doi：10.13413/j.cnki.jdxblxb.2023298

基于融合特征 ADRMFCC 的语音识别方法

朵 琳，马 建，韦贵香，唐 剑
（昆明理工大学 信息工程与自动化学院，昆明 650500）

摘要：针对在复杂噪声环境下语音识别准确率低和鲁棒性差的问题，提出一种基于增减残差 Mel 倒谱融合特征的语音识别方法. 该方法首先利用增减分量法筛选关键语音特征，然后将其映射到 Mel 域-残差域空间坐标系中生成增减残差 Mel 倒谱系数，最后将这些融合特征用于训练端到端模型. 实验结果表明，该方法在不同噪声类型和信噪比条件下均显著提高了语音识别准确率及性能，在－5 dB 低信噪比条件下，语音识别准确率达 73.13%，而在其他噪声条件下的平均语音识别准确率达 88.67%，充分证明了该方法的有效性和鲁棒性.

关键词：语音识别；残差 Mel 倒谱系数；特征筛选；增减分量法
中图分类号：TP391；TN912.3 **文献标志码**：A **文章编号**：1671-5489(2024)04-0943-08

Speech Recognition Method Based on Fusion Feature ADRMFCC

DUO Lin, MA Jian, WEI Guixiang, TANG Jian
(*Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming* 650500, *China*)

Abstract: Aiming at the problem of low accuracy and poor robustness of speech recognition in complex noise environment, we proposed a speech recognition method based on Mel cepstrum fusion feature of increasing and decreasing residuals. This method first used the increase and decrease component method to screen the key speech features, and then mapped them to the Mel domain-residual domain spatial coordinate system to generate the increase and decrease residual Mel cepstral coefficients. Finally, these fusion features were used to train the end-to-end model. The experimental results show that the proposed method significantly improves the accuracy and performance of speech recognition under different noise types and signal-to-noise ratio conditions. Under the low signal-to-noise ratio condition of －5 dB, the speech recognition accuracy reaches 73.13%, while the average speech recognition accuracy under other noise conditions reaches 88.67%, which fully proves the effectiveness and robustness of the proposed method.

Keywords: speech recognition; residual Mel cepstral coefficient; feature screening; increase and decrease component method

随着科技的发展和人工智能的兴起，语音识别技术已成为人们日常生活中不可或缺的一部分. 但传统的语音识别技术在噪声干扰、说话人变化等方面仍存在一定的局限性，因此如何提高语音识别准确率的研究备受关注. 特征提取是语音信号处理中的关键步骤，直接影响后续任务的效果.

收稿日期：2023-07-12.

第一作者简介：朵 琳(1974—)，女，彝族，博士，副教授，从事语音识别和信号处理的研究，E-mail：duolin2003@126.com.
通信作者简介：马 建(1998—)，男，汉族，硕士研究生，从事语音识别和信号处理的研究，E-mail：2703729898@qq.com.

基金项目：国家自然科学基金(批准号：61962032).

目前，主流的语音特征主要包括基于声学层特征和音素层特征，例如，Mel频率倒谱系数(Mel-scale frequency cepstral coefficients，MFCC)[1]，Gammatone频率倒谱系数(Gammatone frequency cepstral coefficients，GFCC)[2]和线性预测倒谱系数(linear predictive cepstral coefficients，LPCC)[3]等. 在嘈杂环境下，这些特征很容易受干扰，导致语音识别效果较差. 基于音素层的识别方法将语音信号分割成若干个音素单元，并将每个音素单元映射到对应音素库中的音素单元，得到一个表示整个语音信号的音素序列. 通过分析该音素序列的特征，例如音素出现的概率和音素之间的转移概率等，对整个语音信号进行识别. 相对于声学层特征，基于音素层特征的语音识别方法受噪声环境的影响较小，但由于音素的切分提取较困难，因此识别性能可能会下降.

随着深度学习被引入语音识别领域，Wang等[4]提出了将MFCC中Mel滤波器进行翻转得到翻转Mel倒谱系数(inverted Mel-frequency cepstral coefficients，IMFCC)特征，该特征可获取语音高频特征信息，结合MFCC特征以表征更全面的语音信息. Zhao等[5]提出了Fbank特征提取时基于滤波器组对音频进行滤波，可以捕获音频的重要信息，但Fbank特征只考虑了音频的频率分布信息，对其他音频的特征信息如时域和能量信息等未涉及，导致识别效果较差. 为克服MFCC和Fbank特征提取的缺点，本文提出在残差Mel倒谱系数(residual Mel-frequency cepstral coefficients，RMFCC)中引入残差信号[6]的概念，提取语音信号中不能被MFCC描述的残余信息，可有效提高语音识别的准确率. 此外，各种深度学习框架也被应用于语音识别任务，包括深度神经网络(deep neural network，DNN)[7]、长短期记忆神经网络(long short-term memory，LSTM)[8]、循环神经网络(recurrent neural network，RNN)[9]和双向循环神经网络(bidirectional recurrent neural network，BiRNN)[10]等神经网络模型.

近期，基于注意力机制的Transformer模型在各种语音识别任务中逐渐取代了传统的循环神经网络模型. 这是因为Transformer模型具有捕获长距离语音特征信息和高度并行训练的能力，而卷积神经网络(CNN)则擅长提取局部细粒度特征. 通过引入注意力机制，Transformer模型能同时处理整个输入序列，而不像RNN模型那样需要按顺序逐步计算. 这使得Transformer模型能高效地并行计算，从而显著加快了训练速度和推理速度. 在此基础上文献[11]提出了Conformer模型，该模型既能捕获长距离信息又能提取局部特征信息，在端到端语音识别任务中展现了优异的识别性能.

针对复杂噪声环境下的语音识别准确率低和鲁棒性差的问题，本文提出一种基于增减残差Mel倒谱系数(addition-deletion residual Mel-frequency cepstral coefficients，ADRMFCC)的语音识别方法. 该方法首先利用基于增减分量法的语音贡献度特征筛选方式对MFCC和RMFCC特征进行筛选，然后将特征映射在由Mel域-残差域组成的空间坐标系中以得到ADRMFCC，并将处理后的融合特征ADRMFCC送入Conformer-CTC端到端模型中进行识别训练. 实验结果表明，在不同的噪声种类和信噪比条件下，本文方法显著提高了语音识别性能.

1 特征提取

1.1 MFCC特征

MFCC是一种常用的语音信号处理特征提取方法. 在特征提取过程中，首先，将语音信号分帧，并对每帧进行加窗处理；其次，对每帧进行快速Fourier变换(FFT)，得到该帧语音信号的频谱；再次，使用一组Mel滤波器将频谱转换为Mel频率谱，并对Mel频率谱取对数运算，得到以dB为单位的对数谱；最后，对对数谱进行离散余弦变换，得到MFCC特征. 一般使用20～40个滤波器，得到20～40维度的特征向量. 在使用MFCC特征时需要对特征进行归一化处理，以保证不同特征维度的重要性相同. 第i帧第j维的MFCC为

$$\boldsymbol{M}_{\mathrm{MFCC}}(i,j)=\sqrt{\frac{2}{M}}\sum_{m=1}^{M}\lg[S_i(m)]\cos\left(\frac{j\pi(m-0.5)}{M}\right), \tag{1}$$

其中：$i=1,2,\cdots,I$为语音参数；$j=1,2,\cdots,J_m$，J_m为MFCC维度；M为滤波器数量；m为滤波器. 将$F\times J_m$维的MFCC特征矩阵表示为$\boldsymbol{M}$.

1.2　RMFCC 特征

残差 Mel 倒谱系数(RMFCC)是对 Mel 频率倒谱系数的一种改进. RMFCC 的计算方式与 MFCC 类似，但在计算 Mel 频率谱时，使用残差信号，即原始音频信号与线性预测编码(linear predictive coding，LPC)[12]预测信号的差. 计算步骤如下：

1) 对语音信号 $x(n)$分帧加窗，使用汉明窗，分帧加窗后的第 i 帧信号为 $x_i(n)$；

2) 对 $x_i(n)$进行离散 Fourier 变换，有

$$S_i(k)=\sum_{n=0}^{N-1}x_i(n)\mathrm{e}^{-j2\pi kn/N}, \tag{2}$$

其中 N 为离散 Fourier 变换的点数；

3) $S_i(k)$的功率谱密度为

$$P_i(k)=|S_i(k)|^2; \tag{3}$$

4) 对每帧的功率谱进行 LPC 分析，得到 LPC 系数，使用 LPC 系数对每帧音频信号进行线性预测编码，得到 LPC 预测信号为

$$\hat{x}(k)=\sum_{i=1}^{P}a_ix(k-i), \tag{4}$$

其中：p 为 LPC 的阶数，即 LPC 系数的数量；a_i 为估计得到的 LPC 系数，$i=1,2,\cdots,p$；

5) 将原始音频信号与 LPC 预测信号做差，得到残差信号为

$$r(k)=x(k)-\hat{x}(k); \tag{5}$$

6) 使用 Mel 滤波器组将残差信号转换为 Mel 频率谱

$$S_m(k,m)=\sum_{i=0}^{N-1}|R(k,i)|^2H_m(i); \tag{6}$$

7) 对 Mel 频率谱进行倒谱变换，得到 RMFCC 为

$$R_{\mathrm{MFCC}}(i,j)=\sqrt{\frac{2}{M}}\sum_{m=0}^{M}\log(S_m(k,m))\cos\left[\frac{j\pi}{M}\left(m-\frac{1}{2}\right)\right], \tag{7}$$

其中 M 是 Mel 滤波器数量，$S_m(k,m)$是第 k 帧残差信号经过第 m 个 Mel 滤波器的响应，j 为 RMFCC 系数阶数，将 $F\times R_r$ 维的 MFCC 特征矩阵表示为 $\boldsymbol{R}$.

1.3　基于增减分量法的融合特征 ADRMFCC

传统的特征融合方式是将单一的底层声学特征进行维度拼接，例如将 MFCC 和 RMFCC 拼接在一起，得到一个维度为 $F\times(J_m+R_r)$的融合特征矩阵：

$$\boldsymbol{X}=((\boldsymbol{M}_1,\boldsymbol{M}_2,\cdots,\boldsymbol{M}_{J_m}),(\boldsymbol{R}_1,\boldsymbol{R}_2,\cdots,\boldsymbol{R}_{R_r})), \tag{8}$$

其中 $\boldsymbol{M}_1$ 和 $\boldsymbol{R}_1$ 分别为第一维 MFCC 和 RMFCC. 虽然这种方式能表征两种声学特征所包含的不同信息，但无法将它们之间的关联关系考虑在内. 为解决该问题，可将相同维度的 MFCC 和 RMFCC 以线性相加的方式进行特征融合，即得到一个维度为 $F\times J_m$ 的融合特征矩阵：

$$\boldsymbol{X}=\boldsymbol{M}+\boldsymbol{R}. \tag{9}$$

这种方式可以增强特征之间的关联，但融合后的特征矩阵维度较高，会增加模型训练和计算的复杂度，同时也可能存在冗余信息，并且在噪声环境下，随着信噪比的降低，语音的声学特征会被破坏，因此仍无法达到理想的语音识别效果. 为解决上述问题，本文提出一种基于增减分量法语音贡献度的特征筛选方式，去除不必要的维度成分，得到 MFCC 和 RMFCC 中含有对语音识别有贡献度的特征维度. 增减分量法的平均贡献度函数如下：

$$G_i=\frac{1}{K}\left[\sum_{i<j}(p(i,j)-p(i+1,j))+\sum_{i>j}(p(i,j)-p(i-1,j))\right], \tag{10}$$

其中 G_i 表示贡献度，$p(i,j)$表示第 i 维到第 j 维特征作为语音特征参数时的识别准确率. 本文实验首先检测特征参数每个维数 $i\sim j$ 组合的识别率，然后由式(10)计算每个维度的贡献度.

由于简单的特征域维度叠加和线性相加并不能更好地发挥两种特征的抗噪性能，本文提出将 Mel 域和残差域分别作为新的语音特征空间的横轴和纵轴，并在保证 MFCC,GFCC 特征在帧长、帧移一致

的情况下，将两种特征进行矩阵乘运算得到融合特征 ADRMFCC，简化后的 ADRMFCC 为

$$x_{ij}=\sum_{t=1}^{F}M_{it}R_{tj}=\sum_{t=1}^{F}M_{it}\sqrt{\frac{2}{M}}\sum_{m=1}^{M}\lg[S_i(m)]\cos\left(\frac{j\pi(m-0.5)}{M}\right), \tag{11}$$

其中：M_{it}为第 t 帧第 i 维 MFCC；R_{tj}为第 t 帧第 j 维 RMFCC；x_{ij}为两种特征中某一维度不同语音特征的加权和，数值越大，二者关系越大.

2 基于 Conformer-CTC 语音识别模型

为实现更好的语音识别模型，本文采用链接时序分类(connectionist temporal classification，CTC)作为解码器，构建 Conformer-CTC 编码解码模型. 图 1 为该模型架构.

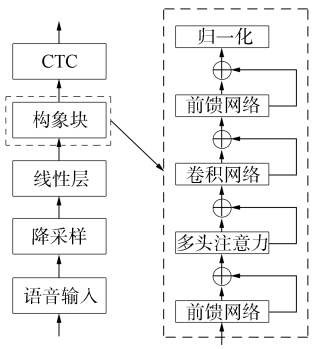

图 1 Conformer-CTC 编码解码模型架构

Fig. 1 Architecture of Conformer-CTC encoding and decoding model

Conformer 模型是一种序列建模架构，它融合了多个关键组件，并通过残差连接实现它们之间的连接. 这些组件包括多头注意力模块、卷积网络模块和前馈网络模块.

多头注意力模块使用类似于 Transformer-XL 的方法计算序列中的位置编码信息，可有效捕捉输入语音特征序列中的重要语音特征信息. 卷积网络模块由逐点卷积网络、ReLU 激活函数和一维深度卷积网络组成，它能有效捕捉输入特征序列中的局部细节语音特征信息. 前馈网络模块在 Conformer 模型中扮演重要角色，它由两个线性变换层和 Swish 激活函数构成，该模块引入了非线性变换，可更好地捕捉输入特征的复杂关系. Conformer 模型借鉴了 Macaron-Net 网络结构的思想，将前馈网络模块分别放置在多头注意力模块之前和卷积网络模块之后. 这种设计使模型可充分利用多头注意力模块对全局上下文的建模能力，以及卷积网络模块对局部细节的建模能力. 同时，通过在各模块之间添加残差连接，有助于信息的传递并减轻梯度消失问题.

该过程首先对输入的语音信号进行特征提取，并对其进行降采样处理，使用多个构象块(conformer blocks，CB)建立编码器部分. 每个 CB 包含自注意力层、前馈神经网络层和卷积层，用于捕捉输入序列的上下文信息和特征表示. 在编码器之后添加一个 CTC 层，将编码器的输出映射到字符序列. CTC 层使用 CTC 损失函数训练模型，无需对齐标签，可处理不定长输入和输出序列. 在训练过程中，使用 CTC 解码器对 CTC 层的输出进行解码，得到最终的识别结果.

3 实验及结果分析

3.1 实验设计

利用 PyCharm 进行仿真实验，使用的软件为 TensorFlow1.15 版，Window10 操作系统，12 GB 内存，处理器为 Intel-i5-12400F. 本文使用的实验数据来自中文数据集 THCS30. 数据集 THCHS30 总持续时间超过 30 h，采样频率为 16 kHz，采样大小为 16 bit. 训练集包含 10 000 条语音数据. 表 1 列出了中文语音数据集 THCHS30 的信息.

表 1 中文语音数据集 THCHS30 的信息

Table 1 Information of THCHS30 Chinese speech dataset

数据集	语音时长/min	句子数	词数
训练集	25	10 000	198 252
验证集	134	893	17 743
测试集	375	2 459	49 085

针对复杂噪声环境下的语音识别，本文实验训练集采用公开噪声数据库 NOISEX-92 中的白噪声

作为背景噪声，每种语音包含的 SNR 等级为[5 dB,10 dB,15 dB,20 dB,25 dB]的语音各 1 000 条和 500 条未加噪语音. 测试集采用 NOISEX-92 中 7 种不同的噪声源：buccaneer1,destroyerops,f16,hfchannel,pink,volvo,white，模拟真实环境下不同的噪声环境. 每种语料库包含的 SNR 等级为[−5 dB,0,5 dB,10 dB,15 dB]的音频，从而构成 35 个测试数据库. 平均信噪比为

$$\mathrm{SNR}=10\lg\frac{\sum_{n=1}^{H}s^2(n)}{\sum_{n=1}^{H}w^2(n)},\tag{12}$$

其中 $\sum_{n=1}^{H}s^2(n)$ 表示语音信号能量总和，H 表示语音的总采样点数，$\sum_{n=1}^{H}w^2(n)$ 表示噪声信号能量总和. 语音识别性能评价指标为

$$\mathrm{WER}=\frac{S+D+I}{N}\times 100\%,\tag{13}$$

其中 S 表示替换，D 表示删除，I 表示插入，N 表示词数目，WER 表示错误率.

3.2　实验参数选取

语音的帧长和帧移是对语音识别性能有重要影响的关键参数. 本文选择 39 维的 MFCC 特征和 24 维的 RMFCC 特征，并在以 10 dB 的白噪声为背景噪声的数据集 THCHS30 中进行实验，以验证不同帧长和帧移对语音识别性能的影响. 表 2 列出了不同帧长和帧移下的语音识别准确率.

表 2　不同帧长和帧移下的语音识别准确率

Table 2　Speech recognition accuracy under different frame lengths and frame shifts

帧长	帧移	特征矩阵	准确率/%	特征矩阵	准确率/%
256	128	$\boldsymbol{M}_{39\times976}$(38 064)	50.86	$\boldsymbol{R}_{976\times72}$(38 064)	61.56
512	128	$\boldsymbol{M}_{39\times974}$(37 986)	52.56	$\boldsymbol{R}_{974\times72}$(37 986)	64.31
512	256	$\boldsymbol{M}_{39\times488}$(19 032)	53.85	$\boldsymbol{R}_{488\times72}$(19 032)	67.24
1 024	256	$\boldsymbol{M}_{39\times486}$(18 954)	55.27	$\boldsymbol{R}_{486\times72}$(18 954)	67.56
1 024	512	$\boldsymbol{M}_{39\times244}$(9 516)	57.98	$\boldsymbol{R}_{244\times72}$(9 516)	68.48
2 048	512	$\boldsymbol{M}_{39\times242}$(9 438)	54.93	$\boldsymbol{R}_{242\times72}$(9 438)	68.15
2 048	1 024	$\boldsymbol{M}_{39\times122}$(4 758)	53.21	$\boldsymbol{R}_{122\times72}$(4 758)	66.72

表 2 由 13 维的静态 MFCC 特征及其 1 阶、2 阶动态差分参数组成，语音帧数为 976. 此外，随着帧长和帧移的增加，特征的识别准确率呈现先增加后降低的趋势. 当帧长和帧移分别为 1 024 和 512 时，两种特征的识别准确率最高，分别为 57.98%和 68.48%. 实验结果表明，在噪声环境下，RMFCC 能更好地表征语音特征，从而提高语音识别的准确性. RMFCC 通过引入残差信息，可捕捉到语音信号中的细微变化和动态特征，对在噪声环境下更稳定地表示语音有益. 而传统的 MFCC 只考虑静态特征，对噪声环境下的语音识别可能会受到干扰.

本文语音识别模型选用 CTC 损失函数度量真实标签与预测标签的差值，CTC 损失函数能处理输入序列和输出序列长度不一致的情况，它通过对齐和计算两个序列之间的差异训练模型. 选用 Adam 优化函数加速模型收敛，并在学习率设为 0.001，迭代次数为 200 时，模型具有较好的收敛效果.

3.3　基于语音识别贡献度 ADRMFCC 选取

本文采用多次实验取均值的形式，将 39 维的 MFCC 特征和 24 维的 RMFCC 在以 5 dB 的白噪声为背景噪声的数据集 THCHS30 及不同模型中进行实验. 图 2 为 MFCC 和 RMFCC 各维度贡献度. 由图 2 可见，39 维 MFCC 和 24 维 RMFCC 特征在不同维度下的贡献度呈下降趋势. 表明增加特征的维度并不一定会提升语音识别性能. 基于此，本文提出两种特征筛选方式.

方式 1：由图 2 可见，当 MFCC 特征在第 27 维时，贡献度快速下降，因此选取前 26 维特征作为待融合 MFCC 特征(eliminate dimensions-MFCC，ED-MFCC)；同理，当 RMFCC 特征在第 16 维时，贡献度快速下降，因此选取前 15 维特征作为待融合 RMFCC 特征(eliminae dimensions-RMFCC，ED-RMFCC).

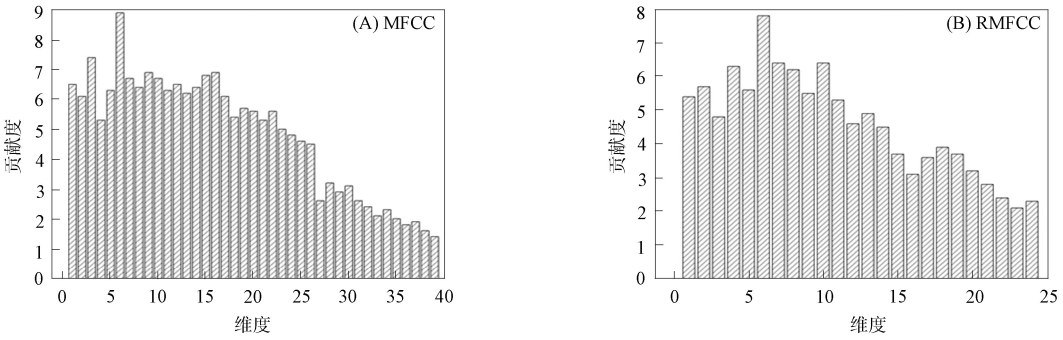

图 2 MFCC 和 RMFCC 各维度贡献度

Fig. 2 Contribution of each dimension of MFCC and RMFCC

方式 2：以贡献度大小依次排序选取特征，将 39 维 MFCC 特征贡献度从大到小排序后选取(6,3,9,16,15,7,10,1,12,8,14,5,11,13,2,17,19,20,22,18,4,21,23,4,24,25)共 26 维特征作为待融合特征；将 24 维 RMFCC 特征贡献度从大到小排序后选取(6,7,10,4,8,2,5,9,1,11,13,3,12,14,18)共 15 维作为待融合特征.

表 3 列出了不同模型下两种方式的识别准确率. 由表 3 可见，在使用数据集 THCHS30 进行实验时，以 5 dB 的白噪声作为背景噪声，Conformer-CTC 作为系统模型时，方式 1 的语音识别准确率为 89.56%，方式 2 的语音识别准确率为 91.23%. 实验结果表明，方式 2 中按照贡献度大小排序后的筛选方式能更好地获取有意义的特征，因此具有更高的语音识别准确率.

表 3 不同模型下两种方式的识别准确率

Table 3 Recognition accuracy of two methods under different models

系统模型	WER/%		系统模型	WER/%	
	方式 1	方式 2		方式 1	方式 2
GMM-HMM	71.68	74.18	BiLSTM-CTC/Attention	81.23	83.46
DNN-HMM	76.41	79.64	Transformer-CTC	85.71	88.14
CNN-HMM	77.49	80.43	Conformer-CTC	89.56	91.23

3.4 测试不同噪声下的语音识别性能

为测试 7 种不同复杂噪声环境下本文方法的有效性和鲁棒性，并分析其优劣原因，设计下列 6 组实验，实验结果列于表 4.

表 4 噪声状态下的语音识别准确率

Table 4 Speech recognition accuracy under noisy conditions

噪声类型	特征	准确率/%				
		15 dB	10 dB	5 dB	0	−5 dB
WN	Fbank	72.31	63.97	58.46	46.35	32.63
	MFCC	66.59	57.98	45.86	32.41	30.67
	RMFCC	75.48	68.48	60.09	54.37	48.56
	ED-MFCC	77.35	69.48	64.23	51.42	44.09
	ED-RMFCC	89.26	74.85	73.74	58.41	54.71
	ADRMFCC	96.45	94.11	91.23	89.41	74.59
VN	ADRMFCC	97.13	96.78	95.45	91.26	89.78
PN	ADRMFCC	96.35	94.26	93.78	90.86	79.48
BFCN	ADRMFCC	96.56	95.41	94.68	90.81	73.13
HFCN	ADRMFCC	95.16	94.84	90.23	88.47	74.23
F16	ADRMFCC	95.78	93.18	91.56	88.47	75.79
DORBN	ADRMFCC	96.47	94.03	93.26	89.87	78.48

实验 1. 提取 200 维 Fbank 特征[13]作为语音特征.

实验 2. 提取 39 维 MFCC 特征[14]作为语音特征.

实验 3. 提取 24 维 RMFCC 特征[15]作为语音特征.

实验 4. 提取 26 维 ED-MFCC 特征作为语音特征.

实验 5. 提取 15 维 ED-RMFCC 特征作为语音特征.

实验 6. 提取融合特征 ADRMFCC 特征作为语音特征.

由表 4 可见：在复杂噪声环境中，随着噪声信噪比的降低，语音信号逐渐被淹没，导致语音识别准确率逐渐下降；对比 ADRMFCC 特征在 7 种不同复杂噪声环境下的识别性能表明，VN 噪声环境下的语音识别准确率均高于其他 6 种噪声，且在－5 dB 信噪比下仍达到 89.78%的识别准确率. 这是因为 VN 噪声为车内噪声，其频率区间在 300 Hz 以下，而人声的主要频率区间在 300～3 400 Hz，故当 VN 噪声叠加到语音信号上时，对语音信号的破坏相对低于其他噪声.

图 3 为不同特征在白噪声不同信噪比下的语音识别性能. 由图 3 可见，ADRMFCC 特征在各信噪比条件下的准确率均高于其他特征. 对比实验 1～3，在－5 dB 信噪比下，39 维的 MFCC 特征准确率最低，仅为 30.67%，这是由于 MFCC 特征对人类听觉系统的感知特征进行了模拟，能很好地表示语音信号的重要频率成分，因此在较清晰的语音环境下性能较好；而实验 1 中 Fbank 特征使用的滤波器数量较多，因此能更好地表示高频和低频信息，具有一定的噪声鲁棒性，相比于 MFCC 特征在 5 种不同信噪比下分别提了 5.72，5.99，12.90，13.97，1.96 个百分点；由于实验 3 中 RMFCC 特征使用了 LPC 预测信号和残差信号，能更好地抑制噪声，因此在高噪声环境下表现出很好的鲁棒性，在 5 种不同信噪比条件下相比于 MFCC 和 Fbank 特征分别提高了 18.90，20.50，14.23，21.96，17.89 个百分点和 13.17，14.51，1.63，8.02，15.93 个百分点.

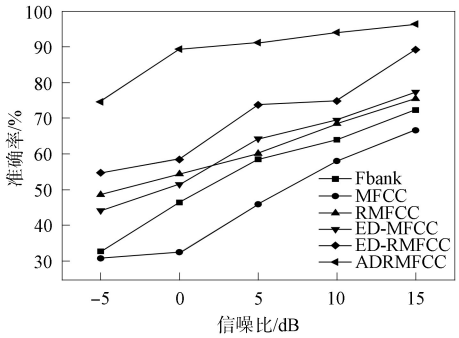

图 3　不同特征在不同信噪比下的语音识别准确率

Fig. 3　Speech recognition accuracy of different features under different signal-to-noise ratios

对比实验 2 和实验 4 可知，39 维的 MFCC 特征中可能包含一些对语音识别意义较小的特征，因此采用增减分量法筛选具有语音贡献度的特征，并从大到小排序提取 26 维的 ED-MFCC 特征，能较好地对特征进行处理，去除不必要的维度成分，减少参数冗余. 在 5 种不同信噪比条件下相比于 ED-MFCC 和 MFCC 特征分别提高了 10.76，11.50，18.37，19.01，13.42 个百分点. 同理，由实验 3 和实验 5 可知，在 5 种不同信噪比条件下相比于 ED-RMFCC 和 RMFCC 特征分别提高了 13.78，6.37，13.65，4.04，6.15 个百分点. 实验 6 中的 ADRMFCC 特征在 5 种信噪比下的识别性能均高于其他 5 种特征性能，相比于 26 维的 ED-MFCC 特征在 WN 噪声下识别准确率提高了 19.10，24.63，27.00，37.99，30.50 个百分点；相比于 15 维的 ED-RMFCC 特征识别准确率提高了 7.19，19.26，17.49，31.00，15.19 个百分点. 实验结果表明，本文针对复杂噪声环境下的语音识别方法具有较好的鲁棒性和识别性能.

图 4 为 7 种不同噪声源下，采用 ADRMFCC 特征和 ED-MFCC，ED-RMFCC 特征的平均识别准确率. 由图 4 可见，在 7 种不同噪声源下，采用 ADRMFCC 特征相对于 ED-MFCC，ED-RMFCC 特征在平均识别准确率上均有提升. 除车内噪声源 VN 外，其他噪声源下语音识别准确率显著提高. 这是因为 VN 属于低频噪声，车内噪声能量主要由其低频部分决定，因此在 VN 源下语音识别准确率提升并不明显. 可见，本文的 ADRMFCC 特征方法

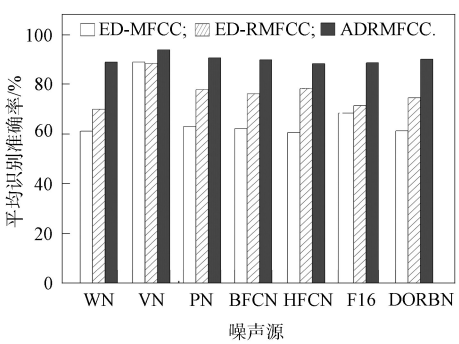

图 4　不同复杂噪声源下的平均识别准确率

Fig. 4　Average recognition accuracy under different complex noise sources

可极大提高在复杂噪声环境下的语音识别准确率，并具有很高的鲁棒性.

综上，针对复杂噪声环境下语音识别准确率低和鲁棒性差的问题，本文提出了一种ADRMFCC的语音识别方法. 该方法先利用基于增减分量法语音贡献度的特征筛选方式对MFCC和RMFCC特征进行筛选，然后将筛选后的特征映射在由Mel域-残差域组成的空间坐标系中以得到增减残差Mel倒谱系数(ADRMFCC)，最后将处理好的融合特征ADRMFCC送入Conformer-CTC端到端模型中进行识别训练. 实验结果表明，本文方法在7种不同噪声源下的语音识别准确率均有提升，且鲁棒性也有增强，因此该方法适用于复杂噪声环境下的语音识别.

参 考 文 献

[1] BISWAS M, RAHAMAN S, AHMADIAN A, et al. Automatic Spoken Language Identification Using MFCC Based Time Series features [J]. Multimedia Tools and Applications, 2023, 82(7): 9565-9595.

[2] CHANDRASEKARAM B. New Feature Vector Based on GFCC for Language Recognition [J]. Journal of Algebraic Statistics, 2022, 13(2): 481-486.

[3] FAÚNDEZ-ZANUY M. Speaker Recognition by Means of a Combination of Linear and Nonlinear Predictive Models [EB/OL]. (2022-05-07)[2023-02-01]. https://arxiv.org/abs/2203.03190.

[4] WANG Z Q, YAN J H, WANG Y F, et al. Speech Emotion Feature Extraction Method Based on Improved MFCC and IMFCC Fusion Features [C]//2023 IEEE 2nd International Conference on Electrical Engineering, Big Data and Algorithms (EEBDA). Piscataway, NJ: IEEE, 2023: 1917-1924.

[5] ZHAO J K, ZHOU H P, LIU H L, et al. Feature Fusion Method for Speaker Recognition Based on Embedding Mechanism [C]//International Conference on Signal Processing and Communication Security (ICSPCS 2022). [S.l.]: SPIE, 2022: 108-113.

[6] SIDDHARTHA S, MISHRA J, PRASANNA S R M. Language Specific Information from LP Residual Signal Using Linear Sub-band Filters [C]//2020 National Conference on Communications (NCC). Piscataway, NJ: IEEE, 2020: 1-5.

[7] WANG D, WANG X D, LÜ S H. An Overview of End-to-End Automatic Speech Recognition [J]. Symmetry, 2019, 11(8): 1018-1044.

[8] ZHAO J F, MAO X, CHEN L J. Speech Emotion Recognition Using Deep 1D & 2D CNN LSTM Networks [J]. Biomedical Signal Processing and Control, 2019, 47: 312-323.

[9] SHEWALKAR A, NYAVANANDI D, LUDWIG S A. Performance Evaluation of Deep Neural Networks Applied to Speech Recognition: RNN, LSTM and GRU [J]. Journal of Artificial Intelligence and Soft Computing Research, 2019, 9(4): 235-245.

[10] HE M H. Application of Bidirectional Recurrent Neural Network in Speech Recognition [J]. Computer and Modernization, 2019(10): 1-6.

[11] ZHANG Y, PUVVADA K C, LAVRUKHIN V, et al. Conformer-Based Target-Speaker Automatic Speech Recognition for Single-Channel Audio [C]//2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2023: 1-5.

[12] DAVE N. Feature Extraction Methods LPC, PLP and MFCC in Speech Recognition [J]. International Journal for Advance Research in Engineering and Technology, 2013, 1(6): 1-4.

[13] DO C T. End-to-End Speech Recognition with High-Frame-Rate Features Extraction [EB/OL]. (2019-06-03)[2023-01-15]. https://arxiv.org/abs/1907.01957.

[14] GARG U, AGARWAL S, GUPTA S, et al. Prediction of Emotions from the Audio Speech Signals Using MFCC, MEL and Chroma [C]//2020 12th International Conference on Computational Intelligence and Communication Networks (CICN). Piscataway, NJ: IEEE, 2020: 87-91.

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

(责任编辑：韩 啸)