doi:10.12068/j.issn.1005－3026.2024.06.005

基于VMD的双通道构音障碍语音特征图谱提取算法

薛珮芸[1,2]，白 静[1]，张 楠[3]，赵建星[1]

（1.太原理工大学 电子信息工程学院，山西 晋中 030600；2.山西高等创新研究院 博士后科研工作站，山西 太原 030024；3.中北大学 信息与通信工程学院，山西 太原 030024）

摘 要：针对在提取构音障碍患者语音有效特征信息不足，导致语音识别率低的问题，提出一种基于变分模态分解（VMD）的多尺度双通道滤波器组（MBCFbank）特征图谱提取算法．首先，为了更好地提取符合人耳听觉结构特性的声学特征，提出一种双通道滤波器组（BCFbank）特征提取算法，该算法采用Mel滤波后做对数变换，同时采用Gammatone滤波后作非线性响度变换；其次，采用VMD来优化BCFbank特征，对分解后的多个语音信号分量筛选出相关系数较高的3个，分别提取其BCFbank特征及其差分特征，同时对未分解的语音信号提取BCFbank特征，从而构成MBCFbank特征图谱；最后，在双路语音识别模型上进行训练和识别．实验结果表明，基于BCFbank特征、MBCFbank特征图谱的语音识别模型准确率最高分别达到了87.82%，94.34%，优于Fbank特征的识别效果．

关 键 词：构音障碍语音识别；变分模态分解；卷积神经网络；MBCFbank特征

中图分类号：TP 912.34；R 741 **文献标志码**：A **文章编号**：1005-3026(2024)06-0793-09

VMD Based Binary Channels Speech Feature Map Extraction Algorithm for Dysarthria

XUE Pei-yun[1,2], *BAI Jing*[1], *ZHANG Nan*[3], *ZHAO Jian-xing*[1]

(1. College of Electronic Information Engineering, Taiyuan University of Technology, Jinzhong 030600, China; 2. Post-doctoral Research Station, Shanxi Academy of Advanced Research and Innovation, Taiyuan 030024, China; 3. School of Information and Communication Engineering, North University of China, Taiyuan 030024, China. Corresponding author: XUE Pei-yun, E-mail: xuepeiyun@tyut.edu.cn)

Abstract: A multiscale binary channels filter banks (MBCFbank) feature extraction algorithm based on variational modal decomposition (VMD) is proposed to address the issue of poor speech recognition caused by insufficient extraction of effective feature information from speech of patients with dysarthria. Firstly, in order to better extract the acoustic features that conform to the structural characteristics of human ears, a binary-channels filter banks (BCFbank) feature extraction algorithm is proposed, which uses Mel filtering and performs logarithmic transformation, simultaneously using Gammatone filtering to perform nonlinear loudness transformation. Secondly, VMD is used to optimize the BCFbank features. Three components with higher correlation coefficients are selected from the decomposed multiple speech signal components, and their BCFbank features and differential features are extracted respectively. At the same time, BCFbank features are extracted from the undecomposed speech signals to form the MBCFbank feature map spectrum. Finally, training and recognition are conducted on a dual channel speech recognition model. The experimental results show that the speech recognition model based on BCFbank features and MBCFbank feature maps has the highest accuracy of 87.82% and 94.34%, respectively, which is superior to the recognition effect of Fbank features.

收稿日期：2023-05-23

基金项目：山西省应用基础研究计划项目(201901D111094)；山西省基础研究项目(青年)(20210302124544).

作者简介：薛珮芸(1990-)，女，山西太原人，太原理工大学讲师，博士；白 静(1965-)，女，山西太原人，太原理工大学教授．

Key words: speech recognition with dysarthria; variational mode decomposition; convolutional neural network; MBCFbank features

随着语音识别技术的发展,将语音信号处理和计算机技术相结合用于解决医疗领域问题成为目前研究的热点[1-5].构音障碍主要是人体发音器官产生病变所引发的发音障碍,此类构音障碍患者能进行发声,但发音迟缓、发出的语音往往含糊不清,导致患者不能进行有效的语言交流,给日常生活带来极大的不便,因此针对构音障碍患者病理语音的研究受到广泛关注. Jiao等[6]提出一种通过使用对抗训练将健康语音转换为构音障碍语音来模拟临床应用的训练数据方法,采用卷积神经网络(convolutional neural networks, CNN)构建构音障碍语音的生成器和鉴别器. Yilmaz等[7]提出一种基于时间和频率的CNN搭建构音障碍语音识别模型,对Gammatone滤波器组特征进行时间和频率的卷积,使得能够联合使用来自声学和发音空间的信息. Zaidi等[8]将提取梅尔频率倒谱系数、梅尔频率频谱系数和感知线性预测特征送入CNN和长短期记忆神经网络模型上进行训练和构音障碍语音识别实验对比,但实验过程中提取的构音障碍语音特征信息不足. Mariya等[9]提出一种基于迁移学习的数据增强技术,用于增强构音障碍语音信号,在UA-Speech数据库上提升构音障碍语音识别效果.李东等[10]提取构音障碍语音的韵律特征及梅尔倒谱系数特征构成融合特征,在随机森林分类器上进行正常语音和病理语音二分类,在提取特征过程中遗漏了语音数据之间的相关性,只能检测出病理语音却不能识别该患者具体说的话.吴丽丹[11]对隐马尔可夫模型、深度神经网络模型进行改进,提出基于对比、多视图学习的构音障碍语音识别,能够捕获构音障碍语音的差异性,但对语音特征提取算法研究较少.尽管国内的学者对构音障碍患者的病理语音已经做了一些研究工作,但是针对构音障碍患者的语音识别技术研究还相对较少,对于构音障碍患者语音特征信息提取不足,使得构音障碍语音识别率难以提升.

通过分析存在的问题,首先,本文提出一种双通道滤波器组(binary channels filter banks, BCFbank)特征提取算法,该算法采用两条支路提取语音特征信息,结合了Mel滤波和Gammatone滤波的优势,能够弥补Mel滤波过程中遗漏掉的有效语音特征信息;其次,对BCFbank特征进行优化,本文提出一种多尺度双通道滤波器组(multiscale binary channels filter banks, MBCFbank)特征图谱提取算法,该算法结合了变分模态分解(variational mode decomposition, VMD)和BCFbank特征两者的优势,能够有效地分解构音障碍语音信号,避免出现模态混叠现象,同时有效地提取符合人耳听觉特性的语音特征信息;最后,设计了不同声学特征送入搭建的基于CNN的双路构音障碍语音识别模型进行训练和语音识别对比实验,来验证本文所提特征提取算法的有效性.

1 基于滤波器组的BCFbank特征提取算法

Gammatone滤波器与人耳耳蜗基底膜滤波器相似,可以反映人类听觉耳蜗滤波器的心理和生理反应,能够仿造人类的听觉特性来处理语音信号.为了能够有效地提取符合人类听觉特性的语音特征,本文提出一种BCFbank特征提取算法,该算法采用两条支路对构音障碍语音信号进行特征参数提取,一条支路利用Mel滤波器组对经过短时傅里叶变换(short-time Fourier transform, STFT)的语音信号进行滤波,再进行对数变换来压缩特征参数;另一条支路采用Gammatone滤波器组对经过快速傅里叶变换(fast Fourier transform, FFT)的语音信号进行滤波,再进行非线性响度变换来模拟人耳耳蜗基底膜的特性;将两条支路提取的特征参数进行叠加得到BCFbank特征参数. BCFbank特征参数提取过程如图1所示.

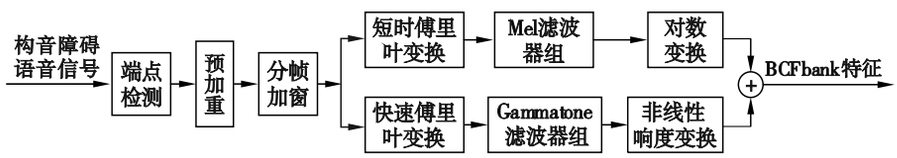

图1 BCFbank特征提取流程图

Fig. 1 BCFbank feature extraction flow chart

BCFbank特征具体提取过程如下:

1) 假设输入构音障碍语音信号为$x(n)$,对$x(n)$进行预处理操作,首先采用数字滤波器对语音信号进行预加重,消除口唇辐射影响,突出语音信号的高频部分,由于语音信号在短时间内(一般为10~30 ms)是平稳的,可将语音信号分成短帧,分帧需要加窗来实现,汉明窗能够让整体语音信息更加连贯,避免语音细节信息丢失,因此本文选用汉明窗与短帧相乘完成分帧加窗处理,从而得到语音信号的序列帧$s(n)$.

2) 其中一条支路对$s(n)$进行STFT得出$S(n)$,再对语音频谱$S(n)$取平方得到能量谱$|S(n)|^2$.

3) $|S(n)|^2$经过Mel滤波后得到能量谱,第m个Mel滤波器的传递函数$H_m(k)$可以表示为

$$H_m(k)=\begin{cases}\dfrac{2\left(k-f(m-1)\right)}{\left(f(m+1)-f(m-1)\right)\left(f(m)-f(m-1)\right)}, & f(m-1)\leqslant k\leqslant f(m);\\ 0, & k=\text{其他};\\ \dfrac{2\left(f(m+1)-k\right)}{\left(f(m+1)-f(m-1)\right)\left(f(m+1)-f(m)\right)}, & f(m)<k\leqslant f(m+1).\end{cases}\tag{1}$$

其中:$f(m)$代表Mel滤波器组中第m个频谱的中心频率,其公式为

$$f(m)=F_{\text{Mel}}\left(f_{\text{l}}\right)F_{\text{Mel}}{}^{-1}\left(\frac{N}{F_{\text{s}}}\right)+F_{\text{Mel}}{}^{-1}m\frac{F_{\text{Mel}}\left(f_{\text{h}}\right)-F_{\text{Mel}}\left(f_{\text{l}}\right)}{M+1}\left(\frac{N}{F_{\text{s}}}\right).\tag{2}$$

其中:f_{h},f_{l},M分别为Mel滤波器的最高频率、最低频率、个数;F_{s}为傅里叶变换的点数.在本文实验中,M取40,F_{s}取512.

$F_{\text{Mel}}(f)$为梅尔频率,可以表示为

$$F_{\text{Mel}}(f)=2\,595\times\lg\left(1+\frac{f}{700}\right).\tag{3}$$

其中:f表示频率,Hz;$F_{\text{Mel}}{}^{-1}$为F_{Mel}的逆函数;关于f求逆,可以表示为

$$F_{\text{Mel}}^{-1}(f)=700\times\left(10^{\frac{f}{2595}}-1\right).\tag{4}$$

4) 采用对数变换对步骤3)所得能量谱进行压缩得到$L(m)$,计算过程为

$$L(m)=\lg\left(\sum_{k=0}^{N-1}H_m(k)\left|S(k)\right|^2\right),\quad 0\leqslant m<M.\tag{5}$$

5) 另一条支路对$s(n)$进行FFT得出语音序列频谱$P(n)$.再对频谱$P(n)$取平方得到能量谱$|P(n)|^2$.

6) $|P(n)|^2$经过Gammatone滤波器组,得到滤波后的能量谱,滤波器组的时域脉冲响应为

$$g_i(t)=At^{n-1}\cos(2\pi f_i t+\phi_i)\times \mathrm{e}^{-2\pi b_i t}u(t),\quad t\geqslant 0,\ 1\leqslant i\leqslant N.\tag{6}$$

其中:A代表滤波器的增益;f_i代表滤波器的中心频率;$u(t)$为阶跃函数;ϕ_i是相位;n是滤波器阶数;N为滤波器个数;b_i代表滤波器衰减因子[12]. b_i的大小决定了滤波器组时域脉冲响应衰减的快慢,可以表示为

$$b_i=24.7\left(\frac{4.37f_i}{1\,000}+1\right).\tag{7}$$

7) 对步骤6)得到的能量谱,再进行非线性响度变换,模拟人耳听觉特性在不同频率段下计算不同的幂指数,计算公式为

$$G(m)=\left(\sum_{k=0}^{N-1}H_{\text{G}}(k)\left|P(k)\right|^2\right)^{\alpha},\quad 0\leqslant m<M.\tag{8}$$

式中:$H_{\text{G}}(k)$代表语音信号经过第k个Gammatone滤波器进行滤波;α代表不同频率段下的幂指数值,计算公式为

$$\alpha=\begin{cases}0.1\left(f-1\,000\right)/1\,000+1/3, & 0\leqslant f<1\,000;\\ 0.1\left(f-2\,000\right)/2\,000+1/3, & 1\,000\leqslant f<2\,000;\\ 0.1\left(f-3\,000\right)/3\,000+1/3, & 2\,000\leqslant f<3\,000;\\ 0.1\left(f-4\,000\right)/4\,000+1/3, & 3\,000\leqslant f<4\,000;\\ 0.1\left(f-5\,000\right)/5\,000+1/3, & 4\,000\leqslant f<5\,000;\\ 0.1\left(f-6\,000\right)/6\,000+1/3, & 5\,000\leqslant f<6\,000;\\ 0.1\left(f-7\,000\right)/7\,000+1/3, & 6\,000\leqslant f<7\,000;\\ 0.1\left(f-8\,000\right)/8\,000+1/3, & 7\,000\leqslant f<8\,000;\\ 1/3, & f\geqslant 8\,000.\end{cases}\tag{9}$$

其中,f代表滤波器组的频率.

8) 将步骤4)得到的$L(m)$与步骤7)得到的$G(m)$进行叠加,得到BCFbank特征参数.

2　基于VMD的BCFbank特征图谱提取算法

VMD方法能够自适应地对构音障碍患者的语音信号进行有效分解、频率划分,对上述

BCFbank特征提取算法进一步优化,本文提出一种MBCFbank特征图谱提取算法,该算法采用VMD方法对构音障碍患者的语音信号进行有效分解,再使用相关系数分析对分解后的语音信号分量(voice signal components,VSC)筛选出3个相关度较高的有效VSC,对筛选出的3个有效VSC分别提取BCFbank特征及其差分特征,同时对未分解的构音障碍语音信号直接提取BCFbank特征,将上述提取的各帧特征进行拼接构成MBCFbank特征图谱. MBCFbank特征图谱提取过程如图2所示.

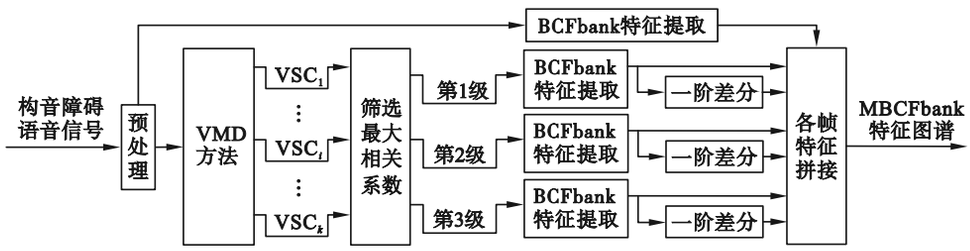

图2 MBCFbank特征图谱提取流程图

Fig. 2 MBCFbank feature map extraction flow chart

MBCFbank特征图谱提取算法的具体实现步骤如下:

1)对构音障碍患者的语音信号先进行预处理操作后,采用VMD方法[13]对语音信号分解出k个VSC.

2)采用Spearman Rank相关系数[14]从k个VSC中筛选出3个相关度较高的有效VSC,Spearman Rank相关系数ρ计算过程为

$$\rho=\frac{\sum_{n=1}^{N}[y_n(t)-\bar{y}][x_n(t)-\bar{x}]}{\sqrt{\sum_{n=1}^{N}[y_n(t)-\bar{y}]^2\sum_{n=1}^{N}[x_n(t)-\bar{x}]^2}}. \tag{10}$$

其中:$\bar{y},\bar{x}$分别表示$y_n(t),x_n(t)$的均值;N表示语音信号分帧的总个数.

3)对3个相关系数较高的有效VSC分别提取BCFbank特征,同时对经过构音障碍语音信号预处理操作后的未分解语音信号直接提取BCFbank特征,BCFbank特征提取过程见图1.

4)对筛选出3个有效VSC的BCFbank特征求取差分特征,一阶差分计算过程为

$$\Delta B_i(n,m)=\frac{\sum_{z=1}^{2}z(B_i(n+z,m)-B_i(n-z,m))}{\sqrt{2\sum_{z=1}^{2}z^2}}. \tag{11}$$

其中:$B_i(n,m)$代表第i个有效VSC的BCFbank特征;$\Delta B_i(n,m)$代表第i个有效VSC的BCFbank特征的差分特征;m表示特征维度;n表示语音信号进行分帧操作后的第n帧.

5)将得到第n帧的未分解构音障碍语音信号的BCFbank特征、VMD方法分解得到相关系数较高的3个有效VSC的BCFbank特征及其差分特征有效地拼接在一起,构成构音障碍语音信号第n帧的组合特征$C(n,d)$,可以表示为

$$C(n,d)=[B(n,m);B_1(n,m);\Delta B_1(n,m);$$
$$B_2(n,m);\Delta B_2(n,m);B_3(n,m);\Delta B_3(n,m)]. \tag{12}$$

其中:d=7m表示特征维度大小;$B(n,m)$表示第n帧构音障碍语音信号的BCFbank特征.

6)将各帧组合特征$C(n,d)$拼接在一起,得到MBCFbank特征图谱.

3 基于CNN的双路构音障碍语音识别模型

CNN在处理二维图形数据上优势突出,能够不断对参数进行更新并使整个网络具有一定的移动性、缩放性和非线性的稳定性;深度可分离卷积(depthwise separable convolution,DSC)能有效地提取网络模型中不同通道以及空间维度上的语音特征信息,同时能够减少特征参数在训练网络模型过程中的参数量以及降低模型内部运算的复杂度.结合CNN和DSC两者优势,本文搭建了一种基于CNN的双路构音障碍语音识别模型结构,该结构采用两条支路对语音特征信息进行学习,一条单路CNN支路采用卷积层、池化层进行串联对特征信息学习,另一条单路CNN+DSC支路前部分使用卷积层、池化层进行串联,后部分利用卷积层、深度可分离卷积层以及池化层交替使用进行特征信息学习;将经过单路CNN支路与单路CNN+DSC支路学习得到的语音特征信息采用Concat函数进行有效地拼接在一起

得到新的隐层语音特征图,送入全连接层,再采用联结时序分类(connectionist temporal classification,CTC)算法构建语音识别网络模型.基于CNN的构音障碍语音识别模型的总体结构如图3所示.

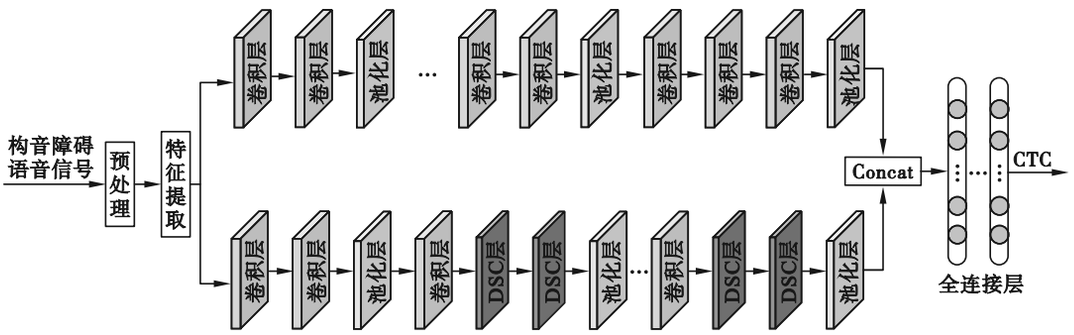

图3　基于CNN的双路构音障碍语音识别模型

Fig. 3　A CNN based binary channel speech recognition model for dysarthria

DSC层分两部分对语音特征信息进行学习,先经过深度卷积层进行卷积运算,再经过点卷积层将不同通道上的语音特征信息有效地拼接在一起,由此搭建的模型能够减少特征参数在训练网络模型过程中的参数量以及降低模型内部运算的复杂度.DSC结构如图4所示.

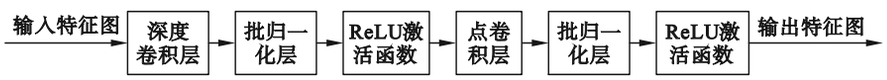

图4　深度可分离卷积结构图

Fig. 4　Deep separable convolutional structure diagram

CTC放在构音障碍语音识别模型最后,对输入的语音信号特征经过构音障碍语音识别网络模型训练之后与真实的构音障碍语音信号输出之间的差别进行一定程度的衡量.

假设输入构音障碍语音序列和对应输出语音序列为$X=(x_1,x_2,\cdots,x_T)$和$Y=(y_1,y_2,\cdots,y_U)$,其中T是语音信号经过加窗长度;x_i表示第i个加窗得到的语音序列;U为输出语音序列的音节个数,$y_i\in L$表示输出的第i个;L表示输出语音序列的集合.CTC引入了一个代表空白的"blank"块来扩充L,对应CTC搜索路径为$\pi=(\pi_1,\pi_2,\cdots,\pi_T)$,整个CTC搜索路径概率可以表示为

$$P(\pi\mid X)=\prod_{t=1}^{T}P\left(\pi_t\mid X_t\right). \tag{13}$$

由于Y能够与多个π相对应,因此用全部的CTC搜索路径概率来表示Y的概率,可以表示为

$$P(Y\mid X)=\sum\nolimits_{p\in\beta(Y)}P(p\mid X). \tag{14}$$

其中,β表示π向Y映射的一种关系.

4　实验结果分析

4.1　构音障碍语音数据集筛选

本文实验采用UA-Speech数据集中的构音障碍患者发音语料,如表1所示,使用了12名患有构音障碍的男性受试者(M01,M04,M05,M06,M07,M08,M09,M10,M11,M12,M14和M16)和3名患有构音障碍的女性受试者(F02,F03和F04)的语音.实验在上述人群中筛选出6 264个语音样本,每个单词发音包含216个样本,其中训练集包含4 640个语音样本,测试集包含1 273个语音样本,验证集包含351个语音样本,每个语音样本在训练、测试、验证集中不会重复出现.

表1　构音障碍患者发音语料表

Table 1　Pronunciation corpus of patients with dysarthria

语料名称	发音项	合计
计算机命令单词	ESCAPE,LEFT,LINE,PARAGRAPH,PASTE,RIGHT,SHIFT,SENTENCE,TAB,ALT,BACKSPACE,COMMAND,CONTROL,COPY,CUT,DELETE,DOWNWARD,ENTER,UPWARD	19
数字单词	ZERO,ONE,TWO,THREE,FOUR,FIVE,SIX,SEVEN,EIGHT,NINE	10

4.2 实验

本文进行构音障碍语音识别实验,分为训练和测试两个部分,训练阶段对训练数据集和验证数据集进行预处理操作,提取相应的特征参数送入基于CNN的双路构音障碍语音识别模型中进行训练和解码,该阶段对训练误差进行梯度下降来学习可训练的权重参数,调整网络模型超参数并对该模型识别结果进行初步预测,由此获得最佳的语音识别模型参数;测试阶段对测试数据集进行预处理操作,提取特征参数后在训练好的语音识别模型上进行语音解码,以此得到测试的语音识别输出结果,验证该模型是否具有泛化性.具体实验流程如图5所示.

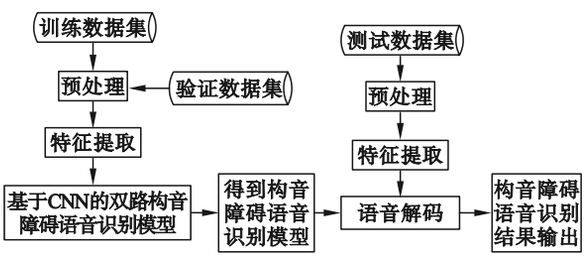

图5 实验流程图

Fig. 5 Experimental flow chart

实验过程中搭建的基于CNN的双路构音障碍语音识别模型见图3,其中单路CNN支路的隐藏层包含9个卷积层、4个池化层,filters与pool_size分别设置为16,32,64,128与2;单路CNN+DSC支路的隐藏层包含6个卷积层、8个深度可分离卷积层、4个池化层,filters与pool_size分别设置为8,16,32,64,128与2;同时采用激活函数为ReLU和softmax函数,优化器为Adam,增添批量归一化层.

本文采用单词识别准确率(word recognition accuracy,WRA)对构音障碍语音识别效果进行评价,可以表示为

$$\mathrm{WRA}=\frac{\text{识别出语音个数}}{\text{需识别语音总个数}}\times 100\%. \quad (15)$$

4.3 实验结果及分析

设计不同声学特征参数分别在不同网络模型上进行训练和构音障碍语音识别实验对比,来验证本文所提的BCFbank特征提取算法以及MBCFbank特征图谱提取算法的有效性,并与MFCC及Fbank特征识别效果进行比较.

不同的声学特征以及不同的网络模型,具体如下:

特征1:120维的Fbank特征;特征2:120维的BCFbank特征;特征3:120维的MBCFbank特征图谱剔除一阶差分部分;特征4:120维的MBCFbank特征图谱;特征5:39维的MFCC特征.

模型1:单路CNN模型;模型2:单路CNN+DSC模型;模型3:基于CNN的双路构音障碍语音识别模型.

不同声学特征在上述三种网络模型上的识别结果对比如表2所示,对于15名构音障碍患者的不同声学特征在模型2与模型3上的语音识别结果对比如表3~表4所示.

表2 5种特征在不同模型上的识别结果

Table 2 Recognition results of five features on different models

特征	WRA/%		
	模型1	模型2	模型3
特征1	83.74	84.45	85.47
特征2	85.94	86.49	87.82
特征3	91.52	92.14	93.48
特征4	92.69	93.24	94.34
特征5	69.91	70.78	71.88

表3 15名患者的4种特征在模型2上的识别结果

Table 3 Recognition results of four features from 15 patients on model 2

构音障碍患者	语音清晰度水平/%	WRA/%			
		特征1	特征2	特征3	特征4
M04	2	71.76	77.65	85.88	87.06
F03	6	79.31	84.48	89.66	91.38
M12	7	74.12	76.47	88.24	89.41
M01	15	70.59	69.12	83.82	85.29
M07	28	88.37	90.70	95.35	93.02
F02	29	84.75	89.83	94.92	94.92
M06	39	86.59	86.59	93.90	95.12
M16	43	87.80	82.93	91.46	93.90
M05	58	90.24	92.68	92.68	95.12
F04	62	85.71	87.30	93.65	95.24
M11	62	86.42	88.89	91.36	92.59
M09	86	84.30	91.74	95.04	95.87
M14	90	92.77	93.98	95.18	96.39
M08	93	89.34	90.98	94.26	95.90
M10	93	93.10	93.97	96.55	97.41

由表2可知,特征4在模型3上的构音障碍语音识别率最高达到了94.34%,相对特征5、特征1、特征2、特征3分别提升了22.46%,8.87%,6.52%,0.86%,特征4在模型3上表现最佳,相比

在模型2、模型1上的构音障碍语音识别率分别提升了1.10%,1.65%.由于Fbank特征是在提取MFCC特征过程中剔除离散余弦变换得到的,特征1相比特征5在模型3、模型2、模型1上的语音识别率分别提升了13.59%,13.67%,13.83%;特征2在模型3、模型2、模型1上的构音障碍语音识别率分别达到了87.82%,86.49%,85.94%,由于Gammatone滤波器组相较于Mel滤波器组更能有效地模拟人耳耳蜗基底膜特性,符合人耳听觉特性,弥补了Mel滤波器组在滤波过程中丢失的有效语音信息,特征2相比特征1在模型3、模型2、模型1上的语音识别率分别提升了2.35%,2.04%,2.20%;由于引入VMD方法对BCFbank特征进行改进,能够有效避免分解过程中出现的模态混叠现象对构音障碍语音信号进行非平稳的有效特性分析,全面表征语音信号包含的信息,特征3相比特征2在模型3、模型2、模型1上的构音障碍语音识别率分别提升了5.66%,5.65%,5.58%;使用一阶差分可以得到构音障碍患者的非平稳语音信号随时间变化的信息和相邻帧与帧之间的信息联系,特征4相比特征3在模型3、模型2、模型1上的构音障碍语音识别率分别提升了0.86%,1.10%,1.17%,由此可见,本文所提特征的识别效果优于MFCC特征及Fbank特征识别效果.

表4　15名患者的4种特征在模型3上的识别结果

Table 4　Recognition results of four features from 15 patients on model 3

构音障碍患者	语音清晰度水平/%	WRA/%			
		特征1	特征2	特征3	特征4
M04	2	72.94	78.82	87.06	88.24
F03	6	81.03	86.20	91.38	93.10
M12	7	75.29	77.65	89.41	90.59
M01	15	70.59	70.59	85.29	85.29
M07	28	90.24	91.86	96.51	94.19
F02	29	86.44	91.52	96.61	96.61
M06	39	87.80	89.02	95.12	96.34
M16	43	89.02	84.15	93.90	95.12
M05	58	91.46	93.90	93.90	95.12
F04	62	87.30	88.89	95.24	96.83
M11	62	87.65	90.12	92.59	93.83
M09	86	84.30	92.56	95.87	96.69
M14	90	92.77	95.18	96.39	97.59
M08	93	89.34	91.80	95.08	96.72
M10	93	93.97	94.83	97.41	98.28

由表3可知,对于每个构音障碍患者,采用特征4在模型2上的构音障碍语音识别效果优于采用特征1、特征2、特征3在该模型上的构音障碍语音识别效果;特征4、特征3、特征2和特征1在模型2上,对于每个构音障碍患者的语音识别率最高分别达到了97.41%,96.55%,93.98%,93.10%;最低分别达到了85.29%,83.82%,69.12%,70.59%.可以看出采用MBCFbank特征图谱的构音障碍语音识别效果是最好的,语音识别率最高达到了97.41%.图6表示不同语音清晰度水平的构音障碍患者在模型2上的语音识别结果对比,可以看出患者语音清晰度水平越高,构音障碍语音识别率也相应越高,在患者语音清晰度高时,特征4、特征3、特征2和特征1在模型2上的构音障碍语音识别率分别达到了96.39%,95.26%,92.67%,89.88%;在患者语音清晰度非常低时,特征4、特征3、特征2和特征1在模型2上的构音障碍语音识别率分别达到了88.29%,86.90%,76.93%,73.95%.

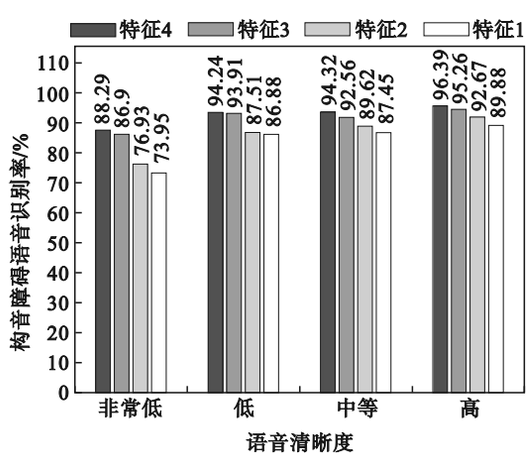

图6　不同语音清晰度的患者在不同特征下的识别结果(模型2)

Fig. 6　Recognition results of patients with different speech intelligibility under different features (model 2)

由表4可知,对于每个构音障碍患者,采用特征4在模型3上的构音障碍语音识别效果优于采用特征3、特征2、特征1在该模型上的构音障碍语音识别效果;特征4、特征3、特征2和特征1在模型3上,对于每个构音障碍患者的语音识别率最高分别达到了98.28%,97.41%,95.18%,93.97%;最低分别达到了85.29%,85.29%,70.59%,70.59%.可以看出采用MBCFbank特征图谱在基于CNN的双路构音障碍语音识别模型上的构音障碍语音识别效果是最好的,语音识别率最高达

到了98.28%.图7表示不同语音清晰度水平的构音障碍患者在模型3上的语音识别结果对比,可以看出在构音障碍患者语音清晰度水平高时,特征4、特征3、特征2和特征1在基于CNN的双路语音识别模型上的构音障碍语音识别率分别达到了97.32%,96.19%,93.59%,90.10%;在患者语音清晰度水平非常低时,特征4、特征3、特征2和特征1在模型3上的构音障碍语音识别率分别达到了89.31%,88.29%,78.32%,73.96%.

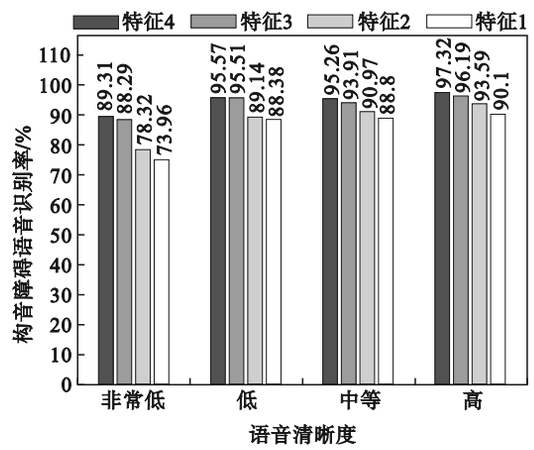

图7 不同语音清晰度的患者在不同特征下的识别结果(模型3)

Fig. 7 Recognition results of patients with different speech intelligibility under different features (model 3)

通过上述分析可以看出,BCFbank特征、MBCFbank特征图谱在单路CNN+DSC模型、基于CNN的双路构音障碍语音识别模型上都显现出很好的网络模型泛化性,BCFbank特征的构音障碍语音识别效果优于传统Fbank特征,该特征弥补了Mel滤波过程中遗漏掉的有效语音特征信息,全面地表征了语音信号包含的信息,在一定程度上提升了构音障碍语音识别率,验证了本文所提BCFbank特征提取算法的有效性.MBCFbank特征图谱的构音障碍语音识别效果优于传统Fbank特征、BCFbank特征和MBCFbank特征图谱剔除一阶差分部分,识别率最高达到了94.34%,该图谱能够有效分解非平稳、非线性的构音障碍语音信号,有效避免分解过程中出现模态混叠现象,能够表达更多丰富的语音信息;同时能够有效地模拟人耳耳蜗基底膜的特性,符合人类的听觉感知特性,弥补了Mel滤波器组在滤波过程中丢失的有效语音信息;该图谱在一定程度上可以反映构音障碍患者的非平稳语音信号随时间变化的信息和相邻帧与帧之间的信息联系,进一步提升了构音障碍语音识别率,验证了本文所提MBCFbank特征图谱提取算法的有效性.

在UA-Speech数据集上,选取文献[15-17]的研究方法与本文算法进行比较,如表5所示.其中,文献[15]提取MFCC特征参数向量,将其映射到固定维度的向量空间上,在转移向量空间(TP-SVM)去构建判别分类器;文献[16]提取MFCC特征作为神经网络ANN和MLP的输入,进行模型训练;文献[17]对语音视觉表征更感兴趣,结合CNN提出语音视觉系统(speech vision,SV),将单词语音转换为视觉特征,在S-CNN架构上进行模型训练.观察表5可知,特征2+模型3的语音识别率比ANN+MLP[15]提升了18.94%;特征4+模型3的语音识别率比ANN+MLP[15],LL-SVM[16]和SV+S-CNN[17]分别提升了25.46%,6.43%,4.80%.通过上述语音识别效果比较可知,MBCFbank特征图谱在基于CNN的双路构音障碍语音识别模型上的识别效果表现更佳,能够进一步提升语音识别率.

表5 本文算法与其他主流方法对比

Table 5 Comparison of the method with other mainstream methods

方法	特征参数	WRA/%
ANN+MLP[15]	MFCC	68.88
LL-SVM[16]	MFCC	87.91
SV+S-CNN[17]	频谱图	89.54
特征2+模型3	BCFbank特征	87.82
特征4+模型3	MBCFbank特征图谱	94.34

5 结 语

为了更好地提取到符合人耳耳蜗基底膜结构特性的声学特征参数,提出一种BCFbank特征提取算法,结合了Mel滤波器组和Gammatone滤波器组两者的优势;由于VMD方法能够对构音障碍语音信号进行有效分解,对BCFbank特征提取算法进行优化,提出一种MBCFbank特征图谱提取算法,结合了VMD和BCFbank特征的优势,能够有效地分析构音障碍语音信号,避免出现模态混叠现象.本文使用5种不同声学特征在不同网络模型上进行构音障碍语音识别的实验对比与结果分析,得到最佳的声学特征参数,能够进一步提升构音障碍语音识别率,验证了本文所提特征提取算法的有效性.

参考文献:

[1] Mohammed S Y, Sid-Ahmed S, Brahim-Fares Z, et al. Improving dysarthric speech recognition using empirical mode decomposition and convolutional neural network[J]. *EURASIP Journal on Audio, Speech, and Music Processing*, 2020, 2020(1): 1-7.

[2] Al-Qatab B A, Mustafa M B. Classification of dysarthric speech according to the severity of impairment: an analysis of acoustic features[J]. *IEEE Access*, 2021(9): 18183-18194.

[3] Liu S, Hu S, Xie X, et al. Recent progress in the CUHK dysarthric speech recognition system [J]. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021, 29(99): 2267-2281.

[4] Yue Z, Loweimi E, Christensen H, et al. Acoustic modelling from raw source and filter components for dysarthric speech recognition[J]. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2022(30): 2968-2980.

[5] 梁正友,黎雨星,孙宇,等.基于多特征组合的构音障碍语音识别[J].计算机工程与设计,2022,43(2):567-572.
(Liang Zheng-you, Li Yu-xing, Sun Yu, et al. Speech recognition with dysarthria based on multi-feature combination[J]. *Computer Engineering and Design*, 2022, 43(2): 567-572.)

[6] Jiao Y, Tu M, Berisha V, et al. Simulating dysarthric speech for training data augmentation in clinical speech applications [C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary Allcerta: IEEE, 2018: 6009-6013.

[7] Yilmaz E, Mitra V, Sivaraman G, et al. Articulatory and bottleneck features for speaker-independent ASR of dysarthric speech[J]. *Computer Speech & Language*, 2019, 58: 319-334.

[8] Zaidi B F, Selouani S A, Boudraa M, et al. Deep neural network architectures for dysarthric speech analysis and recognition[J]. *Neural Computing and Applications*, 2021, 33(15): 9089-9108.

[9] Mariya T A, Vijayalakshmi P, Nagarajan T. Data augmentation techniques for transfer learning-based continuous dysarthric speech recognition [J]. *Circuits, Systems, and Signal Processing*, 2023, 42(1): 601-622.

[10] 李东,张雪英,段淑斐,等.结合语音融合特征和随机森林的构音障碍识别[J].西安电子科技大学学报,2018,45(3):149-155.
(Li Dong, Zhang Xue-ying, Duan Shu-fei, et al. Articulation disorder recognition based on speech fusion features and random forest[J]. *Journal of Xidian University*, 2018, 45(3): 149-155.)

[11] 吴丽丹.基于深度时序网络的多视图构音障碍语音识别[D].上海:华东师范大学,2021.
(Wu Li-dan. Multi-view articulation disorder speech recognition based on deep temporal network[D]. Shanghai: East China Normal University, 2021.)

[12] 王赵国,韦存海,彭雅妮,等.基于GFCC-SVM-RFE的电力设备声音特征提取方法[J].电力信息与通信技术,2022,20(9):34-42.
(Wang Zhao-guo, Wei Cun-hai, Peng Ya-ni, et al. Sound feature extraction method of Power Equipment based on GFCC-SVM-RFE [J]. *Electric Power Information and Communication Technology*, 2022, 20(9): 34-42.)

[13] Dragomiretskiy K, Zosso D. Variational mode decomposition [J]. *IEEE Transactions on Signal Processing*, 2014, 62(3): 531-544.

[14] Fritsch J, Magimai-Doss M. Utterance verification-based dysarthric speech intelligibility assessment using phonetic posterior features [J]. *IEEE Signal Processing Letters*, 2021(28): 224-228.

[15] Shahamiri S R, Salim S. Artificial neural networks as speech recognisers for dysarthric speech: identifying the best-performing set of MFCC parameters and studying a speaker-independent approach [J]. *Advanced Engineering Informatics*, 2014, 28(1): 102-110.

[16] Rajeswari N, Chandrakala S. Generative model-driven feature learning for dysarthric speech recognition [J]. *Biocybernetics & Biomedical Engineering*, 2016, 36(4): 553-561.

[17] Shahamiri S R. Speech vision: an end-to-end deep learning-based dysarthric automatic speech recognition system [J]. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2021(29): 852-861.