doi: 10.3969/j.issn.0490-6756.2020.05.010

基于 CGRU 多输入特征的地空通话自动切分

郭东岳[1], 林 毅[1], 杨 波[1,2]
(1. 四川大学视觉合成图形图像技术国防重点学科实验室, 成都 610065; 2. 四川大学计算机学院, 成都 610065)

摘 要: 自动语音切分是语音识别、声纹识别、语音降噪等语音应用中非常重要的预处理环节,切分算法的优劣直接影响了系统输出结果的精度. 在空管地空通话中,传输信道噪声、天气因素以及说话人工作状态均会对语音信号产生影响,进而在一定程度上影响语音切分性能. 在分析空管地空通话语音特性基础上,提出了一种基于 CGRU 网络多输入特征的自动语音切分方法. 该方法结合地空通话的特点,采用深度学习的方法进一步提取语音信号的时域和频域非线性特征,将语音信号帧分类为语音帧、结束帧以及其他帧三类. 实验对比了多种语音特征作为输入对切分效果的影响,同时验证了 GMM、CNN、CLDNN、CGRU 等切分算法在真实地空通话测试集上的表现,并提出了一种简单预测结果平滑算法. 实验结果表明,文中提出的自动切分方法在地空通话中具有明显优势,分类模型的 AUC 值达到了 0.98.

关键词: 语音切分; 语音端点检测; 地空通话; 卷积神经网络; 循环神经网络
中图分类号: TP301 **文献标识码**: A **文章编号**: 0490-6756(2020)05-0887-07

Automatic speech segmentation for air-ground communication based on multi-input CGRU neural network

GUO Dong-Yue[1], *LIN Yi*[1], *YANG Bo*[1,2]
(1. National Key Laboratory of Fundamental Science on Synthetic Vision, Sichuan University, Chengdu 610065, China; 2. College of Computer Science, Sichuan University, Chengdu 610065, China)

Abstract: Automatic Speech segmentation is a very important pre-processing approach in many large-scale applications such as speech recognition, speaker recognition and speech noise reduction. The performance of the segmentation algorithm directly affects the accuracy of the system output. In the air traffic control, the quality of the channel, the weather factor and the workload level of the speaker hugely affect the speech segmentation performance. In this paper, by analyzing the speech feature of air-ground communication, an automatic speech segmentation approach is proposed based on CGRU network. The proposed method analyzes the characteristics of air-ground communication, and uses the deep learning method to further extract the time-domain and frequency-domain nonlinear features of the speech signal, and classifies the speech signal frame into three categories: speech, end signal and others. The experiment compares the effects of multiple speech features as input on the segmentation effect, and verifies the performance of GMM, CNN, CLDNN, CGRU and other segmentation algorithms on the air-ground communication test dataset, a simple prediction result smoothing algorithm is presented. The experimental results show that the automatic segmentation method proposed in this paper has obvious

收稿日期: 2019-06-25
基金项目: 国家自然科学基金委员会与中国民用航空局联合项目(U1833115)
作者简介: 郭东岳 (1994—), 男, 山东济宁人, 硕士研究生, 主要研究领域为计算机应用. E-mail: 2017226049015@stu.scu.edu.cn
通讯作者: 杨波. E-mail: boyang@scu.edu.cn.

advantages in air-ground communication, the AUC value of the classification model reaches 0.98.
Keywords: Speech segmentation; VAD; Air-ground communication; CNN; RNN

1 引 言

随着空管自动化概念的提出,许多前沿技术都在空中交通管制中进行了探索与应用.其中,地空通话语音识别、声纹识别等是研发空中管制安全辅助系统[1]、通话数据分析系统的主要技术手段.地空通话实时自动切分是从地空通话语音流中将不同说话人的语音切分出来,为语音降噪、语音识别、声纹识别等应用提供可靠的语料信息,是大型空管语音应用系统中不可或缺的环节.

目前主流的语音切分方法一般是基于语音端点检测 VAD(Voice Activity Detection)方法实现,从技术原理来看主要分为三类:(1) 是基于声音能量特征,比如过零率、短时能量、双门限法[2]等,这类方法抗噪性较差,只能进行简单的声音与静音的检测,适用于语音信道噪声较小的场景;(2) 是基于语音统计学特征,比如高斯混合模型 GMM(Gaussian Mixed Model)[3]、隐马尔科夫模型 HMM(Hidden Markov Model)[4]、谱熵法[5]等,这类方法抗噪性较好,能区分一般噪声与人声的区别,就鲁棒性而言要优于第一类.其中,Google 开源的基于 GMM 的 webrtcvad 语音切分算法以其普适性、灵活性在工业界颇受欢迎.但是这类方法不能应对特殊噪声,如电话铃声,特殊设备噪声等;(3) 是基于深度神经网络 DNN(Deep Neural Network)、卷积神经网络 CNN(Convolutional Neural Network)、循环神经网络 RNN(Recurrent Neural Network)等深度学习的方法[6-14].这类方法通过监督学习训练分类模型以区分语音帧与非语音帧的特征,既可以适用于普通语音环境下,又可以针对特殊环境下的语音信道进行采样、学习,以适应特殊信道、提高切分的准确率.

本文前期研究[6]证明了地空通话语音切分中基于 CNN 深度神经网络的方法的性能优于基于 GMM 的方法.该方法在帧级别上对语音帧和非语音帧进行了区分,加入噪声帧训练后,模型具有一定的抗噪性,能有效规避噪声的干扰.但是,该方法在通话停顿时间较长时仍然会将语句切断.基于以上不足之处,本文做了以下几点研究.

(1) CGRU 神经网络:基于本文的前期研究工作[6],改进了 CNN 卷积结构,并在 CNN 卷积之后加入了基于 RNN 的门控循环单元 GRU(Gated Recurrent Unit)网络层,即本文提出的 CGRU 结构.实验结果表明,CGRU 网络进一步提高了卷积核的音频特征提取能力,同时提高了帧级别的分类精度.

(2) 地空通话结束帧:经分析大量的地空通话数据,发现地空通话设备 PTT(Push To Talk)[15]在通话结束时会产生一种特殊的音频帧,标志着通话结束,本文将之定义为结束帧.因此,本文将语音信号分为三类:语音帧、结束帧、其他帧.

(3) 多特征输入:对比了 LPS(log-power Spectrum)、MFCC(Mel Frequency Cepstral Coefficents)、Fbank(Filter Bank)以及 MFE(MFCC、Fbank、Energy) 联合特征在地空通话语音切分中的性能.

实验结果证明,本文提出的方法网络参数较少,在保证实时切分的前提下准确度明显提高,同时在未训练过的地空通话信道中表现良好,是一种稳定、高效的地空通话自动切分的方法.

2 地空通话的特点

地空通话主要依靠高频无线电收发语音信号,实时性强,但是易受天气、设备等因素干扰,从而影响通话质量.地空通话自动切分的主要难点是从实时的语音流中检测一句话的开始与结束,尤其是语音结束点.飞行流量、说话人习惯、语速、信道质量等都是影响判断语音端点的关键.根据大量通话数据分析发现,地空通话语音切分技术较一般语音切分主要有以下难点.

(1) 地空通话往往是以对话的形式出现,管制员发出管制指令后,飞行员要马上复诵以确认指令.如图 1(a)所示,由于应答时间间隔短,对话产生粘连,传统方法难以切分对话.

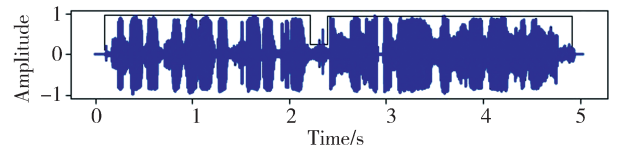
(2) 飞行流量高峰时段信道中说话人较多、通话密集,各说话人语速、习惯等不尽相同,对算法的鲁棒性要求较高.

(3) 信道易受天气、通话设备等因素的影响,在恶劣的生产环境下信道中出现大量随机不稳定噪声.

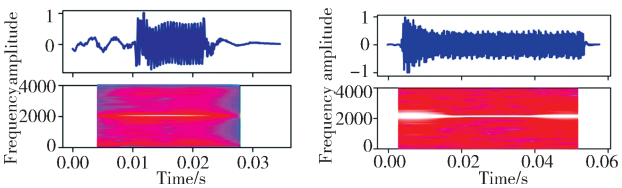
本文通过对成都、北京、太原等地区管制中心的大量历史通话数据分析发现,在多数信道中每人

通话结尾均存在一种特殊的音频帧，帧长在 30～60 ms 不等. 经分析验证，该帧是释放通话设备(PTT)开关时产生的一种特殊信号. 一般而言，该帧的出现标志着说话人释放了 PTT 开关，即说话结束，本文定义该帧为结束帧. 以结束帧作为语音结束标志并结合静噪音检测将大幅度提高语音端点检测的准确性，从而提高切分精度. 在实际应用中，由于各管制中心通话设备不尽相同，结束帧也存在一定的差异，目前数据集中约存在 6 类结束帧，其波形-频谱图样例如图 1(b)和(c)所示，而随着应用场景的增加结束帧的类别也将随之增加. 目前亟需一种通用的技术手段以辨别不同的结束帧，提高切分方法鲁棒性，以达到自适应切分各地区地空通话的目的.

因此，针对地空通话的特性，本文从各管制中心历史地空通话语音中采集特殊噪音、结束帧样本，经数据清洗、人工标注后加入数据集进行训练深度神经网络，旨在应对极端天气或复杂环境下的不稳定噪声，提高语音切分准确率.



(a)通话粘连波形图



(b) 结束帧 1　　(c) 结束帧 2

图 1　地空通话的特点

Fig. 1　Examples audio of air-ground communication

3　地空通话自动切分方法

根据地空通话的特点，需要设计一种抗噪性强、鲁棒性较好、计算速度快的网络结构完成地空通话实时切分任务. 文献[16]研究表明，Convolutional Recurrent Neural Networks 在音频分类任务中的表现优异. 本文在改进前期研究[6]中的 CNN 卷积结构的同时，加入了 GRU 网络层，将模型的输出类别为三类，包含语音帧、结束帧和其他帧，并提出了一种简单平滑算法.

3.1　CGRU 网络结构

为了保证空管安全辅助系统的实时性、降低切分时延，与一般使用上下文多帧输入的深度学习方法不同，本文采用帧长 35 ms，步长 15 ms 的单帧预测策略，分别提取 13 维的 MFCC 特征、Fbank 特征、短时能量三种特征组成 3×13 维的 MFE 联合特征，特征向量经数据归一化后作为神经网络的输入. MFE 联合特征能够有效弥补单帧预测引起的输入信息不足，同时 MFE 联合多种音频特征作为模型输入，音频信号经过 MFE 联合特征抽取的预处理，初步抽象出了音频信号的高维特征，其计算代价要远远小于原始波形、LPS 等特征，大大减少了模型的计算时间.

首先，MFE 联合特征经过 3 层卷积模块，每个卷积模块包含 Conv2D、BatchNorm、MaxPooling 和 Dropout 等 4 个部分，每层使用 ELU[17-18]非线性激活函数. 其表示如式(1)所示，X 和 Y 分别表示卷积模块的输入和输出矩阵，$\beta(x)$、$\varphi(x)$、$\delta(x)$ 分别表示归一化、非线性激活和下采样的过程. $\mathrm{conv}(x,\boldsymbol{W})$是卷积层，其主要作用是进一步的聚合 MFE 联合特征，得到高维语音信号的时域和频域非线性特征. 其中，$\boldsymbol{W}$ 是权重矩阵，$\boldsymbol{b}$ 是偏置矩阵. $\beta(x)$基于卷积操作参数共享的优势对卷积结果进行归一化以减小数据分布的离散度，可以加快模型收敛速度，大大减少模型训练时间. $\delta(x)$是对特征进行下采样，在保留主要特征的同时，对数据降维处理. 同时，下采样操作可以有效防止过拟合，减少网络参数，增强模型的泛化能力. Dropout 负责剪枝不必要的网络参数，加快模型计算速度.

$$Y=\delta(\varphi(\beta(\mathrm{conv}(X,\boldsymbol{W})+\boldsymbol{b}))) \tag{1}$$

随后，将 CNN 卷积模块提取出的非线性特征馈入 GRU 网络层[19-20]. GRU 门控循环神经网络是 RNN 的变体，它引入了重置门(reset gate)和更新门(update gate)概念. 假设给定时间步 t 的语音高维聚合特征 X_t 和上一时间步的隐藏状态 H_t-1，重置门 R_t、更新门 Z_t 的计算如式(2)和式(3)所示.

$$\boldsymbol{R}_t=\sigma(X_t\boldsymbol{W}_{xr}+H_t-1\boldsymbol{W}_{hr}+b_r) \tag{2}$$

$$\boldsymbol{Z}_t=\sigma(X_t\boldsymbol{W}_{xz}+H_t-1\boldsymbol{W}_{hz}+b_z) \tag{3}$$

其中，$\sigma(x)$为激活函数，$\boldsymbol{W}_{xr}$、$\boldsymbol{W}_{hr}$、$\boldsymbol{W}_{xz}$、$\boldsymbol{W}_{hz}$是权重矩阵，$\boldsymbol{b}_r$、$\boldsymbol{b}_z$ 是偏置矩阵. GRU 层通过可学习的门控单元控制信息流动，捕捉短时平稳的音频信号内部的变化关系，有助于提高分类精度. 并且 GRU 在保持 RNN 特性的同时又拥有更加简单的结构，大大减少了训练时间和训练难度. 最后，由 softmax 层输出音频帧的所属类别的概率.

本文提出的 CGRU 网络结构如图 2(a)所示，网络参数细节如表 1 所示. CGRU 网络改进了前期研究[6]CNN 网络结构(图 2(c))中的卷积模块，采用 3×3 的小卷积核，在保证足够感受野的前提下，减少了网络参数，并且在卷积过程中加入 batch normalization 层，以提升训练速度和模型精度. 并且在卷积模块之后加入 GRU 网络层捕获音频信号的时序变化，使得网络的特征提取能力显著提升. 在实验阶段，本文也实现了文献[9]中的 RAW CLDNN(图 2(b))方法，与之相比本文提出的 MFE 输入特征经过音频信号预处理更加适用于复杂环境，而原始波形作为输入易受环境影响，泛华能力相对较弱. 并且，使用 GRU 网络层代替 LSTM，可以缩减训练时间，降低训练难度，更适用于工程应用.

表 1 CGRU 网络参数表

Tab. 1 Configurations of the proposed CGRU

ConvLayers	Conv1	Conv2	Conv3
Filter outputs	16	32	64
Filter size	3×3	3× 3	3×3
Pooling size	1× 3	1× 6	1× 9
Params	264	4688	18520
GRULayers	GRU1	GRU2	
Hidden units	32	32	
Params	9 312	6 240	
Prediction Layer	Softmax		
Hidden units	3		
Total params	39 123		

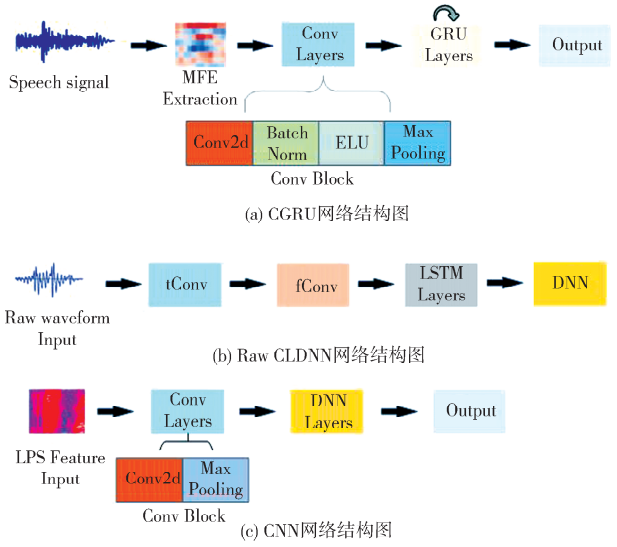


图 2 网络结构图

Fig. 2 Network structure

实验结果表明，CGRU 网络结构中的 CNN 卷积模块可以抽取地空通话语音中的语音帧、结束帧以及不稳定的噪声帧的局部特征，GRU 门控循环单元能捕捉帧内信息短时变化的依赖关系，能较好的完成帧分类任务. 并且该网络结构简单，模型总参数不足 40 K，能够满足实时切分的需要.

3.2 平滑算法

本文从地空通话内话系统引接音频信号到专业音频采集设备，编程读取实时语音流，并进行音频信号分帧、预处理等操作，然后馈入训练好的模型预测所属类别，完成切分任务. 为了提高切分的准确率，降低语音帧间的短暂停顿、信道噪声等因素对切分效果影响，本文提出了如下平滑算法. 对于输入音频帧序列 $X_n=\{x_1,x_2,x_3,\cdots\cdots,x_{n-1},x_n\}(n>0)$，分类模型预测类别序列为 $Y_n=\{y_1,y_2,y_3,\cdots\cdots,y_{n-1},y_n\}$，$y_i=\{l_s,l_e,l_o\}$. 其中，$y_i$ 为 one-hot 编码，当 l_s 为真时表示输出结果为语音帧；l_e 为真时表示输出结果为结束帧；l_o 为真时表示输出结果为其它帧. 帧的最终标签由预测本身以及上下文共同决定. 当音频帧 x_i 的预测结果 y_i 为语音，且语音帧子序列$\{x_{i-1}\cdots\cdots x_{i-m}\}$对应的预测序列$\{y_{i-1}\cdots\cdots y_{i-m}\}$中的语音帧 l_s 之和大于 ξ，则认为检测到语音开始. 其中，m，ξ 为适应性参数，根据当前传输信道质量等因素设置. 当语音开始后，若检测到 x_i 结束帧 l_e，认为语音结束，单句实时切分完成；同理，若检测到 x_i 为其他帧，当语音帧子序列$\{x_{i+1}\cdots\cdots x_{i+m'}\}$预测序列$\{y_{i-1}\cdots\cdots y_{i+m'}\}$中其他帧 l_o 之和大于 μ，认为语音结束，否则认为是不平稳的短噪声. 定义语音开始端点为 ssp(Speech start point)，结束端点为 sep(Speech end point)，则语音端点 L_i 计算方法如式(4)所示.

$$L_i=\begin{cases}\text{ssp, if } \sum_{k=i-m}^{i} y_k \geqslant \xi \\ \qquad (m>0,0<\xi\leqslant m) \\ \text{sep, if } y_k=l_e \text{ or } \sum_{k=i-m}^{i} y_k \geqslant \mu \\ \qquad (m>0,0<\mu\leqslant m)\end{cases} \tag{4}$$

通过调整平滑算法中的参数能够避免信道中的不稳定噪声以及通话时的短暂停顿引起的抖动，避免将句子切断，从而保证语料的完整性，为后端应用提供可靠的输入. 因此，平滑算法能一定程度上提升语音切分的准确率.

4 实 验

4.1 数据集与实验环境

与文献[6]中数据集不同，本文实验数据取自成都、太原、北京和上海等区域管制中心的历史地

空通话语音数据，数据中复杂环境下的带噪语音占比较高. 该数据经过人工清洗、标注后作为实验数据集. 本文采用 8 K 采样率、16 bit 采样精度的原始音频样本数据，总时长约 100 h. 其中包括语音数据时长约 45 h，静音/噪音时长约 50 h，结束帧时长约 5 h. 实验中将原始数据分为以下子集以验证模型性能：80% 为训练集，10% 为验证集，其余 10% 用作测试集.

训练服务器采用 Ubuntu 16.04 操作系统，NVIDIA GTX 1080 显卡提高模型训练速度. 测试环境严格仿真地空通话生产环境，采用模拟音频信号仿真地空通话内话系统作为专业音频采集设备的输入.

4.2　对比实验设计

GMM-webrtcvad：webrtcvad 是 google 开源的语音端点检测工具，该算法基于 GMM 提取音频帧子带能量对语音/非语音建立统计学模型，使用假设检验的方法确定音频帧的类型，是一种无监督的学习方法. 其主要特点是简单易用、适用场景广泛，并且模型参数根据时间上下文实时更新，目前在工业界颇受欢迎. 经多次实验，将其初始化参数设置为 2(aggressive mode)，帧长设置为 30 ms，在地空通话信道中效果达到最优，在本次对比实验中均采用最优参数.

CNN：实验把本文的前期研究[6]中的 CNN 网络的 softmax 层的输出神经元修改为 3 个，选取帧长 32 ms 为一帧，提取 1×256 的 LPS 特征向量作为网络输入. 损失函数尊选取交叉熵函数，优化器选择 SGD(Stochastic Gradient Descent)算法，网络参数采用 glorot uniform 算法进行初始化，配置学习率为 0.01，batch size 设置为 80，训练至网络收敛.

Raw Waveform CLDNN：本文实现了文献[9]中表现较好 CLDNN_100 K 的网络，将输出改为 3 个神经元，旨在探究其在地空通话中的应用效果. 实验选取帧长 35 ms 的音频帧作为输入，使用 ASGD(Asynchronous Stochastic Gradient Descent)算法作为优化器、交叉熵损失函数.

CGRU：为验证 MFE 联合特征的有效性，除本文提出的 MFE 特征输入的 CGRU 网络结构之外，实验中还对比了 MFCC、Fbank 单独作为输入特征的分类效果. 实验中均取 35 ms 帧长，网络细节及参数与表 1 描述相同.

此外，实验对比了上述所有分类器原生切分效果和加入本文提出的平滑算法后的切分效果，以验证平滑算法的有效性.

4.3　实验结果与分析

4.3.1　评价标准　ROC(Receiver Operator Characteristic Curve)曲线又称受试者工作特征曲线，是反映敏感度和特异度连续变量的综合指标，其特点是在数据样本不均衡的情况下可以直观的评估分类器性能. 而受限于地空通话数据特点，数据集结束帧样本占比较低. 因此，实验使用 ROC 曲线作为分类器性能的评估方法. AUC (Area Under Curve)值是指 ROC 曲线下的面积，是定量评价分类器性能的指标.

4.3.2　实验结果与分析　实验结果如表 2 所示，其中 Accuracy 指未使用本文提出的平滑算法的准确率，Accuracy-S 代表平滑过后的切分准确率，Delay 代表预测一帧的时间代价. 实验结果表明，本文提出的平滑算法根据模型的性能不同，将切分的准确率提升了约 1%～9% 不等. 同时，平滑后的准确率提升幅度可以作为衡量各方法稳定性的依据，准确率提升幅度越高，说明相邻帧之间预测结果抖动越大，方法在测试集上越不稳定.

表 2　实验结果

Tab. 2　Experimental result

Method	AUC	Accuracy /%	Accuracy-S /%	Params /K
Webrtcvad (GMM)	—	83.2	88.1	—
CNN (LPS)	0.87	89.7	93.2	5.6
CGRU (LPS)	0.91	90.0	94.9	4.0
CGRU(MFCC)	0.89	89.4	93.6	4.0
CGRU(Fbank)	0.94	94.4	98.0	4.0
CLDNN (Raw)	0.84	82.3	91.5	100
CGRU (MFE)	0.98	98.5	99.3	3.9

从实验结果来看，基于 GMM 无监督学习的 webrtcvad 并不适用于复杂环境的地空通话语音切分，在仿真测试集上准确率仅有 83.2%，加入平滑算法后准确率约提升了 5%，在不稳定噪声环境下预测结果抖动较大. 在基于深度学习的方法中，基于 LPS 特征的 CNN、CGRU 网络以及基于 MFCC 的 CGRU 网络准确率在 90% 左右，经平滑后效果提升约 3%，帧之间预测结果也存在抖动. 基于 Fbank 特征的 CGRU 网络模型准确率表现良好，AUC 值达到了 0.95，平滑后切分准确率达到 98%. 基于原始波形输入的 Raw CLDNN 网络在

地空通话中准确率仅有 82.3%,经平滑后准确率提升了约 9%,幅度较大,ROC 曲线对比图如图 3 所示.经分析,该方法使用原始采样数据作为输入,原始采样数据在地空通话中受不稳定噪声、采样设备、说话人等因素的影响较大,导致测试集输出结果与训练集差别较大,同时,该网络参数较多,时间代价约是其他网络的一倍,并不适用于地空通话的切分.基于 MFE 联合特征的 CGRU 网络在仿真测试集上表现最好,分类器准确率达到 98.5%,AUC 值为 0.98,经平滑后切分准确率约 99.3%,预测输出较稳定.

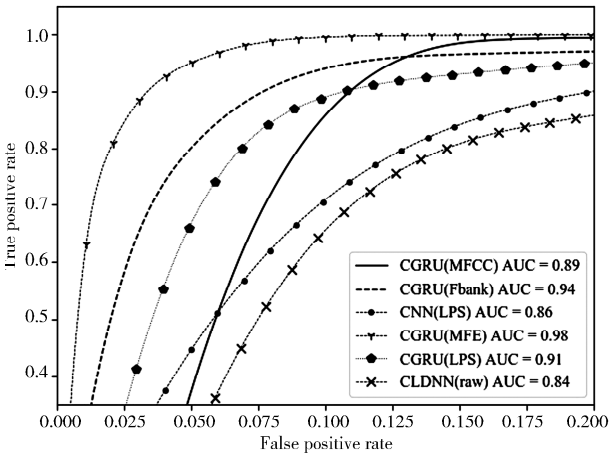


图 3 ROC 曲线对比图
Fig. 3 Chart of ROC curve

由图 3 可知,本文提出的 MEF 联合特征在音频信息有限的单帧预测策略上具有明显优势,并且 CGRU 网络结构在进一步深入挖掘音频信号隐藏信息的同时,优化了模型参数,缩短了模型预测的时间代价,是一种稳定、高效的地空通话实时切分方法.

5 结 论

本文在基于空管语音识别的安全防护系统的应用背景下,提出了一种基于 CGRU 神经网络的地空通话语音实时切分的方法.该方法基于对地空通话特点的全面分析以及地空通话语音特征的深入挖掘的基础上,经过多次对比试验,采用 MFE 联合特征输入的方式训练语音帧分类器.同时,在严格、精确地对语音帧分类情况下,为了应对信道中的不稳定噪声、不同说话人的语速习惯等,采用单帧预测、多帧预测结果平滑的方法,从一定程度上辅助语音切分,提高了语音切分准确率.与已有语音切分方法相比,本文提出的方法具有明显优势,为后端语音降噪、语音识别、声纹识别和语义理解等应用提供了可靠的语料输入.但平滑算法参数需要人为参照生产环境的复杂度设置,并非自适应参数,语音帧中语种、说话人性别等信息还待进一步挖掘.因此,平滑算法自适应参数的改进,继续挖掘语音帧中的隐藏信息将是下一步工作的重点.

参考文献:

[1] Lin Y, Tan X, Yang B, *et al*. Real-time controlling dynamics sensing in air traffic system [J]. Sensors, 2019, 19: 679.

[2] 路青起, 白燕燕. 基于双门限两级判决的语音端点检测方法[J]. 电子科技, 2012, 25: 13.

[3] Misra A. Speech/nonspeech segmentation in web videos[C]//Thirteenth Annual Conference of the International Speech Communication Association. Portland, OR, USA: INTERSPEECH, 2012.

[4] Hwang Y, Mun-Ho J, Oh S R, *et al*. Applying the Bi-level HMM for robust voice-activity detection [J]. J Electr Eng Technol, 2017, 12: 373.

[5] Chaitanya K, Sinha R. Energy and entropy based switching algorithm for speech endpoint detection in varying SNR conditions [C]// Proceedings of the Interspeech, Conference of the International Speech Communication Association. Brisbane, Australia, September: DBLP, 2008.

[6] 郭东岳, 杨波, 高登峰, 等. 基于 CNN 的空管地空通话自动切分[C]//第一届空中交通管理系统技术学术年会论文集. 北京: 电子工业出版社, 2018.

[7] Vesperini F, Vecchiotti P, Principi E, *et al*. Deep neural networks for multi-room voice activity detection: advancements and comparative evaluation [C]// Proceedings of the International Joint Conference on Neural Networks. Vancouver, BC, Canada: IEEE, 2016.

[8] Drugman T, Stylianou Y, Kida Y, *et al*. Voice activity detection: merging source and filter-based information [J]. IEEE Signal Proc Lett, 2016, 23: 252.

[9] Zazo R, Sainath T N, Simko G, *et al*. Feature learning with raw-waveform CLDNNs for voice activity detection [C]// Interspeech 2016, 17th Annual Conference of the International Speech Communication Association. San Francisco, CA, USA: Interspeech, 2016.

[10] Silva D A, Stuchi J A, Violato R P V, *et al*. Ex-

ploring convolutional neural networks for voice activity detection [M]//Cognitive technologies. Cham: Springer, 2017.

[11] Hughes T, Mierle K, Hughes T, *et al*. Recurrent neural networks for voice activity detection[C]// Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver, BC, Canada: IEEE, 2013.

[12] Eyben F, Weninger F, Squartini S, *et al*. Real-life voice activity detection with LSTM recurrent neural networks and an application to hollywood movies [C]//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver, BC, Canada: IEEE, 2013.

[13] Kim J, Kim J, Lee S, *et al*. Vowel based voice activity detection with LSTM recurrent neural network[C] // Proceedings of the 8th International Conference on Signal Processing Systems. Auckland, New Zealand USA: Association for Computing Machinery, 2016.

[14] Thomas S, Ganapathy S, Saon G, *et al*. Analyzing convolutional neural networks for speech activity detection in mismatched acoustic conditions [C]// Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Florence, Italy: IEEE, 2014.

[15] 王万乐，赵琦. 空中交通管制员无线电陆空通话[M]. 北京：清华大学出版社，2016.

[16] Choi K, Fazekas G, Sandler M, *et al*. Convolutional recurrent neural networks for music classification [C]// Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New Orleans, LA, USA: IEEE, 2017.

[17] Clevert D A, Unterthiner T, Hochreiter S. Fast and accurate deep network learning by exponential linear units (elus)[J]. arXiv preprint arXiv, 2015, 1511: 07289.

[18] 刘宇晴，王天昊，徐旭. 深度学习神经网络的新型自适应激活函数[J]. 吉林大学学报：理学版，2019，57：857.

[19] Chung J, Gulcehre C, Cho K H, *et al*. Empirical evaluation of gated recurrent neural networks on equence modeling[J]. arXiv preprint arXiv, 2014, 1412: 3555.

[20] 吴昊，平鹏，孙立博，等. 基于改进 LRCN 模型的驾驶行为图像序列识别方法[J]. 江苏大学学报：自然科学版，2018，39：303.

引用本文格式：

中　文：郭东岳，林毅，杨波. 基于 CGRU 多输入特征的地空通话自动切分[J]. 四川大学学报：自然科学版，2020，57：887.

英　文：Guo D Y, Lin Y, Yang B. Automatic speech segmentation for air-ground communication based on multi-input CGRU neural network [J]. J Sichuan Univ: Nat Sci Ed, 2020, 57: 887.