计算机科学
COMPUTER SCIENCE

基于改进MFCC和能量算子倒谱的语种识别

陈思竹, 龙华, 邵玉斌

引用本文

陈思竹, 龙华, 邵玉斌. 基于改进MFCC和能量算子倒谱的语种识别[J]. 计算机科学, 2024, 51(11A): 231000065-6.

CHEN Sizhu, LONG Hua, SHAO Yubin. Language Recognition Based on Improved MFCC and Energy Operator Cepstrum [J]. Computer Science, 2024, 51(11A): 231000065-6.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.231000065

基于改进 MFCC 和能量算子倒谱的语种识别

陈思竹[1,2]　龙　华[1]　邵玉斌[1]
1 昆明理工大学信息工程与自动化学院　昆明 650500
2 云南省无线电监测中心　昆明 650228
(1652720478@qq.com)

摘　要　针对广播语音信号低信噪比下语种识别准确率低和鲁棒性差的问题,提出了基于小波包变换改进 MFCC 和能量算子倒谱特征的语种识别算法。首先,采用小波包变换代替 MFCC 中的傅里叶变换和 Mel 滤波得到 WMFCC 特征参数。在保留人耳听觉感知特性的基础上提升语音信号的高频分析能力和分析精确度,克服傅里叶变换的局限性。其次,提取 Teager 能量算子倒谱,得到语音瞬时能量的特性,与改进的 MFCC 特征参数融合得到新的特征参数 TWMFCC。最后,为进一步提升低信噪比语音的识别效果,提出了 VMD 自适应维纳滤波去噪算法。通过实验对比了所提特征与传统特征的识别效果,所提特征的平均识别准确率显著提升,带噪语音在未进行语音去噪处理的情况下较传统 MFCC 高 13.02%,有效改善了传统特征在低信噪比下识别准确率低的问题,具有较强的抗噪性和鲁棒性。

关键词:语种识别;MFCC;小波包变换;能量算子倒谱;GMM-UBM
中图分类号　TN912.34

Language Recognition Based on Improved MFCC and Energy Operator Cepstrum

CHEN Sizhu[1,2],LONG Hua[1] and SHAO Yubin[1]
1 Faculty of Information Engineering and Automation,Kunming University of Science and Technology,Kunming 650500,China
2 Radio Monitoring Center of Yunnan Province,Kunming 650228,China

Abstract　Aiming at the problem of low accuracy and poor robustness of language recognition under low signal-to-noise ratio of broadcast speech signals,a language recognition algorithm based on wavelet packet transform to improve MFCC and energy operator cepstrum features is proposed. Firstly,the WMFCC feature parameters are obtained by using wavelet packet transform instead of Fourier transform and Mel filter in MFCC. On the basis of retaining the auditory perception characteristics of the human ear,the high-frequency analysis ability and analysis accuracy of the speech signal are improved,and the limitations of the Fourier transform are overcomed. Secondly,the Teager energy operator cepstrum is extracted to obtain the characteristics of the instantaneous energy of the speech,which is fused with the improved MFCC feature parameters to obtain a new feature parameter TWMFCC. Finally,in order to further improve the recognition effect of low SNR speech,a VMD adaptive Wiener filtering denoising algorithm is proposed. The experiment compares the recognition effect of the proposed features with the traditional features. The average recognition accuracy of the proposed features is significantly improved,which is 13.02 % higher than that of the traditional MFCC without speech denoising. It effectively alleviates the problem of low recognition accuracy of traditional features under low signal-to-noise ratio,and has strong anti-noise and robustness.

Keywords　Language recognition,MFCC,Wavelet packet transform,Energy operator cepstrum,GMM-UBM

1　引言

随着国际化、全球化趋势的不断加强和现代科技的快速发展,多语言交流需求不断增加。云南省作为我国西南边陲,与缅甸、老挝、越南相邻,随着面向南亚东南亚辐射中心的建设,近年来云南省边境地区无线电业务发展十分迅速,电磁环境日趋复杂。如何通过技术手段快速准确掌握广播信号发射语言种类成为无线电监管部门的一项工作内容。目前,针对云南边境小语种的识别研究较少,本文将建立基于云南边境语种的广播语音数据集,开展云南边境广播语种识别研究。

语种识别系统的性能取决于特征参数的类型和所使用的分类判别器。如果特征参数不能很好地表征语音中的信息,无论使用何种分类判别器,系统的性能都将很差。特征的选择和提取对于语种识别系统获得良好的识别性能非常重要。在以往的语种识别研究中,基于底层声学特征的语种识别方法一直是语种识别领域的研究热点[1-5]。语种识别研究常用的声学特征有梅尔频率倒谱系数(Mel Frequency Cepstrum Coefficients,MFCC)[6]、感知线性预测系数(Perceptual Linear Predictive Coefficients,PLP)[7]、线性预测倒谱系数(Linear Predictive Cepstral Coefficient,LPCC)[8]、移位差分倒谱

基金项目:云南省媒体融合重点实验室开放基金(320225403)
This work was supported by the Yunnan Key Laboratory of Media Convergence Open Fund(320225403).
通信作者:龙华(1670931890@qq.com)

(Shifted Delta Cepstrum, SDC)[9]等。MFCC梅尔频率倒谱系数是说话人识别、语种识别和语音识别中最为常用的特征[10-11],以倒谱系数的形式表示声音信号。研究者进行了大量基于MFCC特征的改进和融合来提高识别率的研究。Manchala等[12]使用MFCC和从语音信号的线性预测(Linear Predictive, LP)分析中提取的共振峰频率进行识别。Mukherjee等[13]提出了一种新的基于第二级梅尔频率倒谱系数(MFCC-2)的特征,以解决MFCC维数大且不均匀的问题。Sangwan等[14]提取MFCC和RASTA-PLP(Relative Spectral Transform-Perceptual Linear Prediction, RASTA-PLP)融合特征输入神经网络进行识别。Liu等[15]提出了一种基于CNN-BLSTM和时间池化单元的MFCC提取方法,通过学习网络隐藏状态与时间序列之间的关系,得到语言的话语级表示,增加特征的时序信息。除此之外,研究者们还提出了基于伽玛通频率倒谱系数(Gammatone Frequency Cepstrum Coefficient, GFCC)、耳蜗滤波倒谱系数(Cochlear Filter Cepstral Coefficients, CFCC)及其与能量算子倒谱特征融合的语种识别方法[16-17],均取得了提升语种识别率的效果。针对广播音频,传统特征提取方法在低信噪比环境下的识别效果并不理想。

本文提出了基于传统MFCC特征的改进算法。针对传统MFCC在低信噪比下语种识别率表现不佳的问题,研究传统MFCC的局限性,采用小波包变换代替MFCC中的傅里叶变换和Mel滤波得到WMFCC特征参数,提高语音信号的高频分析能力和分析精确度,并提取Teager能量算子倒谱,获取语音信号的能量信息,与改进的MFCC特征参数融合得到新的特征参数TWMFCC,并使用VMD自适应维纳滤波去噪算法进一步提升低噪语音识别效果。

2 融合特征提取

2.1 小波包变换

在傅里叶变换中,三角函数是一个无限长的信号,在时域上没有局部性,因此在对信号进行傅立叶变换后,只能获得信号的频域信息,故傅里叶变换对非平稳信号的处理能力有限。小波变换由一系列正交、迅速衰减、有限长的小波函数基进行拟合。小波函数基可通过其平移和伸缩变化,获得不同的频率和时间位置。小波函数的一般形式为:

$$\psi_{a,b}(t)=\frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right),a,b\in\mathbb{R} \tag{1}$$

小波包变换在小波变换的基础上引入了更加精细的分解过程。小波包通常由两组相互正交的小波基滤波器系数生成。设U_n代表小波包变换的小波包集,其中$n\in Z$,小波包变换函数表示为:

$$\begin{cases}U_{2n}(t)=\sqrt{2}\sum_k h(k)U_n(2t-k)\\U_{2n+1}(t)=\sqrt{2}\sum_k g(k)U_n(2t-k)\end{cases} \tag{2}$$

根据式(2)可知,小波包变换在小波变换的基础上,在每一级信号分解时,对低频子带和高频子带同时进行进一步分解,从而提供更多的时频局部信息,还能通过最小化代价函数自适应选择最优分解路径。通过这种方式,可以根据信号的特点和变化,自适应地选择最合适的频段进行分解,从而获得更好的信号分析结果。

2.2 改进MFCC提取

提取MFCC的过程是对语音信号做快速傅里叶变换得到频域信号,再通过一组Mel滤波器进行滤波,用小波包变换来取代这两个步骤,构造新的特征参数WMFCC(Wavelet packet transform Mel-scale Frequency Cepstral Coefficients),提取过程如图1所示,小波包分解算法如下:

$$c(a,b)=\int_{-\infty}^{\infty}x(t)\psi_{(i,j)}(t)\mathrm{d}t \tag{3}$$

其中,$x(t)$是原始信号,$\psi_{(i,j)}(t)$是小波包函数,i代表小波包分解后所在层数,j代表小波包分解节点在该层的位置。小波包分解后每一个节点都包含大量精确数据。

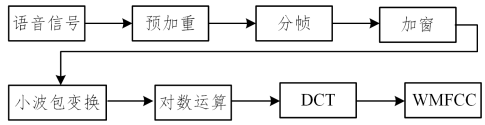

图1 WMFCC提取流程图

Fig. 1 WMFCC extraction process

1)小波包变换:对每帧信号进行小波包变换,得到小波包信号$E_{l,r}$,每个子带的能量P_l的计算式为:

$$P_l=\frac{\sum_{r=1}^{N_l}E_{l,r}}{N_l},l=1,2,\cdots,L \tag{4}$$

其中,l和r对应第l个子带中第r个小波包系数;N_l为第l个子带内小波包系数的个数;L为子带总数。

2)取对数:对每个子带的能量P_l取对数,提高信号的鲁棒性。

$$Q_l=\log P_l,l=1,2,\cdots,L \tag{5}$$

3)DCT变换:对Q_l进行离散余弦变换,映射到低维空间:

$$C_{\mathrm{WMFCC}}(i)=\sqrt{\frac{2}{N}}\sum_{j=1}^{L}Q_L\cos\left(\frac{\pi i(j-0.5)}{L}\right),i=1,2,\cdots,D \tag{6}$$

由此得到新的特征参数WMFCC,其中D为WMFCC的维数,本文设置为12。

本文实验中的语音信号采样频率为16 000 Hz,即根据奈奎斯特定理,最高频率为8 000 Hz,这里小波包分解的频率范围为0～8 000 Hz,采用db6小波进行6层分解,共127个节点。梅尔滤波器组通常由24个滤波器构成,用小波包变换代替傅里叶变换和梅尔滤波器组,对分解的子带系数进行不同层次的取舍,选取24个小波包分解节点构建小波包树,如图2所示。

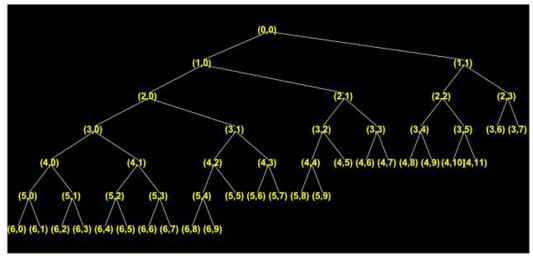

图2 小波包分解树

Fig. 2 Tree of wavelet packet factorization

小波包分解节点频率划分如表1所列,对于能量集中的低频段,取所有的最低层次系数,每段子带带宽为125 Hz,次低频段取次低层的小波系数,每段子带带宽为250 Hz,剩余依

次划分不同层子带频率范围。在子带划分上与 Mel 尺度的频率响应相匹配,24 波段小波包子带类似于 24 波段 Mel 滤波器,在能量密集的低频段分配了更多的子带[18]。

表 1 小波包分解节点频率划分

Table 1 Wavelet packet decomposition node frequency division

序号	节点	频带范围/Hz	中心频率/Hz	频带带宽/Hz
1	[6,0]	0～125	62.5	125
2	[6,1]	125～250	187.5	125
3	[6,2]	250～375	312.5	125
4	[6,3]	375～500	437.5	125
5	[6,4]	500～625	562.5	125
6	[6,5]	625～750	687.5	125
7	[6,6]	750～875	812.5	125
8	[6,7]	875～1000	937.5	125
9	[6,8]	1000～1125	1062.5	125
10	[6,9]	1125～1250	1187.5	125
11	[5,5]	1250～1375	1312.5	125
12	[5,6]	1375～1500	1437.5	125
13	[5,7]	1500～1750	1625	250
14	[5,8]	1750～2000	1875	250
15	[5,9]	2000～2250	2125	250
16	[4,5]	2250～2500	2375	250
17	[4,6]	2500～2750	2625	250
18	[4,7]	2750～3000	2875	250
19	[4,8]	3000～3500	3250	500
20	[4,9]	3500～4000	3750	500
21	[4,10]	4000～5000	4500	1000
22	[4,11]	5000～6000	5500	1000
23	[3,6]	6000～7000	6500	1000
24	[3,7]	7000～8000	7500	1000

2.3 Teager 能量算子倒谱

Teager 能量算子是 Teager 提出的一种非线性信号算子,Kaiser 对其进行了形式化定义和性质研究[19],该算子可以有效地衡量信号的瞬时能量。

对于输入信号 $x(n)$,Teager 能量算子定义为:

$$\Psi[x(n)]=x(n)^2-x(n+1)x(n-1) \tag{7}$$

本文实验在白噪声环境下进行,带噪语音信号 $x(n)$ 可以表示为原始语音信号 $s(n)$ 与零均值噪声信号 $\omega(n)$ 之和,即:

$$x(n)=s(n)+\omega(n) \tag{8}$$

$x(n)$ 的 Teager 能量算子可以表示为:

$$\Psi[x(n)]=\Psi[s(n)]+\Psi[\omega(n)]+2\widetilde{\Psi}[s(n),\omega(n)] \tag{9}$$

其中,$\widetilde{\Psi}[s(n),\omega(n)]$ 是 $s(n)$ 与 $x(n)$ 的互 Teager 能量,且有:

$$\widetilde{\Psi}[s(n),\omega(n)]=s(n)\omega(n)-0.5s(n-1)\omega(n+1)-0.5s(n+1)\omega(n-1) \tag{10}$$

由于 $s(n)$ 和 $\omega(n)$ 均为零且两者相互独立,因此有:

$$E\{\widetilde{\Psi}[s(n),\omega(n)]\}=0 \tag{11}$$

推导出:

$$E\{\Psi[x(n)]\}=E\{\Psi[s(n)]\}+E\{\Psi[\omega(n)]\} \tag{12}$$

一般情况下,与原始语音信号的 Teager 能量相比,噪声的 Teager 能量可以忽略不计,因此可以得到:

$$E\{\Psi[x(n)]\}\approx E\{\Psi[s(n)]\} \tag{13}$$

由此可见,Teager 能量算子可以消除零均值噪声的影响,达到语音增强的目的。

经过预处理的语音信号根据式(7)求出每帧语音信号的平均 Teager 能量,进行归一化处理并取对数得到:

$$\hat{\Psi}[x(n)]=\log\{\Psi[x(n)]/\max(\Psi[x(n)])\} \tag{14}$$

然后进行 DCT 变换再求均值得到一维 Teager 能量算子倒谱参数 TEOCC(Teager Energy Operator Cepstrum Coefficient,TEOCC)。

2.4 特征融合

为了构造更有效的语种识别特征,本文将改进的 MFCC 特征和非线性能量特征进行融合,在提取基于小波包变换改进的 MFCC 特征 WMFCC 的基础上,加入反映信号能量变化的 Teager 能量算子倒谱参数 TEOCC,得到融合特征 TWMFCC。该特征在保留了人耳听觉感知特性的基础上提升了语音信号的高频分析能力和分析精确度,又结合了语音瞬时能量的特性,还能在一定程度上降噪,更完整更清晰地描述语音的特性。

3 基于 VMD 的自适应维纳滤波去噪

自适应维纳滤波算法在进行降噪处理时,在滤波过程中利用信号 $x(n)$ 的局部均值 m_x 和局部方差 σ_x^2 调整滤波器的输出,最小化去噪信号和原始信号之间的均方误差。它可以根据信号的局部统计特性进行自适应滤波,与线性滤波器相比,可以保留更多原始信号的非平稳特征。自适应维纳滤波的具体算法如下。

设带噪语音为:

$$x(n)=s(n)+v(n) \tag{15}$$

其中,$s(n)$ 为原始语音信号,$v(n)$ 为噪声。

本文实验在白噪声环境下进行研究,因此加性噪声 $v(n)$ 的均值为零,方差为 σ_v^2,因此功率谱可近似为:

$$P_v(\omega)=\sigma_v^2 \tag{16}$$

考虑一小段语音信号,其中假定信号 $x(n)$ 是平稳的,信号 $x(n)$ 可以表示为:

$$x(n)=m_x+\sigma_x w(n) \tag{17}$$

其中,m_x 和 σ_x 是 $x(n)$ 的局部平均值和标准差;$w(n)$ 是均值为零的单位方差噪声。因此,原始信号 $s(n)$ 的平均值 m_s 等于 m_x。

在这一小段语音中,维纳滤波器的传递函数可以近似为:

$$H_w(\omega)=\frac{P_s(\omega)}{P_s(\omega)+P_v(\omega)}=\frac{\sigma_s^2}{\sigma_s^2+\sigma_v^2} \tag{18}$$

由式(18)可以得到维纳滤波器的脉冲响应。

$$h(n)=\frac{\sigma_s^2}{\sigma_s^2+\sigma_v^2}\delta(n) \tag{19}$$

由式(20),该段得到增强的语音信号 $\hat{s}(n)$ 可以表示为:

$$\hat{s}(n)=m_s+(x(n)-m_s)\frac{\sigma_s^2}{\sigma_s^2+\sigma_v^2}\delta(n)=m_s+\frac{\sigma_s^2}{\sigma_s^2+\sigma_v^2}(x(n)-m_s) \tag{20}$$

如果假设 m_s 和 σ_s 在每个语音小段上都是变化的,那么:

$$\hat{s}(n)=m_s(n)+\frac{\sigma_s^2(n)}{\sigma_s^2(n)+\sigma_v^2}(x(n)-m_s(n)) \tag{21}$$

我们可以通过式(21),由 $x(n)$ 估算 $m_s(n)$,表达式如下:

$$\hat{m}_s(n)=\frac{1}{(2A+1)}\sum_{k=n-A}^{n+A}x(k) \tag{22}$$

其中,$(2A+1)$ 是估算中使用的短语音段数。

为计算语音系统的原始信号统计量,有 $\sigma_x^2=\sigma_s^2+\sigma_v^2$,因此有:

$$\hat{\sigma}_s^2(n)=\begin{cases}\hat{\sigma}_x^2(n)-\hat{\sigma}_v^2, & \text{if } \hat{\sigma}_x^2(n)>\hat{\sigma}_v^2\\ 0, & \text{else}\end{cases} \tag{23}$$

其中，

$$\hat{\sigma}_x^2(n)=\frac{1}{(2A+1)}\sum_{k=n-A}^{n+A}(x(k)-\hat{m}_x(n))^2 \tag{24}$$

通过该方法，得到了传递函数基于语音信号局部统计量变化的自适应滤波器。研究显示[20]，该方法与传统维纳滤波、谱减法和小波去噪法相比，在加性高斯白噪声和有色噪声情况下均取得了很好的效果。

为了提高非平稳信号的滤波效果，本文采用基于变分模态分解(Variational Mode Decomposition，VMD)的自适应滤波算法，首先利用 VMD 分解带噪信号，得到一系列不同中心频率模态分量的信号，再利用自适应滤波对各模态分量去噪，得到滤波后的模态分量，最后对滤波后的模态分量进行重构，得到去噪后的语音信号。

4 GMM-UBM 模型

高斯混合模型(Gaussian Mixture Model，GMM)[21]是一种多维的概率密度函数，由多个单高斯密度分布(Gaussian)进行线性叠加组成，每个单高斯密度分布称为一个成员函数。高斯混合模型可以拟合任何形状的数据分布。语种识别的声学特征复杂，无法用简单聚类方法进行拟合，高斯混合模型能很好地表示语种分布特性。然而，为每种语言单独建立能准确描述该语言信息的 GMM 模型需要足够的语音数据量。基于此提出了通用背景模型(Universal Background Model，UBM)，UBM 模型本质上是一个大型 GMM 模型，在基于 GMM 的语种识别系统中，每一种语言都有自己的 GMM，而基于通用背景模型的 GMM 系统对所有语言使用单一背景模型，这种 GMM 被称为 UBM，从 UBM 中自适应得到每种语言的单独 GMM。训练 UBM 模型的过程如下：首先通过 k-mean 算法对模型参数进行初始化，再通过 EM 算法迭代更新，得到 UBM 模型后，各个目标语种模型是在 UBM 模型的基础上利用该语种的数据进行自适应。

GMM-UBM 模型是语音处理研究中常用的模式识别方法，尤其是在文本无关的说话人识别研究中[22]，后引入语种识别研究。GMM-UBM 模型的具体流程如图 3 所示。

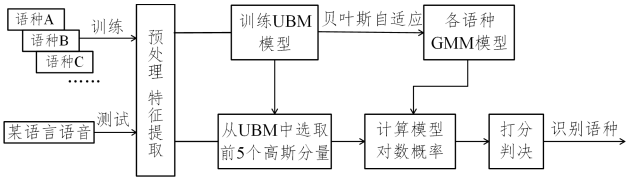

图 3 GMM-UBM 模型

Fig. 3 GMM-UBM model

5 实验设计及结果分析

5.1 数据集及性能评价指标

本文的研究任务来源于实际需求，语料库由真实广播音频构成，分别包含缅甸语、越南语、老挝语、柬埔寨、英语，基本囊括了云南边境地区所有可能接收到的越界信号语种。每个语种语料均采集自多个频道、不同时间段、不同说话人的广播音频。通过人工对广播音频语料进行筛选，去除静音和音乐片段、去噪、分割，得到纯净语音语料。每种语言各 2000 条，统一处理为采样率 16 kHz、精度 16 bit、时长 3 s 的单声道 wav 音频。同时，为了研究噪声对语种识别的影响，提升带噪语音的识别效果，以 NoiseX-92 公共噪声库中的白噪声为噪声源，构建了信噪比分别为 15 dB，10 dB，5 dB，0 dB，−5 dB 的带噪语音语料。其中，每种语言每种信噪比各有 1400 条训练集和 600 条测试集语音文件。训练集和测试集的构成如表 2 和表 3 所列。

表 2 训练集

Table 2 Training set

语种	训练集/条					
	原语音	15 dB	10 dB	5 dB	0 dB	−5 dB
老挝	1400	1400	1400	1400	1400	1400
缅甸	1400	1400	1400	1400	1400	1400
越南	1400	1400	1400	1400	1400	1400
柬埔寨	1400	1400	1400	1400	1400	1400
英语	1400	1400	1400	1400	1400	1400

表 3 测试集

Table 3 Test set

语种	训练集/条					
	原语音	15 dB	10 dB	5 dB	0 dB	−5 dB
老挝	600	600	600	600	600	600
缅甸	600	600	600	600	600	600
越南	600	600	600	600	600	600
柬埔寨	600	600	600	600	600	600
英语	600	600	600	600	600	600

本文采用 NIST 评测标准中的多语种识别正确率作为评价指标，其表达式为：

$$\overline{Acc}=\frac{\sum_{i=1}^{G}R}{N},i=1,2,\cdots,G,G=5 \tag{25}$$

其中，R 为每种语种识别正确数，N 为测试集总数。

5.2 实验设计及结果分析

本文实验采用 MATLAB R2020b 作为测试平台，验证本文提出的融合特征和去噪算法的有效性。

实验 1 验证无噪声环境下所提算法的有效性。由纯净语音分别提取 MFCC、MFCC 加上一阶差分和二阶差分(MFCC-Delta-Acceleration，MFCC-D-A)、WMFCC、TWMFCC 这 4 种特征，输入 GMM-UBM 模型进行训练，对比测试结果。原语音语种识别准确率如表 4 所列。

表 4 原语音语种识别准确率

Table 4 Accuracy of original speech language recognition

(%)

特征参数	语种					平均识别准确率
	老挝	缅甸	越南	柬埔寨	英语	
MFCC	72.11	72.52	69.16	73.18	86.28	74.65
MFCC-D-A	73.58	78.54	74.61	70.73	84.13	76.32
WMFCC	75.64	80.67	75.35	78.75	88.50	79.78
TWMFCC	78.86	82.33	78.00	80.33	90.83	82.07

由表 4 可见，本文提取的 TWMFCC 特征平均识别率最高，相比原 MFCC 特征提高了 7.42%；基于小波包变换的 MFCC 特征 WMFCC 在原 MFCC 的基础上提高了 5.13%；WMFCC 与 TEOCC 融合得到的 TWMFCC 比单一 WMFCC 提高了 2.29%。小波包变换对语音的处理更加细腻，更大程度地保留了语音的特性，MFCC 引入小波包变换得到的 WMFCC 特征对每种语言的识别率均较 MFCC 有很大提升。Teager 能量算子反映了语音信号的能量变换，对语种识别也

有提升效果。

该实验结果证明，对于纯净语音，本文提出的 TWMFCC 特征较传统 MFCC 特征能有效提升语种识别准确率。

实验 2　验证白噪声环境下所提算法的有效性。将不同信噪比的带噪语音，直接进行特征提取，得到 MFCC，MFCC-D-A，WMFCC，TWMFCC 4 种特征，输入 GMM-UBM 模型进行训练识别，对比测试结果。然后使用 VMD 自适应维纳滤波降噪后，再进行特征提取，同样提取上述 4 种特征，对比测试结果。带噪语音语种识别准确率如表 5 所列。

表 5　带噪语音语种识别准确率

Table 5　Accuracy of noisy speech language recognition

（%）

特征参数	降噪处理	信噪比/dB					平均识别准确率
		15	10	5	0	−5	
MFCC	否	68.33	50.60	35.78	23.86	20.67	39.85
	是	72.75	68.33	63.15	58.74	48.89	62.37
MFCC-D-A	否	71.50	53.52	38.10	26.53	22.46	42.42
	是	75.53	69.85	65.26	60.50	51.80	64.59
WMFCC	否	75.36	58.80	50.36	41.60	30.56	51.34
	是	78.23	72.82	68.50	65.36	57.16	68.41
TWMFCC	否	77.20	60.66	52.15	43.02	31.30	52.87
	是	79.36	75.13	70.26	68.16	60.05	70.59

由表 5 可见，对于带噪语音，本文提取的 TWMFCC 特征参数在语种识别中仍然具有较高的识别率。对比不同特征参数的语种识别率，未进行语音去噪处理的情况下，TWMFCC 特征参数的平均识别率较传统 MFCC 高 13.02%，WMFCC 特征参数的平均识别率较传统 MFCC 高 11.49%，具有较好的抗噪性能；在进行语音去噪处理的情况下，TWMFCC 特征参数的平均识别率较传统 MFCC 高 8.22%，WMFCC 特征参数的平均识别率较传统 MFCC 高 6.04%。对比同一特征参数不同信噪比下的识别准确率，当信噪比低于 5dB 时，语音信号信息损失严重，语种识别率随信噪比降低而大幅降低。对比 VMD 自适应维纳滤波降噪处理前后的语种识别率，每种特征参数均在进行降噪处理后识别准确率得到很大提升，VMD 自适应维纳滤波有效去除了语音中的大量噪声，很大程度上还原了语音的原始信息，提升了语种识别效果。

该实验结果证明，对于带噪语音，本文提出的 TWMFCC 特征较传统 MFCC 特征能有效提升语种识别准确率，具有更佳的抗噪性能；本文提出的 VMD 自适应维纳滤波降噪法，能有效去除噪音，还原语音信息，提升识别效果。

结束语　针对广播语音信号低信噪比下语种识别准确率低和鲁棒性差的问题，本文提出了基于小波包变换改进 MFCC 和能量算子倒谱特征的语种识别算法。该特征在保留了人耳听觉感知特性的基础上提升了语音信号的高频分析能力和分析精确度，又结合语音瞬时能量的特性，还具有一定降噪性能。为进一步提升低信噪比语音的识别效果，提出了 VMD 自适应维纳滤波去噪算法。实验对比了传统特征的识别效果，所提算法有效改善了传统特征在低信噪比下识别准确率低的问题，且具有较强的抗噪性和鲁棒性。由于本文实验仅针对白噪声语音进行，且小波包变换增大了特征提取过程的算力开销和时间开销，因此在未来研究中，可以加入其他复杂噪声进行研究，并进一步优化算法，以提升语种识别系统的实用性。

参 考 文 献

[1] LI H，MA B，LEE K A. Spoken Language Recognition：From Fundamentals to Practice[J]. Proceedings of the IEEE，2013，101(5)：1136-1159.

[2] DESHWAL D，SANGWAN P，KUMAR D. Feature Extraction Methods in Language Identification：A Survey[J]. Wireless Personal Communications，2019，107(4)：2071-2103.

[3] SRINIVAS N S S，SUGAN N，KAR N，et al. Recognition of Spoken Languages from Acoustic Speech Signals Using Fourier Parameters[J]. Circuits，Systems，and Signal Processing，2019，38(11)：5018-5067.

[4] TAWAQAL B，SUYANTO S. Recognizing Five Major Dialects in Indonesia Based on MFCC and DRNN[J]. Journal of Physics：Conference Series，2021，1844(1)：012003.

[5] GUPTA J，PATHAK S，KUMAR G. Deep Learning(CNN) and Transfer Learning：A Review[J]. Journal of Physics：Conference Series，2022，2273(1)：012029.

[6] BISWAS M，RAHAMAN S，AHMADIAN A，et al. Automatic spoken language identification using MFCC based time series features[J]. MultimediaTools and Applications，2023，82(7)：9565-9595.

[7] DEEPTI D，PARDEEP S，DIVYA K. A Language Identification System using Hybrid Features and Back-Propagation Neural Network[J]. Applied Acoustics，2020，164：107289.

[8] ZHU J，LIU Z. Analysis of Hybrid Feature Research Based on Extraction LPCC and MFCC[C]// 2014 Tenth International Conference on Computational Intelligence and Security. 2014：732-735.

[9] TZUDIR M，BAGHEL S，SARMAH P，et al. Analyzing RMFCC Feature for Dialect Identification inAo，an Under-Resourced Language[C]// 2022 National Conference on Communications (NCC). 2022：308-313.

[10] SUYANTO S，ARIFIANTO A，SIRWAN A，et al. End-to-End Speech Recognition Models for a Low-Resourced Indonesian Language[C]// 2020 8thInternational Conference on Information and Communication Technology(ICoICT). Yogyakarta，Indonesia：IEEE，2020：1-6.

[11] ALKHATIB B，KAMAL EDDIN M M W. Voice Identification Using MFCC and Vector Quantization[J]. Baghdad Science Journal，2020，17(3(Suppl.))：1019.

[12] MANCHALA S，KAMAKSHI PRASAD V，JANAKI V. GMM based language identification system using robust features[J]. International Journal of Speech Technology，2014，17(2)：99-105.

[13] MUKHERJEE H，OBAIDULLAH S M，SANTOSH K C，et al. A lazy learning−based language identification from speech using MFCC-2 features[J]. International Journal of Machine Learning and Cybernetics，2020，11(1)：1-14.

[14] SANGWAN P，DESHWAL D，DAHIYA N. Performance of a language identification system using hybrid features and ANN learning algorithms[J]. Applied Acoustics，2021，175：107815.

[15] LIU X，CHEN C，HE Y. Temporal feature extraction based on CNN-BLSTM and temporal pooling for language identification [J]. Applied Acoustics，2022，195：108854.

[16] LIU J，SHAO Y，LONG H，et al. Language identification based on GFCC and energy operator cepstrum[J]. Journal of Yunnan

University(Natural Science Edition). 2022,44(2):254-261.
[17] SHI Y,BAI J. Speech recognition combining CFCC and Teager energy operator cepstral coefficients [J]. Computer Science, 2019,46(5):286-289.
[18] FAROOQ O,DATTA S. Mel filter-like admissible wavelet packet structure for speech recognition[J]. IEEE Signal Processing Letters,2001,8(7):196-198.
[19] PRÉAUX Y,BOUDRAA A O,LARKIN K G. On the positivity of Teager-Kaiser's energy operator[J]. Signal Processing,2022, 201:108702.
[20] ABD EL-FATTAH M A,DESSOUKY M I,ABBAS A M,et al. Speech enhancement with an adaptive Wiener filter[J]. International Journal of Speech Technology,2014,17(1):53-64.
[21] DOUGLAS A R,RICHARD C R. Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models [J]. IEEE Transactions on Speech and Audio Processing,1995, 3(1):72-83.
[22] REYNOLDS D A,QUATIERI T F,DUNN R B. Speaker Verification Using Adapted Gaussian Mixture Models[J]. Digital Signal Processing,2000,10(1/2/33):19-41.

CHEN Sizhu, born in 1996, postgraduate. Her main research interests include wireless signal processing and language recognition.

LONG Hua, born in 1963, Ph.D, professor, is a member of CCF(No. B3460M). Her main research interests include Audio signal processing and analysis, big data and wireless network.