

1　生成式对抗网络简介

1.1　原始对抗网络

生成式对抗网络(GAN)是 Goodfellow 等[22]在 2014 年提出的一种生成式模型,其优化过程是一个极小极大博弈(minimax game)问题,优化目标是达到纳什均衡[23],得到全局最优解。GAN 由一个生成器 G 和一个判别器 D 构成,G 获取真实样本数据的概率分布生成新的数据,D 作为分类器,对输入数据进行分类。

目前,GAN 被广泛应用到图像和视觉领域,已经可以生成数字、人脸,构成各种逼真的室内外场景,根据轮廓图恢复图像,从低分辨率图像生成高分辨率图像等[24],GAN 也已经开始被应用到语音和自然语言处理[25-26]问题的研究中。

GAN 的判别器的损失函数如公式(1)所示:

$$L_D = E_x[\log(D(x))] + E_z[\log(1 - D(G(z)))] \tag{1}$$

生成器的损失函数表达式如公式(2)所示:

$$L_G = E_z[\log(1 - D(G(z)))] \tag{2}$$

1.2　Wasserstein 对抗网络

Arjovsky 等[27]从理论上阐述了原始 GAN 存在训练不稳定、梯度消失等问题。由于 Wasserstein 距离相对 KL 散度与 JS 散度具有优越的平滑特性,理论上可以解决梯度消失问题,可以提供有意义的梯度。因此,Arjovsky 等人用 Wasserstein 距离代替原始 GAN 目标函数中的 KL 散度、JS 散度。WGAN 的目标函数由 Wasserstein 距离产生,Wasserstein 距离定义如式(3)所示:

$$W(P_r,P_g) = \inf_{\gamma \in \prod(P_r,P_g)} E_{(x,y)\sim\gamma}[\,\|x - y\|\,] \tag{3}$$

其中,(x,y) 分别表示真实样本和虚假样本,$\prod(P_r,P_g)$ 表示 P_r 和 P_g 组合起来的所有可能的联合分布的集合。对于每一个可能的联合分布 γ 而言,可以从中采样 $(x,y) \sim \gamma$,得到一个真实样本 x 和一个生成样

本 y,并算出这对样本的距离 $\|x-y\|$,所以可以计算该联合分布 γ 下样本对距离的期望值 $E_{(x,y)\sim\gamma}[\|x-y\|]$。在所有可能的联合分布中能够对这个期望值取到的下界 $\inf_{\gamma\in\prod(P_r,P_g)} E_{(x,y)\sim\gamma}[\|x-y\|]$。用 Wasserstein 距离代替 JS 散度构造 GAN 模型的损失函数得到生成器的损失函数如式(4)所示:

$$L_{WG}=-E_{x\sim p_g}[f_w(x)] \tag{4}$$

判别器的损失函数如式(5)所示:

$$L_{WD}=E_{x\sim p_g}[f_w(x)]-E_{x\sim p_r}[f_w(x)] \tag{5}$$

式(4)和式(5)中的 $f_w(x)$ 表示判别器。

经过生成式对抗网络的简单介绍,在语音关键词识别任务中,该文采用 WGAN 来实现关键词的识别。

2 WGAN 识别关键词的基本思想和方法

由于端到端的关键词识别算法包括特征提取模块、神经网络模块和输出后验得分的计算模块,因此,该文使用 WGAN 识别语音关键词也由三个模块组成,即:特征提取模块、WGAN 神经网络模块和关键词识别模块,基本思想如图 1 所示。

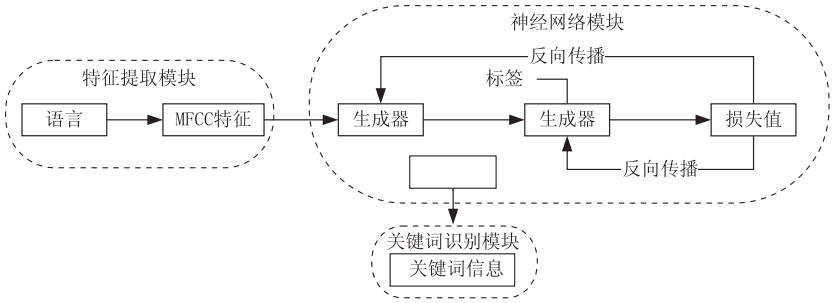


图 1 WGAN 识别关键词过程

在识别阶段,首先提取语音特征,之后将特征送入训练好的 WGAN 模型中,输出经过处理后的各个帧属于关键词特征的后验概率,最后依据生成序列分析语音中是否存在关键词。

特征提取模块,该文使用梅尔频率系数作为语音信号的特征,主要从定位损失函数、模型训练和关键词识别等方面来介绍 WGAN 识别关键词的基本思想和方法。

2.1 定位损失函数

由于所设计的方法需要获取关键词的位置信息,因此需要为 WGAN 制定一个目标定位损失函数。在这之前需要了解标签和生成序列这两个名称。文中的标签是按照以下步骤制作的:

(1)假设音频的总时长为 T,存在关键词且在语音中对应的开始和结束时间分别为 s 和 e,那么出现关键词的区间为 $[\lfloor \frac{s}{T}\times M\rfloor,\lceil \frac{s}{T}\times M\rceil]$,其中 M 表示语音提取特征后的帧长。

(2)将关键词区间对应的帧全记为 1,表示该区域存在关键词特征,其余区间的帧记为 0,最终得到大小为 $1\times M$ 的标签。

(3)由于语音信号时长 T 大小不一,使得 M 也不同,最终导致标签大小不一致。为了解决这个问题,获取所有语音的标签,从中选出最长的标签作为标杆,将其余标签序列的末尾填充 0 至与标杆一致。

生成序列是 WGAN 的生成器 G 输出的序列经过以下步骤处理得到的:

(1)假设 G 的输出值为 $y=\{y_1,y_2,\cdots,y_M\}$,其中 y_i 表示第 i 帧特征是关键词特征的概率。

(2)若 $y_i\geqslant 0.5$ 时,则 y_i 置为 1,否则将 y_i 置为 0,这样就得到了只有值为 0 和 1 且大小为 $1\times M$ 的生成序列。

为了更形象地描述标签和生成序列,如有以下生成序列:{000000011111101010110000},则标签序列为{000001111111111100000000},可以看出标签中包含关键词特征的帧区间全部为 1,不含关键词的区间全为 0。在生成序列中关键词所在区域,除了 1 之外,还有少数的 0,因此为了使得生成序列更加真实,该文为 G 定义了一个定位损失函数,如下:

$$L'_{WG}=\frac{\lambda}{M}\times\left(\sum_{i=1}^{M}|y_i-y'_i|\right) \tag{6}$$

其中,$|y_i-y'_i|$ 表示标签和生成序列之间第 i 个值的绝对误差;λ 是常数,增加常数 λ 可有效防止定位损失函数在训练时出现 0 导致梯度消失的现象,通常 λ 取值为 0.000 1。

综上,在语音关键词识别任务中 WGAN 的 G 的损失函数如式(7)所示:

$$\bar{L}_{WG}=L_{WG}+L'_{WG} \tag{7}$$

由于判别器 D 的作用仍是分类,所以在语音关键词检测任务中,D 的损失函数不变。

2.2 模型训练

音频关键词识别算法中,生成式对抗网络的 G 和 D 的结构与 LeNet[28]类似,在语音关键词识别任务中,WGAN 的 G 有 10 个卷积层,用以充分获取关键词特

征, G 的结构与 LeNet 最大的区别是前者不含全连接层和 Sigmoid 层, G 经过最后一层卷积层的输出值是大小为 $1 \times M$ 矩阵,在将其传入 D 之前需要做一个简单的变换,即使其变成大小不变的向量。D 包含 7 个卷积层、3 个全连接层,不包含 Sigmoid 层。文中每层卷积层经过了卷积、池化、BatchNorm、ReLu 和 Dropout 操作。

WGAN 的训练过程如图 2 所示。

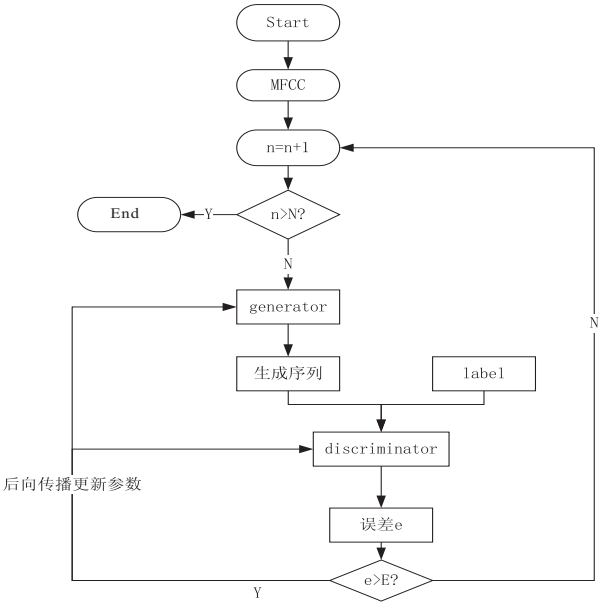


图 2　音频关键词识别训练流程

训练模型的时候需要预先设置网络训练步数 N, 若训练步数达到 N 则训练结束,否则继续训练。在训练网络的过程中,通过前向传播获得误差并判断误差是否小于阈值 E, 若误差超过阈值则依据误差反向传播更新 G 和 D 的参数,直至误差小于 E 后进行下一步训练。由于一个良好的 D 不仅可以监督 G 生成效果更好的生成序列,还能加快模型的训练速度。因此在训练 WGAN 时,一般先训练 D, 当 D 的参数更新若干次之后,将 D 的参数固定,之后才开始训练 G。G 和 D 的参数更新通过反向传播算法实现。

2.3　关键词识别

基于音频的关键词识别算法的模型训练好之后,就可检测音频中是否存在关键词,若生成序列中有连续的若干个值为 1 (如:{000001111110000000000100101}),则定义它为一个连通区域为{111111}。

若生成序列中存在连通区域,还不能判断语音中有关键词,此时还需要判断连通区域的长度是否超过阈值 Th。若生成序列中存在连通区域长度大于 Th 的情况,则视为语音中是存在关键词的,否则认定语音中不存在关键词。假设生成序列为 $y = \{y_1, \cdots, y_i, \cdots, y_j, \cdots, y_M\}$, 其中 $y_1 \sim y_{i-1}$ 的值为 0, $y_i \sim y_j$ 的值为 1, 且 $j - i \geqslant \mathrm{Th}$, $y_{j+1} \sim y_M$ 的值为 0,根据生成序列与音频之间的对应关系,把连通区域映射到音频,从而得到关键词在语音中的定位结果,关键词在语音中的确切位置如公式(8)所示:

$$\left[\max\left(0, \left\lfloor \frac{i}{M} \times T \right\rfloor\right), \min\left(T, \left\lfloor \frac{i}{M} \times T \right\rfloor\right)\right] \quad (8)$$

对于阈值 Th 的大小要依据关键词的情况而定,比如设定的关键词语音平均持续时长为 1 秒,那么阈值 Th 应设置在 4 ~ 10 这个范围内(因为 1 秒的音频数据可生成大约 10 帧长度为 22 维的特征)。若语音信号的语速较快,则阈值 Th 的值可设置一个较小的值(0 ~ 4 之间的任意一个值),否则 Th 的值应大于 4。

3　数据集及环境配置

3.1　数据集

文中自行建立了语料库,所包含的语言有普通话、四川话和粤语。共设定 10 个中文关键词,分别为:一带一路、互联网时代、民族尊严、一国两制、人民代表大会、国家主席、中华人民共和国、体制改革、环境治理、自然灾害。语音来源包括录音及网络广播两种方式。在安静环境下录音,拟定包含关键词的语句内容,每个关键词涉及的语句内容有 10 条,分别用普通话、四川话和粤语朗诵若干次。参与录音的人数为 30 人,其中男生 20 人,女生 10 人,年龄分布均在 18 岁到 26 岁之间,这 50 个志愿者均会四川话和粤语。从网络上下载的广播,有的片段有关键词,有的片段没有关键词,两者之间的比例为 8 : 2。

根据上述方法得到语音的均为 WAV 格式,单声道,采样频率 16 kHz。文中设置的每个关键词发音时长平均在 1 ~ 2 秒内,含关键词语音(句子)长度均为 5 ~ 15 秒。语音中含有普通话、四川话和粤语三种关键词语音文件。普通话、四川话和粤语分别包含5 600 条人工采集数据和 4 400 条网络语音数据作为训练集,各类别的方言语音分别使用 400 条人工采集数据和 200 条网络语音数据作为测试集。

这些关键词都是一些名词。在获得音频之后,根据 2.1 小节中的标签制作方法制作标签,并以与音频对应的名称单独保存在一个文件夹内。

3.2　实验环境配置

文中使用 Python 的深度学习库 Tensorflow 编程实现 WGAN,经过一系列实验,WGAN 的一些超参数设置如下:学习速率为 0.000 3,Dropout 正则化的保留概率 keep-prob 设置为 0.5,优化算法使用 RMSProb。实验的硬件配置为:Intel core CPU @ 2.6 GHz+Nvidia GeForce GTX 1080 Ti (11 GB) 以及 32 GB RAM,软件环境为:Ubuntu 16.04 及 Python3.5.0。

4 实验与结果

4.1 评价指标

文中采用准确率、召回率、错误接受率和错误拒绝率[29]作为评估关键词识别的性能指标。TP 表示本属于正例的样本被正确预测为正例的样本数,FN 表示正例被错误判别为负例的样本数,TN 表示负例被正确判断为负例的样本数,FP 表示负例被错误判断为正例的样本数。

准确率,衡量某一检索任务判断正确的概率,其定义如公式(9)所示:

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

召回率,表示所有正例中被判断为正例的概率,其定义如式(10)所示:

$$\text{recall} = \frac{TP}{TP + FN} \tag{10}$$

错误接受率,表示不包含关键词的样本中,错误检测出有关键词样本所占比例,其定义如公式(11)所示:

$$FAR = \frac{TP}{TP + FP} \tag{11}$$

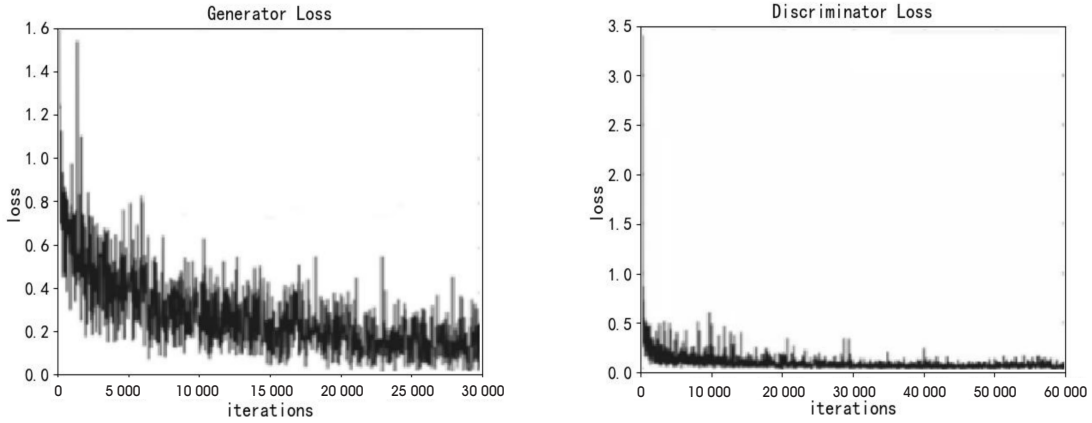
错误拒绝率,表示在有关键词的语音样本中,没有检测到管检测的语音样本所占比例,其表达式如公式(12)所示:

$$FRR = \frac{FN}{TP + FN} \tag{12}$$

4.2 实验结果评估

WGAN 模型在自制的语料库上进行训练。图 3 展示了 G 和 D 的损失值随着训练步数的变化情况,图中是每隔 50 步统计一次数据得到的值。

分析图 3 知道,WGAN 模型的生成器大概在 2.5w 步时得到了收敛,继续训练生成器,它的损失值有所波动,但不影响总体变化趋势。判别器直到 3w 步左右收敛,继续训练网络损失值也无明显变化,因此文中将训练到 3w 步时得到的模型作为最优模型。



(a)生成器损失值变化情况 (b)判别器损失值变化情况

图 3 模型损失值变化情况

为了设置合适的阈值 Th,观察 Th 的值在 0 ~ 10 之间,模型识别关键词的准确率的变化情况,在普通话这个数据集上的实验结果如图 4 所示。

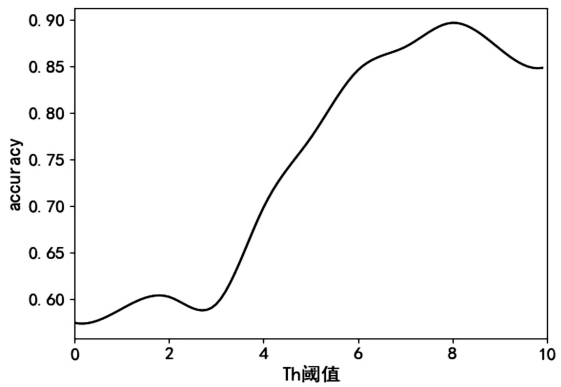


图 4 准确率与 Th 值之间的关系

从图 4 分析可知,当 Th 的值设置过低时(0 ~ 4),准确率低于 0.7,当 Th 的值逐渐增加时,accuracy 也在逐渐上升,直到 Th 的值为 8 时达到了顶点,Th 再继续增大,准确率反而降低。因此,文中将连通区域的阈值 Th 设置为 8。

此模型是用普通话、四川话和粤语三种语言的音频训练得到的,因此,它在这三种语言的测试集上进行验证,其结果如表 1 所示。

表 1 三种语言的测试结果

语言	评价指标			
	recall	FAR	FRR	accuracy
普通话语	0.861	0.020	0.103	0.897
四川话	0.844	0.018	0.112	0.885
粤语	0.839	0.023	0.092	0.880

根据表 1 的结果表明,WGAN 可以识别四川话和粤语这两种方言,并且识别准确率和召回率均到达了 80%,说明 WGAN 是有能力识别连续语音中的关键词

的,并且可以识别四川话和粤语这两门方言中是否存在关键词。

此外,文中还提到 WGAN 模型可以准确获取关键词的时间信息,因此,含关键词的语句内容为“随着社会的快速发展,我国已进入了互联网时代”,其中“互联网时代”为关键词,WGAN 检测该段语句内容的语音得到的结果为:关键词在音频中开始出现时间和结束时间分别为 8.5 秒和 9.8 秒。

从上述实验结果可知,生成式对抗网络不仅能识别语音中的关键词,还可以定位出关键词在语音中的具体位置。

与模板匹配算法的对比见表 2。

表 2　与模板匹配算法对比结果

算法	评价指标		
	recall	accuracy	平均识别时间/s
文献[1]	0.873	0.852	16.9
文献[5]	0.865	0.903	8.7
WGAN	0.861	0.897	4.3

据表 2 的结果分析,文献[1]中的快速模板匹配的准确率略低于 WGAN 模型,WGAN 识别关键词的速度比它快了将近 4 倍。虽然文献[5]中的模板匹配方法准确率是三个方法中最好的,但是其花费的时间却是文中所提方法的 2 倍。这就表明,基于 WGAN 的音频关键词识别方法识别速度比模板匹配快。

接下来,与文献[14]中的基于 DNN-HMM 的识别方法进行比较,其实验结果如表 3 所示。

表 3　与语音识别算法对比结果

算法	评价指标		
	accuracy	recall	FAR
文献[14]	0.931	0.898	0.035
WGAN	0.897	0.861	0.020

通过实验对比可以看到,基于语音识别的关键词识别方法的准确率可高达 0.931,但仅仅比所研究的方法高了 0.034,而 WGAN 识别关键词的错误接受率 FAR 却比语音识别低了 0.015,因此可得出这样的结论:WGAN 识别关键词的性能与基于 DNN-HMM 的语音识别的性能相差无几。

此外,本小节还做了一组鲁棒性实验,以验证所提方法的抗噪能力。对普通话这个数据集的测试集加入信噪比(signal-to-noise ratio,SNR)分别为 20 dB、15 dB、10 dB 和 5 dB 的高斯白噪声,查看模型在不同强度的噪声情形下识别关键词的能力,实验结果如图 5 所示。

可以看到,在噪声不大的情况下,WGAN 识别关键词的准确率基本上与安静环境下相差不大,但当噪声的信噪比变为 15 dB 时,准确率严重下降,当 SNR 为 5 dB 时,准确率低至 0.365。这个表明,基于 WGAN 的音频关键词识别方法具有微弱的鲁棒性。

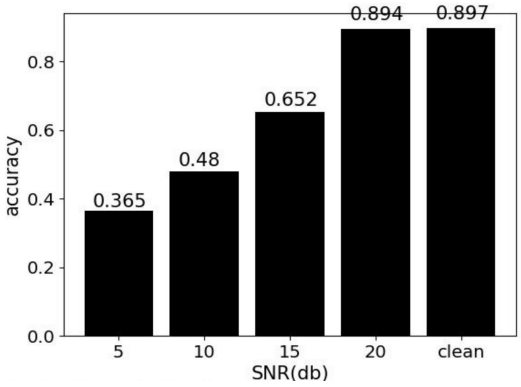


图 5　模型准确率与噪声之间的关系

5　结束语

通过分析基于语音识别的关键词识别技术,发现该方法的工作量大,对无文字语言的关键词识别不适用,并且无法获得关键词的具体位置。针对这些问题,提出了一种基于 WGAN 的音频关键词检测方法,利用 WGAN 的生成器生成定位关键词的掩码序列,用于分析音频中有无关键词以及关键词的位置信息。在包含普通话、四川话和粤语的混合数据集上训练了 WGAN 模型,从准确率、召回率、错误接受率和错误拒绝率分析了所提方法的性能。虽然模板匹配算法也能识别出无文字语言的关键词,但是其识别速度低于文中所提方法。另外,WGAN 识别关键词的性能与基于 DNN-HMM 的语音关键词识别方法相差不大。这就说明基于 WGAN 的音频关键词识别方法可作为无文字语言关键词识别方法的一种替代方法。由于所研究的方法可以获得关键词的位置信息,因此,文中所提方法在隐私保护等领域具有应用前景。由于基于 WGAN 的音频关键词识别方法抗噪能力低,下一步的工作将研究如何提升模型的鲁棒性。

参考文献:

[1] 朱国腾,孙　伟. 基于模板匹配的快速语音关键词检出方法[J]. 计算机应用,2013,33(11):3138-3140.

[2] ZHANG Y,GLASS J R. Unsupervised spoken keyword spotting via segmental DTW on Gaussian posteriorgrams[C]// Proceedings of IEEE workshop on automatic speech recognition & understanding. Merano,Italy:IEEE,2009:398-403.

[3] WU H,LI M,CAI Z,et al. Unsupervised query by example spoken term detection using features concatenated with self-organizing map distances[C]//Proceedings of 2018 11th international symposium on Chinese spoken language processing (ISCSLP). Taipei:[s. n.],2018:1-5.

[4] YUSUF B, GUNDOGDU B, SARACLAR M. Low resource keyword search with synthesized crosslingual exemplars[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 27(7): 1126-1135.

[5] RAM D, MICULICICH L, BOURLARD H. CNN based query by example spoken term detection[C]//Proceedings of interspeech. [s. l.]: IEEE, 2018: 92-96.

[6] SALVADOR S, CHAN P. Toward accurate dynamic time warping in linear time and space[J]. Intelligent Data Analysis, 2007, 11(5): 561-580.

[7] WEINTRAUB M. LVCSR log-likelihood ratio scoring for keyword spotting[C]//Proceedings of international conference on acoustics, speech and signal processing. Detroit, MI, USA: IEEE, 1995: 297-300.

[8] JANSEN A, NIYOGI P. Point process models for spotting keywords in continuous speech[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2009, 17(8): 1457-1470.

[9] SZOKE I, SCHWARZ P, MATEJKA P, et al. Phoneme based acoustics keyword spotting in informal continuous speech[C]//Proceedings of international conference on text. Berlin, Heidelberg: Springer, 2005: 302-309.

[10] CHIU J, WANG Y, TRMAL J, et al. Combination of FST and CN search in spoken term detection[C]//Annual conference of the international speech communication association (interspeech). [s. l.]: IEEE: 2014: 2784-2788.

[11] CAN D, COOPER E, SETHY A, et al. Effect of pronounciations on OOV queries in spoken term detection[C]//Proceedings of IEEE international conference on acoustics, speech and signal processing. Taipei: IEEE, 2009: 3957-3960.

[12] HINTON G, DENG L, YU D, et al. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups[J]. IEEE Signal Processing Magazine, 2012, 29(6): 82-97.

[13] 侯一民, 周慧琼, 王政一. 深度学习在语音识别中的研究进展综述[J]. 计算机应用研究, 2017, 34(8): 2241-2246.

[14] 王朝松. 基于深度学习的汉语语音关键词检测方法研究[D]. 哈尔滨: 哈尔滨工业大学, 2015.

[15] FERNÁNDEZ S, GRAVES A, SCHMIDHUBER J. An application of recurrent neural networks to discriminative keyword spotting[C]// Proc. of ICANN. Porto, Portugal: Springer, 2007: 220-229.

[16] GRAVES A, MOHAMED A R, HINTON G. Speech recognition with deep recurrent neural networks[C]//2013 IEEE international conference on acoustics speech & signal processing. Vancouver, BC, Canada: IEEE, 2013: 6645-6649.

[17] 王山海, 景新幸, 杨海燕. 基于深度学习神经网络的孤立词语音识别的研究[J]. 计算机应用研究, 2015, 32(8): 2289-2291.

[18] 黄光许, 田 垚, 康 健, 等. 低资源条件下基于 i-vector 特征的 LSTM 递归神经网络语音识别系统[J]. 计算机应用研究, 2017, 34(2): 392-396.

[19] CHEN G, PARADA C, HEIGOLD G. Small-footprint keyword spotting using deep neural networks[C]//IEEE international conference on acoustics, speech and signal processing. Florence, Italy: IEEE: 2014: 4087-4091.

[20] TARA N S, CAROLINA P. Convolutional neural networks for small-footprint keyword spotting[C]//Proceedings of 16th annual conference of the international speech communication association. Dresden, Germany: [s. n.], 2015: 1478-1482.

[21] ARIK S O, KLIEGL M, CHILD R, et al. Convolutional recurrent neural networks for small-footprint keyword spotting[C]//Annual conference of the international speech communication association. [s. l.]: IEEE, 2017: 1606-1610.

[22] GOODFELLOW I J, POUGET A J, MIRZA M, et al. Generative adversarial nets[C]//Proceedings of international conference on neural information processing systems. Montreal, Canada: NIPS, 2014: 2672-2680.

[23] RATLIFF L J, BURDEN S A, SASTRY S S. Characterization and computation of local Nash equilibria in continuous games[C]//Proceedings of 51st annual Allerton conference on communication, control, and computing (Allerton). Monticello, IL, USA: IEEE, 2013: 917-924.

[24] GOODFELLOW I J. NIPS 2016 tutorial: generative adversarial networks [EB/OL]. [2016]. https://arxiv. org/pdf/1701. 00160. pdf.

[25] LI J W, MONROE W, SHI T L, et al. Adversarial learning for neural dialogue generation[C]//EMNLP 2017: conference on empirical methods in natural language processing. Stroudsburg, PA: ACL, 2017: 1-13.

[26] YU L T, ZHANG W N, WANG J, et al. SeqGAN: sequence generative adversarial nets with policy gradient[C]//Proceedings of the 31th conference on artificial intelligence. Palo Alto, USA: AAAI Press, 2016: 2852-2858.

[27] ARJOVSKY M, CHINTALA S, BOTTOU L. Wasserstein GAN[DB/OL]. [2018]. https://arxiv. org/abs/1701. 07875.

[28] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[29] POWERS D M W. Evaluation: from precision, recall and F-factor to ROC, informedness, markedness & correlation[J]. Journal of Machine Learning Technologies, 2011, 2(s1): 37-63.