基于改进 CFCC 特征提取的语种识别算法研究

龙华，黄张衡，邵玉斌，杜庆治，苏树盟
（昆明理工大学信息工程与自动化学院，云南 昆明 650500）

摘 要：针对在低信噪比下语种识别准确率低的问题，提出一种基于分数阶小波变换的语种识别算法。首先，在特征提取前端采用自适应滤波法对带噪信号进行噪声滤除，以减小噪声对特征提取的影响，提升系统对带噪信号的处理能力。其次，采用新型分数阶小波变换作为小波基函数来模拟信号在耳蜗基底膜上的传播过程，利用非线性幂函数对信号进行压缩处理。最后，通过模拟人耳听觉过程提取改进耳蜗滤波器倒谱系数（CFCC）。实验结果表明，改进 CFCC 与传统 CFCC 相比显著提升了语种识别准确率，在 0 dB 信噪比下语种识别准确率平均提升了 11.1%，充分验证了所提算法的有效性和稳健性。

关键词：语种识别；自适应滤波；分数阶小波变换；神经网络；耳蜗滤波器倒谱系数
中图分类号：TN912.34
文献标志码：A
DOI: 10.11959/j.issn.1000−436x.2022234

Research on language recognition algorithm based on improved CFCC feature extraction

LONG Hua, HUANG Zhangheng, SHAO Yubin, DU Qingzhi, SU Shumeng
Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China

Abstract: Aiming at the problem of low language recognition rate under low signal-to-noise ratio, a language recognition method based on fractional wavelet transform was proposed. Firstly, the adaptive filtering algorithm was used to filter the noise of the noisy signal, so as to reduce the influence of noise on the feature extraction and improve the processing ability of the system for non-stationary signals. Secondly, the motion of the signal on the basilar membrane of the cochlea was simulated, and then the signal was compressed by a nonlinear power function. Finally, the improved CFCC were extracted by simulating the human hearing process. Experiments show that compared with the traditional CFCC, the language recognition rate is significantly improved, and the language recognition rate is increased by 11.1% on average under the 0 dB signal-to-noise ratio, which verifies the effectiveness and robustness of the proposed algorithm.

Keywords: language recognition, adaptive filtering, fractional wavelet transform, neural network, cochlear filter cepstral coefficient

0 引言

语种识别（LID, language identification）作为语音信号处理的前端系统，根据语音序列所表征的不同特征信息识别出所属的语言类别，广泛应用于语种识别系统、智能系统等[1]。传统的语种识别系统通常基于不同语音所具有的音素层特征与声学层特征，音素层特征采用了每个语种之间的音节和音素的出现频率与组合方式的差异作为分类依据来进行语种识别[2]。声学特征则通过对语音信号进行一系列的变换提取出能够表征整个语音信号的局部特征来进行分类识别。随着科技的进步以及研究者之间的密切联系合作，对语种识别技术的精确度要求也越来越高，特别是在噪声环境下能够有效提

收稿日期：2022−07−28；**修回日期：**2022−11−30
通信作者：黄张衡，hzh29@qq.com
基金项目：国家自然科学基金资助项目（No.61761025）
Foundation Item: The National Natural Science Foundation of China (No.61761025)

取具有稳定性与稳健性的声学特征尤为关键。

常用的声学特征包括基于梅尔滤波器的梅尔频率倒谱系数（MFCC, Melfrequency cepstral coefficient）[3-4]、伽玛通频率倒谱系数[5]（GFCC, Gammatone frequency cepstral coefficient）以及耳蜗滤波器倒谱系数[6]（CFCC, cochlear filter cepstral coefficient）。文献[7]首次提出使用小波变换作为耳蜗滤波器的冲激响应函数来模拟人耳听觉过程提取传统 CFCC 特征并应用于语音识别，取得了一定的识别效果。但由于小波变换主要在时域对信号进行分析，在分数域并不能很好地对信号进行有效处理，且在低信噪比下传统 CFCC 的抗噪性能并不理想[8]。

为了提升 CFCC 在低信噪比下的抗噪稳健性，李晶皎等[9]利用信号相位匹配方法消除语音信号噪声，再将 Teager 能量算子融合 CFCC 特征组成新的特征参数，相较单一特征，融合特征提升了语种识别准确率。文献[10]将语音相位特征与 CFCC 特征相融合应用于说话人识别系统来提高系统的识别准确率和稳健性。虽然融合特征的识别准确率以及抗噪性有所提升，但是其仅单纯地进行特征融合，语音信号时域信息的固有不足以及信号时频域、分数域信息未能被有效地表征[11]，需要考虑信号的时频域以及分数域信息。Patel 等[11]提出基于对数非线性函数和瞬时频率来提取 CFCC 特征参数进行语音信号的检测，其提取的特征具有较高的抗噪性，且弥补了传统 CFCC 特征不能有效提取信号中时频域信息的缺陷，但其未能有效分析信号分数域中的信息[8]。为了进一步提升低信噪比下语音识别性能，文献[12]在特征提取前端引入语音增强技术，通过谱减法与特征提取相融合，提取更具稳健性的特征。其在特征提取前端进行降噪处理，在特征提取的过程中进行了非线性信号压缩，但也忽略了特征中的分数域信息。

上述方法提取 CFCC 特征参数并未有效考虑噪声环境下语音信号的时频域、分数域信号信息以及语音信号中所含有的声压强度对特征参数的影响。本文首先在特征提取前端引入自适应滤波[13]对语音信号进行增强处理。然后采用新型分数阶小波变换代替小波变换作为小波基函数来模拟信号在耳蜗基底膜上的传播过程，以弥补小波变换不能有效在分数域表征特征的缺陷，且能够在时频域以及分数域对信号进行多辨分析。另外，基于小波变换以及分数阶小波变换的耳蜗滤波函数都未能表现出基底膜滤波器的非对称性与声压强度[14]，因此，在分数阶小波变换滤波函数中引入能够反映声音强度的啁啾参数[15]以更有效地反映语音信号在耳蜗中的声压强度，使提取到的特征更具区分性。再利用非线性幂函数对信号进行压缩处理，将其由能量值变为感知响度，得到基于自适应滤波的新型分数阶耳蜗滤波器倒谱特征（NFCFCCAF, new fractional cochlear filter cepstral coefficient based on adaptive filtering）。该特征突破了传统 CFCC 特征基于小波变换与立方根线性函数局限于时频域分析信号的缺点，在能够继承多分辨分析优点的同时还可以对噪声信号在时频域和分数域进行多辨分析[14]。最后，将提取到的特征语谱图输入分类网络 FcaNet-MobileNetV2 中进行分类识别。

1 CFCC 提取

CFCC 是基于听觉感知模拟人耳的听觉过程提取的，传统的 CFCC 特征采用听觉正变换模拟声音从外界传入人耳经过鼓膜放大声波振动能量，再通过镫骨底板的活塞运动传入内耳耳蜗引起耳蜗基底膜上的振动。文献[6]采用小波基函数作为耳蜗滤波函数通过小波变换来模拟信号在耳蜗基底膜上的运动，使信号通过耳蜗滤波器组、毛细胞窗口、非线性响度变换以及离散余弦变换（DCT, discrete cosine transform）来实现 CFCC 特征提取。

小波变换能够突破时频域的局限，更好地处理分析非线性信号，设原始时域语音信号 $x(t)$，经听觉变换输出 $T(a,b)$ 定义为

$$T(a,b)=\int_{-\infty}^{+\infty}x(t)\psi_{a,b}(t)\,\mathrm{d}t \tag{1}$$

其中，耳蜗基底膜上的冲击响应函数 $\psi_{a,b}(t)$ 定义为

$$\psi_{a,b}(t)=\frac{1}{\sqrt{a}}\left[\frac{t-b}{a}\right]^{\alpha}\exp\left[-2\pi f_{\mathrm{L}}\beta\left(\frac{t-b}{a}\right)\right]\cdot \cos\left[2\pi f_{\mathrm{L}}\left(\frac{t-b}{a}+\theta\right)\right]u(t-b) \tag{2}$$

其中，$\alpha>0,\beta>0$，α 和 β 决定了 $\psi_{a,b}(t)$ 的时频域形状和宽度，一般情况下，α=3 且 β=0.2 时降噪效果较佳。a 为尺度因子，$0<a\leqslant 1$，由耳蜗滤波器组的最低中心频率 f_{L} 与中心频率 f_{C} 的比值决定，即 $a=\frac{f_{\mathrm{L}}}{f_{\mathrm{C}}}$；$b$ 为位移因子，为随时间可变的实

数；θ 为控制冲激响应角度的初始相位；$u(t)$ 为单位阶跃函数，是单位冲激函数的积分。

毛细胞函数用来模拟人耳耳蜗基底膜上由大量毛细胞构成的螺旋器（柯蒂氏器），通过螺旋器毛细胞的换能作用把声波的机械振动能量转变为生物电能最终转化为大脑可分析的电信号。其模拟过程为

$$h(a,b)=[T(a,b)]^2 \tag{3}$$

当前滤波器中心频率响应相关神经穗就可以用每个波段的毛细胞输出 $S(i,j)$ 来表示，即

$$S(i,j)=\frac{1}{d}\sum_{b=l}^{l+d-1}h(i,b),l=1,L,2L,\cdots;\forall i,j \tag{4}$$

其中，L 为帧移，实验中一般取 $L=\frac{d}{2}$；j 为窗的个数；可变窗长 $d=\max\{3.5\tau_i,20\text{ ms}\}$，$\tau_i$ 为第 i 个滤波器中心频带中心频率的时间长度，即 $\tau_i=\frac{1}{f_C}$。

毛细胞输出通过非线性的立方根响度函数模拟非线性响度变换，由能量值变为感知响度[16]，得到非线性响度变换的输出

$$y(i,j)=[S(i,j)]^{\frac{1}{3}} \tag{5}$$

最后，将非线性响度变换输出经 DCT 进行去相关得到传统 CFCC。

2　改进特征提取

2.1　基于 VMD 的自适应滤波降噪

本节主要研究低信噪比环境下的语种识别，定义采样后带噪声的语音信号为

$$s(n)=x(n)+g(n) \tag{6}$$

其中，$x(n)$ 为原时域语音信号 $x(t)$ 采样后的信号，$g(n)$ 为零均值高斯白噪声，其平均信噪比定义为

$$\text{SNR}=10\lg\frac{\sum_{n=1}^{H}x^2(n)}{\sum_{n=1}^{H}g^2(n)} \tag{7}$$

其中，$\sum_{n=1}^{H}x^2(n)$ 为原语音信号能量和，$\sum_{n=1}^{H}g^2(n)$ 为白噪声能量和，H 为全语音的总采样点数。信噪比越低，语音信号被噪声淹没的部分越大，信号中的特征越不容易被提取。因此在低信噪比下带噪语音信号中提取的语种识别特征的识别准确性不高。

为了进一步提高带噪语音信号的识别性能，可以在特征提取前端对语音信号进行滤噪处理。由于本文实验采用添加零均值高斯白噪声后的语音信号，高斯白噪声属于平稳噪声，而常用的频域滤波法可以对带噪信号进行处理，但对于带内噪声其降噪效果并不佳[13]。对于平稳噪声，自适应滤波却能够不完全依赖噪声信号的先验统计特性而根据算法自适应调整参数，使输出信号达到最优，且对带内噪声有更好的处理效果[16-17]。本文实验对带噪语音信号进行变模态分解（VMD, variational mode decomposition）处理，然后通过基于归一化最小均方（NLMS, normalization least mean square）自适应滤波器降噪，该方法对平稳噪声有较好的处理效果。基于 VMD 的自适应滤波系统如图 1 所示，其中，$s(n)$ 为带噪语音信号，$y(n)$ 为自适应滤波器的输出，$\boldsymbol{W}$ 为滤波器的权值系数向量，$e(n)$ 为误差信号，对输入带噪语音信号进行端点检测后，取出语音信号中的所有无话帧并求取均值作为信号的参考噪声 $n(n)$，因此参考信号为

$$d(n)=s(n)-n(n) \tag{8}$$

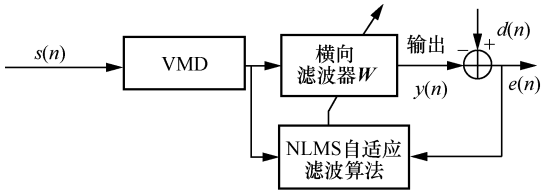


图 1　基于 VMD 的自适应滤波系统

当均方误差达最小时，滤波器的系数向量为最佳权值，滤波器的滤波效果最佳。对于 M 阶滤波器，输出 $y(n)$ 可表示为

$$y(n)=\sum_{m=1}^{M}w_m(n)s(n-m+1) \tag{9}$$

其中，$w_m(n)$ 为权值系数，则自适应滤波的权值系数向量为

$$\boldsymbol{W}=[w_1,w_2,\cdots,w_M]^{\mathrm{T}} \tag{10}$$

滤波器的输入信号向量为

$$\boldsymbol{S}(n)=[s(n-1),s(n-2),\cdots,s(n-M)]^{\mathrm{T}} \tag{11}$$

则输出信号为

$$y(n)=\boldsymbol{S}^{\mathrm{T}}(n)\boldsymbol{W}=\boldsymbol{W}^{\mathrm{T}}\boldsymbol{S}(n) \tag{12}$$

误差信号 $e(n)$ 可表示为

$$e(n)=d(n)-y(n)=d(n)-\boldsymbol{S}^{\mathrm{T}}(n)\boldsymbol{W} \tag{13}$$

则误差平方的数学期望为

$$\sigma(n)=\mathrm{E}[d^2(n)]-2\mathrm{E}[d(n)\boldsymbol{S}^{\mathrm{T}}(n)]\boldsymbol{W}+\boldsymbol{W}^{\mathrm{T}}\mathrm{E}[\boldsymbol{S}(n)\boldsymbol{S}^{\mathrm{T}}(n)]\boldsymbol{W} \tag{14}$$

令 $R_{\mathrm{ds}}=\mathrm{E}[d(n)\boldsymbol{S}^{\mathrm{T}}(n)]$，$R_{\mathrm{ss}}=\mathrm{E}[\boldsymbol{X}(n)\boldsymbol{S}^{\mathrm{T}}(n)]$，则式(14)改写为

$$\sigma(n)=\mathrm{E}[d^2(n)]-2R_{\mathrm{ds}}\boldsymbol{W}+\boldsymbol{W}^{\mathrm{T}}R_{\mathrm{ss}}\boldsymbol{W} \tag{15}$$

取 $\sigma(n)$ 的瞬时估计值 $\hat{\sigma}(n)=0.5e^2(n)$，则 NLMS 算法权值向量迭代式为

$$\boldsymbol{W}(n+1)=\boldsymbol{W}(n)+\frac{\mu}{\gamma+\|\boldsymbol{S}(n)\|^2}\boldsymbol{S}(n)e(n) \tag{16}$$

其中，μ 为收敛系数，主要控制算法的收敛速度与稳定误差；γ 取 0.001。

为了测试自适应滤波法在低信噪比下的降噪效果，设计实验在 –10~0 dB 噪声下对带噪语音信号进行降噪处理。本文实验引入占空比来描述自适应滤波的降噪效果，占空比定义为一段带噪语音信号中纯噪声时间与语音信号时间的比值，主要反映了纯噪声时间的长短对自适应滤波降噪的影响。首先采用 –10~0 dB 信噪比的带噪语音信号各 1 000 条，在每一信噪比下分别对带噪语音信号进行端点检测，求出其占空比，实验发现语音信号的占空比在 20%~40%之间。因此将每一信噪比下语音信号以 5%占空比为刻度分为 5 类，并对每一类占空比下语音信号进行降噪滤波后求取改善信噪比均值。其在不同信噪比、不同占空比下的改善信噪比和均方根误差分别如图 2 和图 3 所示。

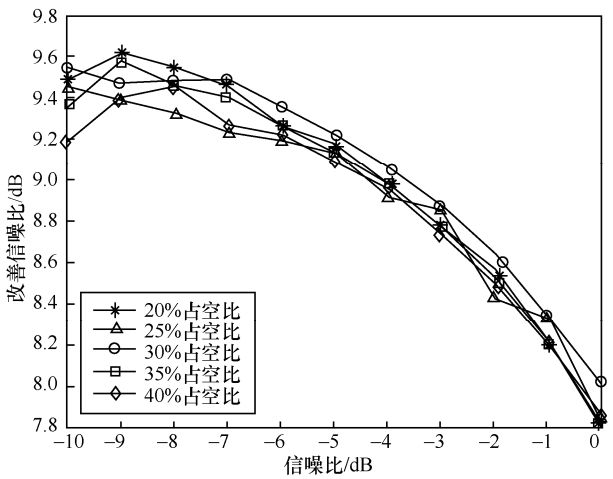


图 2　不同信噪比、不同占空比下的改善信噪比

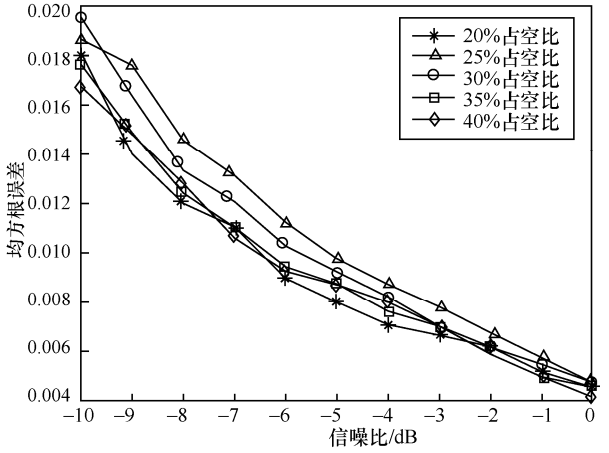


图 3　不同信噪比、不同占空比下的均方根误差

从图 2 和图 3 中可以看出，自适应滤波在低信噪比下对不同占空比的带噪语音信号降噪效果相对稳定，当占空比为 30%时，其滤噪后改善信噪比较其他占空比要高，且整体相对稳定。

在信噪比为–5 dB、不同占空比下自适应降噪过程中的收敛情况如图 4 所示，分析不同占空比下前 10 000 个采样点、100 次重复实验时的平均均方根误差。从图 4 中可以看出，随着迭代次数的增加，不同占空比下的曲线很快便收敛，其中当占空比为 20%时，收敛速度最快，在迭代 2 000 次时便收敛，其滤噪效果较佳。在不同信噪比、不同占空比下的实验结果表明，采用自适应滤波降噪在不同占空比下均有较快的收敛速度，且降噪效果比较稳定。

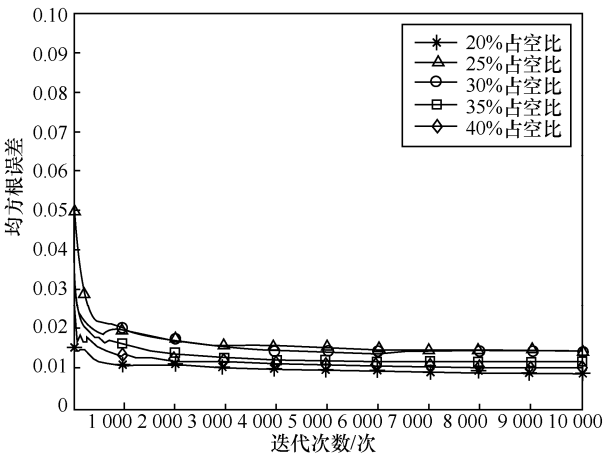


图 4　不同占空比下收敛曲线

2.2　基于新型分数阶小波变换的 NFCFCCAF 特征提取

传统的小波变换虽然能够突破时频域限制对噪声信号进行有效处理，但小波变换主要是基于时

频域信号，不具有分数傅里叶变换分数域表征的缺陷[18]，其在分数域中并不能有效表征信号局部特征。而分数阶小波变换不仅可以在时频域与分数域分析信号，而且在继承多分辨分析优点的同时还可以对信号在时频域与分数域进行多辨分析，更具抗噪性。

设带噪语音信号 $s(n)$ 经自适应滤波降噪后的连续语音信号为 $s(t)$，采用分数阶母小波函数 $\psi_{p,a,b}(t)$ 作为耳蜗基底膜上的耳蜗滤波函数，则语音信号经听觉变换输出 FRWT(p,a,b) 为

$$\mathrm{FRWT}(p,a,b)=\int_{-\infty}^{+\infty}s(t)\psi_{p,a,b}(t)\mathrm{d}t \tag{17}$$

耳蜗滤波函数 $\psi_{p,a,b}(t)$ 定义为

$$\psi_{p,a,b}(t)=\mathrm{e}^{-\mathrm{j}\frac{t^2-b^2}{2}\cot\omega}\psi_{a,b}(t) \tag{18}$$

其中，$\omega=\dfrac{p\pi}{2}$；p=[0,1]，当 $p=1$ 时，式(18)退化为式(2)传统小波耳蜗滤波函数，本文中阶数 p 取 0.5。

设时间函数为 $h(t)$，在分数傅里叶变换下，分数阶卷积定义为

$$s(t)\Theta_p h(t)\xleftrightarrow{f^p}\sqrt{2\pi}S_p(u)H(u\csc\omega) \tag{19}$$

其中，Θ_p 为分数阶卷积算子，$S_p(u)$ 与 $H(u\csc\omega)$ 分别为 $s(t)$ 与 $h(t)$ 的 p 阶分数傅里叶形式。则分数阶小波变换分数域形式表示为

$$\begin{aligned}\mathrm{NFRWT}(\omega,a,b)=&\int_{-\infty}^{+\infty}\sqrt{2\pi a}S_p(u)\cdot\\&\psi^*(au\csc\omega)\Gamma_{-p}(u,b)\mathrm{d}u\end{aligned} \tag{20}$$

其中，$\psi^*(au\csc\omega)$ 为 $\psi(t)$ 的 FT（变换元进行了尺度 $\csc\omega$ 伸缩），核函数 $\Gamma_p(u,b)$ 定义为

$$\Gamma_p(u,b)=\begin{cases}\sqrt{\dfrac{1-\mathrm{j}\cot\omega}{2\pi}}\mathrm{e}^{\frac{\mathrm{j}}{2}(b^2+u^2)\cot\omega-\mathrm{j}bu\csc\omega}, & p\neq k\pi\\ \delta(b-u), & p=2k\pi\\ \delta(b+u), & p=(2k+1)\pi\end{cases} \tag{21}$$

其中，u 为分数频率。因此，式(17)可以改写为

$$\mathrm{NFRWT}(p,a,b)=\mathrm{e}^{-\mathrm{j}\frac{b^2}{2}\cot\omega}\int_{-\infty}^{+\infty}s(t)\mathrm{e}^{\mathrm{j}\frac{t^2}{2}\cot\omega}\psi_{a,b}(t)\mathrm{d}t \tag{22}$$

由于特征提取中耳蜗滤波函数的幅频响应曲线关于中心频率对称，其并未有效体现人耳基底膜曲线的非对称性，且其幅频响应曲线也与强度无关，这与基底膜的强度相关特性并不相符[15]。因此，为了更有效地体现出人耳基底膜曲线的非对称性且符合人耳基底膜的强度相关特性，使函数能够对语音信号进行有效处理，在耳蜗滤波器函数中引入一个能够反映声音强度的啁啾参数 $\xi\ln\left(\dfrac{t-b}{a}\right)$，$\ln\left(\dfrac{t-b}{a}\right)$ 为对时间的对数，啁啾因子 ξ 随着声压强度 P_s（单位为 dB）的变化而变化[15]。

$$\xi=3.38-0.107P_s \tag{23}$$

$$P_s=20\lg\frac{P_e}{P_0} \tag{24}$$

其中，$P_0=2\times10^{-5}$ Pa 为参考声压，P_e 为有效声压。

$$P_e=\sqrt{\frac{1}{N}\sum_{n=1}^{N}x_n^2} \tag{25}$$

其中，N 为所取的采样点数，x_n 为对语音信号 $x(t)$ 的采样点。语音信号的声压级曲线如图 5 所示。

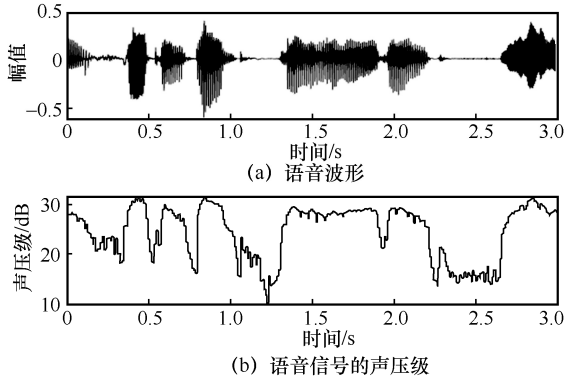


图 5　语音信号的声压级曲线

声压作为声扰动而产生的逾量压强，是空间位置和时间的函数，可以定量描述声波，声压强度级能够随着声音的不同而变化，更好地反映了人耳对声音强弱的变化。因此，在耳蜗滤波函数中加入反映声压强度的啁啾参数 $\xi\ln\left(\dfrac{t-b}{a}\right)$ 的新型耳蜗滤波器函数定义为

$$\varphi_{p,a,b}(t)=\frac{1}{\sqrt{a}}\left(\frac{t-b}{a}\right)^{\alpha}\mathrm{e}^{-\mathrm{j}\frac{b^2-t^2}{2}\cot\omega}r(t)\phi(t) \tag{26}$$

$$r(t)=\exp\left[-2\pi f_{\mathrm{L}}\beta\left(\frac{t-b}{a}\right)\right] \tag{27}$$

$$\phi(t)=\left[\cos2\pi f_{\mathrm{L}}\left(\frac{t-b}{a}+\xi\ln\left(\frac{t-b}{a}\right)\right)+\theta\right]u(t-b) \tag{28}$$

其中，$a=\dfrac{f_{\mathrm{L}}}{f_{\mathrm{C}}}$，$b$ 为位移因子，α=3，β=0.2；θ 为控制冲激响应角度的初始相位；$u(t)$ 为单位阶跃函数，当 ξ=0 时，新型滤波函数 $\varphi_{p,a,b}(t)$ 变为原耳蜗滤波函数 $\psi_{p,a,b}(t)$。新型耳蜗滤波器组函数的频率响应如图 6 所示，其呈现出了非常明显的非对称性，相较于传统耳蜗滤波函数频率响应曲线，这也体现了人耳基底膜曲线的非对称性。

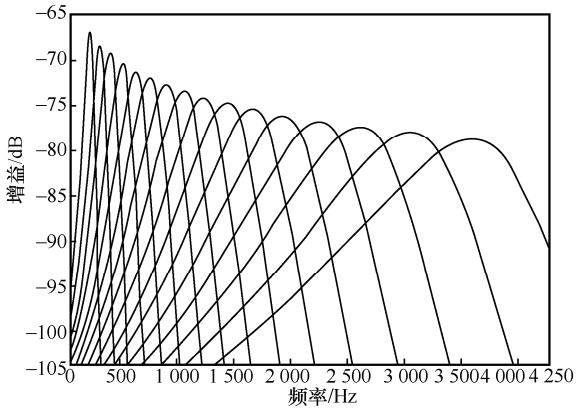


图 6 新型耳蜗滤波器组函数的频率响应

则语音信号经听觉变换输出 NFRWT(p,a,b) 由式(22)改为

$$\mathrm{NFRWT}(p,a,b)=\int_{-\infty}^{+\infty}s(t)\varphi_{p,a,b}(t)\mathrm{d}t \tag{29}$$

毛细胞滑动窗口函数定义为

$$h(p,a,b)=[\mathrm{NFRWT}(p,a,b)]^2 \tag{30}$$

$$S(i,\lambda)=\frac{1}{d}\sum_{b=l}^{l+d-1}h(i,b),l=1,L,2L,\cdots;\forall i,\lambda \tag{31}$$

其中，L 为帧移，实验中取 $L=\dfrac{d}{2}$；λ 为窗的个数；可变窗长 $d=\max\{3.5\tau_i,20\ \mathrm{ms}\}$，$\tau_i$ 为第 i 个滤波器中心频带中心频率的时间长度，即 $\tau_i=\dfrac{1}{f_{\mathrm{C}}}$。

传统 CFCC 特征提取过程中响度函数采用式(5)立方根形式，其可以有效地模拟信号，还可以为非线性幂函数或对数形式。通过实验对比可知，在噪声环境下立方根函数与对数函数较非线性幂函数的识别效果并不理想[12]，因此本文实验采用非线性幂函数来模拟人耳听觉特性。由于非线性幂函数更符合人耳听觉神经的压缩感知，通过非线性幂函数来对毛细胞输出信号进行非线性响度变换，使其由能量值变为感知响度[12]。实验中通过对比调整不同幂函数，当幂函数的指数取 0.25 时识别性能最佳，因此，实验中模拟过程为

$$y(i,\lambda)=[S(i,\lambda)]^{0.25} \tag{32}$$

将非线性响度变换输出经离散余弦变换进行去相关，即

$$\mathrm{CFCC}_{(n,j)}=\sqrt{\frac{2}{M}}\sum_{m=1}^{M-1}y(m,\lambda)\cos\left(\frac{\pi n(m-0.5)}{N}\right) \tag{33}$$

其中，n 为特征变换后每帧特征的维数，M 为耳蜗滤波器个数，$0<n<N,0\leqslant m\leqslant M$。

由于人耳在不同频率声波之间的听觉敏感度存在差异，频率较低的声音在人耳的耳蜗基底膜上行波传递的距离远大于频率较高的声音。因此，通过升半正弦倒谱提升来减少低维中分量的占比，进而可提升高维分量的作用，升半正弦倒谱窗函数定义为

$$w(i)=0.5+0.5\sin\left(\frac{\pi i}{N}\right),1\leqslant i\leqslant N \tag{34}$$

倒谱提升后的 CFCC 为

$$\mathrm{NFCFCC}_i=\mathrm{CFCC}(i)w(i) \tag{35}$$

最后得到新的特征参数 NFCFCCAF，其提取过程如图 7 所示。

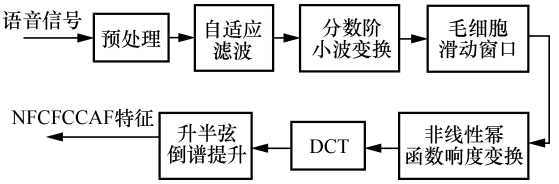


图 7 NFCFCCAF 提取过程

2.3 动态特征提取

采用新型分数阶小波基函数作为耳蜗基底膜上的耳蜗滤波函数，模拟人耳听觉过程提取的特征参数反映了语音信号的静态特性，而人耳听觉过程具有动态特性，为了更好地模拟听觉过程，本文实验提取了能够表征语言信号动态特性的一阶差分 ΔNFCFCCAF 特征，再将其与 NFCFCCAF 特征进行融合构成融合特征 NFCFCCAF-DS，以提升语种识别准确率。最后将融合特征与 NFCFCCAF 特征作为不同分类网络的输入进行语种识别比较，以验证本文算法提取特征的稳健性与有效性。

3　FcaNet-MobileNetV2 识别模型

目前，ResNet 广泛应用于语种识别且能够表现出高竞争力的识别性能，但其在小样本数据集中网络的宽度和深度容易出现过拟合现象，降低整个网络的泛化能力[19]。因此实验采用基于倒残差结构的 MobileNetV2 网络来进行准确率测试，以对小样本数据集进行有效处理，提取特征中更多的高维信息[20]。且所提取 NFCFCCAF 特征中包含了大量的频域信息[8]，因此通过在网络模型中引入频域注意力机制使网络模型能够捕捉更多的频域信息，提升网络的区分性。

本文将轻量化卷积神经网络 MobileNetV2[20]作为识别模型的主干网络，在其 17 个瓶颈结构中添加了注意力机制模块 FcaNet[21]构成 FcaNet-Mobile- NetV2 分类识别模型。其中，FcaNet 作为频域通道注意力机制，是对 SENet[22]的改进，由于 SENet 的全局平均池化（GAP, global average pooling）为二维离散余弦变换的低频部分，而特征图中大量的中高频信息被舍弃了。因此，文献[21]提出了多谱注意力模块 FcaNet，将通道注意力机制的压缩扩展到了频域，进而引入更多的频率分量信息以达到识别度提升的目的。MobileNetV2 属于轻量级识别网络，其是在 MobileNetV1 网络的基础上改进反向残差块与线性瓶颈而来的，在保留了 MobileNetV1 网络中深度可分离卷积加速网络思想的同时能够更好地提取关键信息提升识别准确率[23]。因此，为了更好地提取出特征语谱图中的关键频域信息且有效地分类识别，本文在 MobileNetV2 主干网络瓶颈中加入频域通道注意力机制模块以辅助其有效地提取特征语谱图中的特征信息，组成 FcaNet-MobileNetV2 识别模型，如图 8 所示。

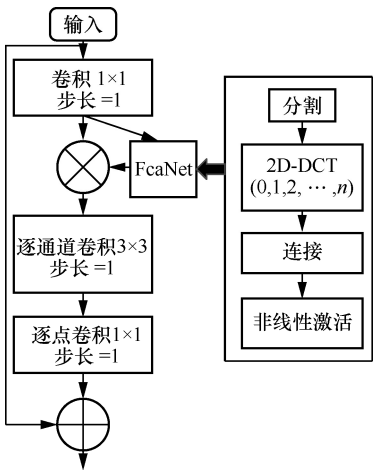


图 8　FcaNet-MobileNetV2 识别模型

4　实验结果及分析

4.1　数据准备及参数设置

1) 数据准备

本文实验采用公共数据集语料库 LibriVox 中的单通道 wav 音频信号文件，其中包括英语、法语、德语、意大利语、西班牙语这 5 个语种。语音的采样率 f_s=16 000 Hz，每条语音信号时长为 3 s，总共有 10 000 条实验语音，每个语种分别有 1 400 条训练集与 600 条测试集语音文件。实验以 NoiseX-92[24]公共噪声库中的白噪声为噪声源，构建了信噪比分别为 –5 dB、0 dB、5 dB、10 dB、15 dB 情况下的训练测试语料并且在每个数据集所对应的语种分别打上标签，如表 1 和表 2 所示。

表 1　不同语种实验训练集

语种	标签	训练集/条				
		–5 dB	0 dB	5 dB	10 dB	15 dB
英语	0	1 400	1 400	1 400	1 400	1 400
法语	1	1 400	1 400	1 400	1 400	1 400
德语	2	1 400	1 400	1 400	1 400	1 400
意大利语	3	1 400	1 400	1 400	1 400	1 400
西班牙语	4	1 400	1 400	1 400	1 400	1 400

表 2　不同语种实验测试集

语种	标签	测试集/条				
		–5 dB	0 dB	5 dB	10 dB	15 dB
英语	0	600	600	600	600	600
法语	1	600	600	600	600	600
德语	2	600	600	600	600	600
意大利语	3	600	600	600	600	600
西班牙语	4	600	600	600	600	600

本文实验采用融合了频域通道注意力机制的轻量化卷积神经网络 FcaNet-MobileNetV2 作为分类网络对特征语谱图进行分类识别。其中，分类网络的训练圈数 epoch 参数设置为 50，batchsize 为 50，worker 数为 4，模型的学习率设置为 0.000 1；将提取出的二维特征输入分类网络中训练 10 次，将每次神经网络最后一次循环的输出节点作为语种的识别模型来对语种测试集进行识别，取 10 次结果的平均值作为语种识别结果。性能评价指标采用美国国家标准与技术研究院语种评测规则，表示为

$$\overline{R_{\mathrm{L}}}=\frac{E_{\mathrm{L}}+F_{\mathrm{L}}+G_{\mathrm{L}}+I_{\mathrm{L}}+S_{\mathrm{L}}}{T_{\mathrm{L}}} \tag{36}$$

其中，E_{L} 为英语的正确识别个数，F_{L} 为法语的正确识别个数，G_{L} 为德语的正确识别个数，I_{L} 为意大利语的正确识别个数，S_{L} 为西班牙语的正确识别个数，T_{L} 为测试集总数，$\overline{R_{\mathrm{L}}}$ 为语种识别准确率。

4.2 实验

1) 实验 1

为了验证本文提出的非线性幂函数对信号进行压缩模拟过程提取的耳蜗滤波器倒谱系数（FCFCC, function of cochlear filter cepstral coefficient）与基于立方根函数、非线性函数提取的 CFCC 的语种识别效果，实验提取了传统 CFCC 特征，采用本文所提非线性幂函数式(32)提取的 FCFCC 特征，文献[11]中基于对数非线性函数的耳蜗滤波器倒谱系数（LCFCC, logarithm of cochlear filter cepstral coefficient）以及文献[12]、文献[25]所提基于非线性幂函数的 CFCC0、CFCC1 特征。上述特征在基于传统 CFCC 特征提取基础上，仅改变了提取过程中的非线性幂函数，最后将其输入分类网络 FcaNet-MobileNetV2 中进行分类识别，不同听觉特性函数识别准确率如表 3 所示。

通过分析发现，在 –5 ~15 dB 信噪比下，本文所提非线性幂函数提取的 FCFCC 特征参数语种识别准确率较文献[12]、文献[25]所提 CFCC0、CFCC1 以及传统 CFCC 有一定的提升。与采用对数非线性函数来模拟听觉特性函数提取的 LCFCC 特征相比，其平均识别准确率提升了 4.79%。实验说明了采用本文所提非线性幂函数模拟人耳听觉特性函数提取的 FCFCC 特征参数在噪声环境下具有一定的抗噪性和有效性。

2) 实验 2

为了验证本文提出的新型分数阶小波变换作为耳蜗基底膜上的耳蜗滤波函数所提取的 NFCFCC 特征参数的稳健性，分别提取不同信噪比下 CFCC 特征、MFCC 特征、GFCC 特征以及 Fbank 特征，将其输入分类网络 FcaNet-MobileNetV2 中进行分类识别比较，不同特征参数识别准确率如表 4 所示。

从表 4 可知，本文利用新型分数阶小波基函数来模拟信号在人耳基底膜上的运动过程，且引入能够反映声音强度的啁啾参数所提取的 NFCFCC 特征较其他特征语种识别准确率有显著提升，与 MFCC 特征相比其平均识别准确率提升了 4.77%，与 CFCC 特征相比其平均识别准确率提升了 6.58%。因此，采用引入啁啾参数的新型分数阶小

表 3 不同听觉特性函数识别准确率

特征函数	听觉特性函数	识别准确率					平均识别准确率
		–5 dB	0 dB	5 dB	10 dB	15 dB	
CFCC	$\frac{1}{3}$	66.8%	70.76%	72.77%	74.16%	79.34%	72.77%
LCFCC	对数	63.73%	68.6%	74.83%	75.06%	78.7%	72.18%
CFCC0	0.101	67.73%	71.46%	75.36%	80.96%	83.46%	75.79%
CFCC1	$\frac{1}{15}$	65.63%	73.8%	76.86%	78.43%	80.1%	74.96%
FCFCC	0.25	68.97%	73.4%	77.5%	80.36%	84.63%	76.97%

表 4 不同特征参数识别准确率

特征参数	识别准确率					平均识别准确率
	–5 dB	0 dB	5 dB	10 dB	15 dB	
MFCC	59.07%	70.03%	76.96%	82.23%	84.6%	74.58%
GFCC	70.9%	72.86%	74.86%	81.5%	83.2%	76.66%
Fbank	67.06%	73.2%	77.16%	80.3%	85.56%	76.65%
CFCC	66.8%	70.76%	72.77%	74.16%	79.34%	72.77%
NFCFCC	71.3%	73.3%	79.96%	84.53%	87.67%	79.35%

波变换来模拟耳蜗滤波器能够有效地突破小波变换对信号进行时频域分析的缺陷，从而能够在时频域分析且在继承多分辨分析优点的同时对信号在时频与分数域进行多辨分析，进一步提升了特征参数的有效性与稳健性。

3) 实验 3

为了进一步提升改进特征参数的稳健性，在实验 2 提取 NFCFCC 特征算法的基础上，在特征提取前端引入自适应滤波对信号进行增强去噪，提取更具抗噪性的 NFCFCCAF 和文献[12]所提基于谱减法与非线性幂函数的 FFPSS 特征参数，分别在不同信噪比下采用分类网络 FcaNet-MobileNetV2 进行语种识别准确率比较，不同特征识别准确率如表 5 所示，各语种在不同信噪比下的识别准确率如表 6 所示。

从表 5 可知，在特征提取前端引入自适应滤波对噪声信号进行语音增强处理，再利用实验 2 提取 NFCFCCAF 特征与文献[12]提取的 FFPSS 特征，分别在 –5~15 dB 信噪比下分别进行语种识别比较，NFCFCCAF 特征参数的语种识别准确率有显著提升。在低信噪比下，NFCFCCAF 特征较 FFPSS 特征其平均识别准确率提升了 2.47%，说明本文算法具有一定的抗噪性与可行性。对比表 4 与表 5 中 NFCFCC 特征与 NFCFCCAF 特征的语种识别准确率可知，引入自适应滤波降噪较未采用自适应滤波降噪所提取的特征显著提升了语种识别准确率，平均识别准确率提升了 3.7%，且较传统 CFCC 特征，其平均识别准确率提升了 10.28%，提升了系统的稳健性。

从表 6 可知，–5 ~15dB 信噪比下各语种的识别效果较好，且识别准确率整体呈上升趋势。说明了采用本文算法对语音信号提取 NFCFCCAF 特征再利用 FcaNet-MobileNetV2 网络进行分类识别能够更好地提取语种之间的相关信息并且达到了较好的区分度。在低信噪比下，相较于其他语种，法语的语种识别准确率较低，平均识别准确率达 80%，而英语的识别准确率最高，平均识别准确率达 88%。这说明法语提取的文本特征区分性较其他语种并不高，而英语提取到的文本特征参数区分性最高。

4) 实验 4

由于 NFCFCCAF 特征参数所表征的为语音信号的静态特征，并不能较好地表现出语音的实际动态特性。为测试本文所提 NFCFCCAF 的语种识别有效性，求取 NFCFCCAF 特征参数的一阶差分系数[24]、再与 NFCFCCAF 特征进行融合处理得到融合特征 NFCFCCAF-DS。

为了验证本文所提分类网络 FcaNet-MobileNetV2 的识别性能，采用不同的分类网络 FcaNet-MobileNetV2、MobileNetV2 以及 ResNet[24]对特征参数进行分类识别。其识别结果如表 7 所示。

从表 7 可知，在–5~15 dB 信噪比下，不同特征在不同分类网络中都表现出较好的识别效果。在 FcaNet-MobileNetV2 分类网络中，2 种特征平均识别准确率达 83.05%和 85.65%；在 MobileNetV2 分类网络中，2 种特征平均识别准确率达 81.04%与 83.6%。在这 2 种特征下 FcaNet-MobileNetV2 网络

表 5　不同特征识别准确率

特征参数	识别准确率					平均识别准确率
	–5 dB	0 dB	5 dB	10 dB	15 dB	
NFPSS	73.8%	77.03%	80.33%	83.77%	87.96%	80.58%
NFCFCCAF	75.5%	81.87%	83.66%	85.8%	88.42%	83.05%

表 6　各语种在不同信噪比下的识别准确率

语种	识别准确率				
	–5 dB	0 dB	5 dB	10 dB	15 dB
法语	71.5%	78.5%	80.5%	82.12%	86.16%
意大利语	70.5%	79.56%	83.16%	84.2%	85.16%
西班牙语	74.16%	82.4%	83.5%	85.16%	87.26%
德语	78.83%	81.56%	82.64%	87%	89.65%
英语	82.5%	87.33%	88.5%	90.5%	91.5%

表 7 不同特征在不同分类网络中的语种识别准确率

特征	分类网络	识别准确率					平均识别准确率
		−5 dB	0 dB	5 dB	10 dB	15 dB	
NFCFCCAF	FcaNet-MobileNetV2	75.5%	81.87%	83.66%	85.8%	88.42%	83.05%
	MobileNetV2	73.62%	79.7%	82.33%	84.37%	85.2%	81.04%
	ResNet	74.2%	80.36%	81.9%	84.5%	85.63%	81.30%
NFCFCCAF-DS	FcaNet-MobileNetV2	81.2%	82.97%	85.93%	87.8%	90.37%	85.65%
	MobileNetV2	79.5%	80.63%	83.82%	85.97%	88.1%	83.60%
	ResNet	77.16%	80.2%	81.2%	84.54%	88.25%	82.27%

较 MobileNetV2 网络平均识别准确率提升了 2.01%与 2.05%。这说明经过引入频域注意力机制使整个网络模型能够集中捕捉特征中的频域信息，加强了特征判别的指向性，提升了网络的识别性能。

在 ResNet 分类网络中，2 种特征平均识别准确率达 81.30%和 82.27%，FcaNet-MobileNetV2 网络较 ResNet 平均识别准确率提升了 1.75%和 3.38%。说明基于倒残差结构 FcaNet-MobileNetV2 网络能够有效处理小样本，且提取到特征中更多的高维信息以及频域信息，避免了特征信息损失，弥补了基于残差结构的 ResNet 对于小样本中不能有效提取整体特征足够多的信息缺陷，验证了 FcaNet-MobileNetV2 网络的可行性与识别优越性。同时通过不同的分类网络也验证了本文算法所提取特征参数的有效性。

另外，从本文所提 NFCFCCAF 特征参数与加上反映其动态特性的 NFCFCCAF-DS 特征参数在不同信噪比下的语种识别准确率可知，在 3 种分类网络下，NFCFCCAF-DS 动态特征参数的识别准确率都要高于静态特征 NFCFCCAF 的识别准确率。且在 FcaNet-MobileNetV2 分类网络下动态特征较静态特征平均识别准确率提升了 2.6%，特别在−5 dB 信噪比下语种识别准确率提升了 5.7%。这说明 NFCFCCAF-DS 特征参数在低信噪比下能够有效反映出语音信号局部特征动态特性，同时有效表征语音信号的完整特性，具有较好的稳健性。

5 结束语

针对低信噪比下语种识别准确率低与稳健性差的问题，提出了一种结合自适应滤波与分数阶小波变换的耳蜗倒谱系数提取算法。实验采用自适应滤波对语音信号进行噪声滤除，再将新型分数阶小波变换作为小波基函数来模拟信号在耳蜗基底膜上的运动，然后通过模拟人耳听觉过程提取出 NFCFCCAF 特征参数，最后将提取出的特征参数作为 FcaNet-MobileNetV2 网络的输入进行分类识别。实验对比了传统 CFCC 特征以及近几年经典的 Fbank 等特征，本文算法的识别准确率都有显著提升，相较于传统 CFCC 语种识别性能提升了 10.28%，有效改善了传统特征在低信噪比下识别准确率低的问题，具有较强稳健性，且更具抗噪性，提高了语种识别准确率。由于本文实验只针对特征提取进行改进，因此在未来的研究中，需要加强对语种识别的模型研究，以进一步提升语种识别性能及稳健性。

参考文献：

[1] IRTZA S, SETHU V, AMBIKAIRAJAH E, et al. Using language cluster models in hierarchical language identification[J]. Speech Communication, 2018, 100: 30-40.

[2] 苗晓晓, 徐及, 王剑. 基于降噪自动编码器的语种特征补偿方法[J]. 计算机研究与发展, 2019, 56(5): 1082-1091.
MIAO X X, XU J, WANG J. Denoising auto encoder-based language feature compensation[J]. Journal of Computer Research and Development, 2019, 56(5): 1082-1091.

[3] DAVIS S, MERMELSTEIN P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1980, 28(4): 357-366.

[4] 龙华, 杨明亮, 邵玉斌. 基于特征流融合的带噪语音检测算法[J]. 通信学报, 2020, 41(4): 134-142.
LONG H, YANG M L, SHAO Y B. Noisy voice detection algorithm based on feature stream fusion[J]. Journal on Communications, 2020, 41(4): 134-142.

[5] QI J, WANG D, JIANG Y, et al. Auditory features based on Gammatone filters for robust speech recognition[C]//Proceedings of 2013 IEEE International Symposium on Circuits and Systems. Piscataway: IEEE Press, 2013: 305-308.

[6] LI Q, HUANG Y. Robust speaker identification using an auditory-based feature[C]//Proceedings of 2010 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE

Press, 2010: 4514-4517.

[7] LI Q, HUANG Y. An auditory-based feature extraction algorithm for robust speaker identification under mismatched conditions[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2011, 19(6): 1791-1801.

[8] 刘影, 韩康康, 钱志鸿. 基于声音空间梯度的高稳健性击键识别方法[J]. 通信学报, 2020, 41(5): 96-103.
LIU Y, HAN K K, QIAN Z H. High-roubustness keystroke recognition method based on acoustic spatial gradient[J]. Journal on Communications, 2020, 41(5): 96-103.

[9] 李晶皎, 安冬, 杨丹, 等. 噪声环境下说话人识别的 TEO-CFCC 特征参数提取方法[J]. 计算机科学, 2012, 39(12): 195-197.
LI J J, AN D, YANG D, et al. TEO-CFCC characteristic parameter extraction method for speaker recognition in noisy environments[J]. Computer Science, 2012, 39(12): 195-197.

[10] 李作强, 高勇. 基于 CFCC 和相位信息的鲁棒性说话人辨识[J]. 计算机工程与应用, 2015, 51(17): 228-232.
LI Z Q, GAO Y. Robust speaker identification based on CFCC and phase information[J]. Computer Engineering and Applications, 2015, 51(17): 228-232.

[11] PATEL T B, PATIL H A. Cochlear filter and instantaneous frequency based features for spoofed speech detection[J]. IEEE Journal of Selected Topics in Signal Processing, 2017, 11(4): 618-631.

[12] 白静, 史燕燕, 薛珮芸, 等. 融合非线性幂函数和谱减法的 CFCC 特征提取[J]. 西安电子科技大学学报, 2019, 46(1): 86-92.
BAI J, SHI Y Y, XUE P Y, et al. CFCC feature extraction for fusion of the power-law nonlinearity function and spectral subtraction[J]. Journal of Xidian University, 2019, 46(1): 86-92.

[13] 吴龙文, 聂雨亭, 张宇鹏, 等. 基于变分模态分解的自适应滤波降噪方法[J]. 电子学报, 2021, 49(8): 1457-1465.
WU L W, NIE Y T, ZHANG Y P, et al. An adaptive filtering denoising method based on variational mode decomposition[J]. Acta Electronica Sinica, 2021, 49(8): 1457-1465.

[14] GUO Y, etal. Novel fractional wavelet transform: principles, MRA and application[J]. Digital Signal Processing, 2021, 110: 102937.

[15] IRINO T, PATTERSON R D. A dynamic compressive gammachirp auditory filterbank[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2006, 14(6): 2222-2232.

[16] SHAO Y, JIN Z Z, WANG D L, et al. An auditory-based feature for robust speech recognition[C]//Proceedings of 2009 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2009: 4625-4628.

[17] LV H, SHAN P F, SHI H F, et al. An adaptive bilateral filtering method based on improved convolution kernel used for infrared image enhancement[J]. Signal, Image and Video Processing, 2022, 16(8): 2231-2237.

[18] 史军, 张乃通, 刘晓萍. 一种新型分数阶小波变换及其应用[J]. 中国科学: 信息科学, 2012, 42(2): 125-135.
SHI J, ZHANG N T, LIU X P. A novel fractional wavelet transform and its applications[J]. Scientia Sinica (Informationis), 2012, 42(2): 125-135.

[19] ZHOU T Y, ZHAO Y, WU J. ResNeXt and Res2Net structures for speaker verification[C]//Proceedings of 2021 IEEE Spoken Language Technology Workshop. Piscataway: IEEE Press, 2021: 301-307.

[20] SANDLER M, HOWARD A, ZHU M L, et al. MobileNetV2: inverted residuals and linear bottlenecks[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 4510-4520.

[21] QIN Z Q, ZHANG P Y, WU F, et al. FcaNet: frequency channel attention networks[C]//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2021: 763-772.

[22] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]// Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 7132-7141.

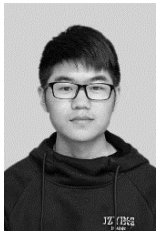
[23] 陈宗阳, 赵辉, 吕永胜, 等. 基于改进 MobileNetV2 网络的涂层表面缺陷识别方法[J]. 哈尔滨工程大学学报, 2022, 43(4): 572-579.
CHEN Z Y, ZHAO H, LYU Y S, et al. A recognition method of coating surface defects based on the improved MobileNetV2 network[J]. Journal of Harbin Engineering University, 2022, 43(4): 572-579.

[24] 陈亮, 邵玉斌, 龙华, 等. 基于时域 Gammatone 滤波特征的广播语种识别[J]. 信号处理, 2022, 38(3): 599-608.
CHEN L, SHAO Y B, LONG H, et al. Language identification for broadcasting signal based on time-domain gammatone filtering features[J]. Journal of Signal Processing, 2022, 38(3): 599-608.

[25] 曾金芳, 徐文涛, 黄费贞. 基于耳蜗倒谱系数的说话人识别[J]. 电子技术与软件工程, 2020, 5: 85-86.
ZENG JF, XU W T, HUANG F Z. Speaker recognition based on cochlear filter cepstral coefficients[J]. Electronic Technology and Software Engineering, 2020, 5: 85-86.

[作者简介]

龙华（1963– ），女，回族，云南大理人，博士，昆明理工大学教授，主要研究方向为无线网络及音频信号处理、语种识别等。



黄张衡（1997– ），男，彝族，云南曲靖人，昆明理工大学硕士生，主要研究方向为音频信号处理、语种识别等。

邵玉斌（1970– ），男，云南曲靖人，昆明理工大学教授，主要研究方向为移动通信和个人通信系统以及信号处理。

杜庆治（1977– ），男，云南楚雄人，昆明理工大学副教授，主要研究方向为语音信号处理、语种识别。

苏树盟（1996– ），男，云南保山人，昆明理工大学硕士生，主要研究方向为音频信号处理、语音识别。