文章编号：2095-2163(2021)10-0061-07 中图分类号：TP183 文献标志码：J

基于 MRACC 特征的鲁棒说话人识别研究

崔 潇，夏秀渝
（四川大学 电子信息学院，成都 610065）

摘 要：为提高噪声环境下说话人识别系统的抗噪性能，提出一种基于 MRACC 特征和 LSTM 网络的鲁棒说话人识别方法。首先采用一种动态调整参数的改进型谱减法进行语音前端降噪处理，接着提取改进的多分辨率听觉倒谱系数特征（Multi-Resolution Auditory Cepstral Coefficient，MRACC），特征提取时采用幂函数代替对数函数模拟人耳的非线性压缩特性。最后将提取的特征参数送入基于长短时记忆网络（Long Short Term Memory，LSTM）的说话人模型进行识别。仿真实验表明，在低信噪比情况下，前端语音降噪处理能有效提高系统的识别性能。在说话人识别系统中 MRACC 特征的识别性能优于传统的 MFCC（Mel Frequency Cepstral Coefficient，MFCC）和 LPCC（linear predictive cepstrum coefficient，LPCC）特征，并且具有一定的鲁棒性。

关键词：MRACC 特征；LSTM 网络；说话人识别；谱减法；鲁棒性

Research on robust speaker recognition based on MRACC features

CUI Xiao，XIA Xiuyu
（College of Electronics and Information Engineering，Sichuan University，Chengdu 610065，China）

【Abstract】In order to improve the anti-noise performance of speaker recognition system in noisy environment，a robust speaker recognition method based on MRACC features and LSTM network is proposed. Firstly，an improved type spectrum subtraction method with dynamically adjusted parameters is used for speech front-end denoising，then improved MRACC is extracted. In feature extraction，power function is used instead of logarithmic function to simulate the nonlinear compression characteristics of human ear. Finally，the extracted feature parameters are fed into the speaker model based on LSTM network for recognition. The simulation results show that the front-end speech noise reduction can effectively improve the recognition performance of the system under the condition of low SNR. The recognition performance of MRACC features in speaker recognition system is better than traditional MFCC and LPCC features，and has a certain degree of robustness.

【Key words】MRACC feature；LSTM network；speaker recognition；spectral subtraction；robustness

0 引 言

近年来，说话人识别的研究获得了迅速发展，在军事、信息安全和通信等领域都有广泛应用[1]。在现实环境中，由于噪声的存在，导致说话人系统的识别率显著下降。因此如何提高系统在噪声环境下的识别性能成为研究热点。

在特征提取方面，常用的说话人识别特征有梅尔频率倒谱系数（Mel Frequency Cepstral Coefficient，MFCC）[2]、线性预测倒谱系数（Linear Prediction Cepstral，LPCC）[3]等。通过对 Gammatone 滤波器的研究，Chen 等人[4]提出多分辨率耳蜗图（Multi-resolution cochleagram，MRCG），相较于 MFCC 和 LPCC 特征，则具有更好的抗噪性能。

随着机器学习算法的发展，神经网络被用于说话人识别中，杨瑶等人[5]使用误差反向传播（Back Propagation，BP）网络进行文本无关的说话人识别研究。盖晁旭[6]利用稀疏（Sparse Autoencoder，SA）提取说话人特征进行说话人识别，循环神经网络（Recurrent Neural Networks，RNN）具有记忆功能，处理时序数据的能力较强。目前，Hochreiter 和 Schmidhuber 提出的长短时记忆网络[7]（Long Short Term Memory，LSTM）是应用最广泛的 RNN 网络之一。LSTM 网络在处理时间范围较大的信息时具有更好的性能，被用于语种识别[8]、语音识别[9]、音素分类[10]、唇语识别[11]等多个领域中。

本文构建了基于 MRACC 特征的说话人识别系统。在噪声环境下，首先使用文中提出的改进型谱减法完成语音的预降噪处理，接着提取基于多分辨率耳蜗图的 MRACC 特征，最后将特征参数输入到

作者简介：崔 潇（1996-），女，硕士研究生，主要研究方向：语音信号处理；夏秀渝（1970-），女，博士，副教授，硕士生导师，主要研究方向：语音信号处理。
通讯作者：夏秀渝 Email：xiaxxy@163.com

收稿日期：2021-08-02

LSTM 网络中实现模型的训练与匹配,通过实验验证了本文提出的说话人识别方法的有效性。

1 说话人识别系统

本文的说话人识别系统框图如图 1 所示。主要包含预处理、特征提取、模型训练、模型匹配和决策判决五个模块。在训练过程中,首先对训练集语音进行特征参数提取,然后利用训练集的特征参数通过模型训练得到说话人模型库。在测试过程中,则对测试集语音信号进行预处理操作,同样提取说话人的特征参数,再通过比对输出概率进行判决,输出概率最高者即为识别的说话人。

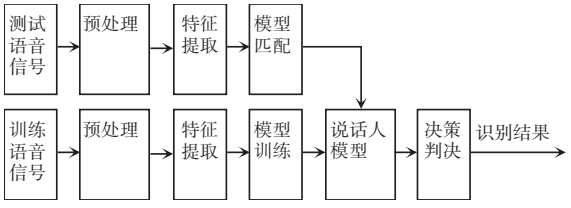

图 1 说话人识别系统框图

Fig. 1 Block diagram of speaker recognition system

1.1 预处理

针对复杂噪声环境,对语音信号进行的预处理包括预加重、前端降噪处理等。

1.1.1 预加重

预加重的目的是提升语音的高频分量。通常使用一阶高通滤波器实现预加重技术, 其传递函数可表示为:

$$H(z) = 1 - a z^{-1} \tag{1}$$

其中, a 为预加重系数。通常, $0.9 < a < 1.0$, 实验中 a 取 0.97。

1.1.2 前端降噪处理

为了解决噪声较大时,说话人系统的识别准确率较低的问题,本文先对含噪语音信号采用前端降噪处理提高识别率。谱减法是最常用的降噪方法,发展较为成熟,简单容易实现。谱减法的基本原理是在假设噪声是统计平稳的前提下估计噪声的频谱值,与含噪语音的频谱值相减,得到纯净语音的频谱估计值。但实际噪声往往是随机非平稳的,语音降噪后常产生“音乐噪声”,针对上述问题,本文提出了一种根据信噪比动态调整参数的改进型谱减法。算法主要步骤如下:

(1)能熵比算法端点检测。在噪声环境下,能熵比法具有较好的端点检测效果,在其计算过程中和谱减法有共用的部分,组合使用运算量小。能熵比的数学定义如下:

$$EEF_i = \sqrt{1 + |AMP_i / H_i|} \tag{2}$$

其中, AMP_i 表示第 i 帧语音的能量, H_i 表示第 i 帧语音的谱熵。

(2)噪声谱实时更新。实际噪声往往是非平稳的,文中采用滑动平均的方法对非语音段内的噪声谱进行实时更新。在静音段,为了得到较小的谱估计方差,对第 i 帧频谱进行如下平滑处理:

$$D(k) = \frac{1}{2M + 1} \sum_{j=-M}^{M} Y_{i+j}(k) \tag{3}$$

其中, $Y_i(k)$ 表示第 i 帧第 k 条谱线的谱值。本文中 $M = 1$, 即在计算 3 帧的平均值。

(3)动态调整参数。根据带噪语音的信噪比动态调整改进谱减法的一组参数,实现抑制噪声和语音失真的折中。改进型谱减法定义如式(4)所示:

$$\hat{S}_i(k) = \begin{cases} (|Y_i(k)|^{\lambda} - \alpha \times |D(k)|^{\lambda})^{\frac{1}{\lambda}} & |Y_i(k)|^{\lambda} \geqslant \alpha \times |D(k)|^{\lambda} \\ (\beta \times |Y_i(k)|^{\lambda})^{\frac{1}{\lambda}} & |Y_i(k)|^{\lambda} < \alpha \times |D(k)|^{\lambda} \end{cases} \tag{4}$$

其中, α 为过减因子, β 为增益补偿因子。同时这里引入了参数 λ。

参数 α 和 λ 值的大小会影响去噪程度, α 和 λ 的值越大,噪声去除得越多,音乐噪声越小,但语音失真也越厉害;增益补偿因子 β 值过大会带来噪声残留,过小会产生“音乐噪声”。

为取得噪声抑制和语音失真之间的平衡,采取根据信噪比动态调整参数 α、β、λ 的方法:

$$\alpha = \begin{cases} 6 & SNR \leqslant -5 \\ 5 - \dfrac{SNR}{5} & -5 < SNR \leqslant 20 \\ 1 & SNR > 20 \end{cases} \tag{5}$$

$$\beta = \begin{cases} 0.05 & SNR \leqslant -5 \\ 0.05 - 0.05(SNR + 4) & -5 < SNR \leqslant 5 \\ 0.001 & SNR > 5 \end{cases} \tag{6}$$

$$\lambda = \frac{1}{1 + e^{(-\sigma(SNR - \tau))}} + 1 \tag{7}$$

其中, σ 是一个控制曲线陡峭程度的参数, τ 是偏差参数。通过实验选取最优参数值为: $\sigma = 0.9$, $\tau = 15$。SNR 为每帧语音的短时信噪比,是一种后验信噪比,计算如下:

$$SNR = 10 \lg \frac{\sum_{n=0}^{N-1} s^2(n)}{\sum_{n=0}^{N-1} d^2(n)} \tag{8}$$

其中，$\sum_{n=0}^{N-1} s^2(n)$ 表示带噪信号的能量，$\sum_{n=0}^{N-1} d^2(n)$ 表示估计的噪声平均能量。

1.2　特征提取

特征提取是说话人识别系统中重要的部分，常用的语音特征包括 MFCC、LPCC 和 MRCG 等。其中，MFCC 利用基于听觉模型的 Mel 滤波器组进行提取，是一种最常用的语音特征参数，而 LPCC 参数是基于声道模型理论，通过线性预测分析得到的一种语音特征参数。说话人识别应用中，LPCC 在纯净语音环境下识别效果较好；相较于 LPCC，MFCC 对噪声环境有一定的鲁棒性，但在低信噪比环境下的识别率仍然较低。MRCG 特征采用 Gammatone 滤波器组模拟人耳听觉模型，有效提取了多分辨率的 cochleagram。该特征既关注了细节性的高分辨率特征，又可把握全局性的低分辨率特征，具有一定的鲁棒性。但都是通过采用对数函数对语音能量进行压缩来模拟人耳对语音强度感知的非线性特性，对数压缩在高音段可以很好地模拟人耳听觉特性，却在低音段会产生较大误差。尤其是在含噪情况下，当噪声较小时，对数压缩会扩大小信号的影响，不利于进行说话人识别。另一方面，MRCG 特征维数较大，计算复杂度高。基于前文分析，本次研究对 MRCG 特征进行了改进，提取一种改进的语音特征参数-多分辨率听觉倒谱系数 MRACC 特征。MRACC 特征提取过程如图 2 所示。

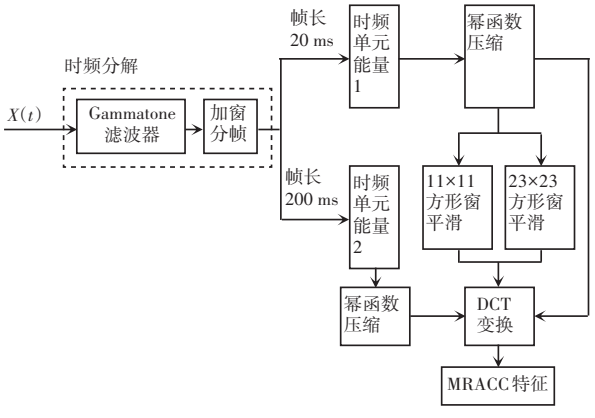

图 2　MRACC 特征提取过程图

Fig. 2　MRACC feature extraction process diagram

MRACC 特征参数提取步骤如下：

(1)时频分解。输入信号 $x(t)$ 经过 Gammatone 滤波器后分解为 64 个子带信号 $G(t, f_c)$，对应公式为：

$$G(t, f_c) = g(t, f_c) \cdot U(t) \cdot x(t) \tag{9}$$

其中，$U(t)$ 表示单位阶跃函数，$g(t, f_c)$ 为 Gammatone 滤波器的频率响应。

对子带信号 $G(t, f_c)$ 进行加窗分帧，得到时频分解表达式 $y_i(t, f_c)$，对应公式为：

$$y_i(t, f_c) = w(t) * G(((i-1) * inc + t), f_c) \tag{10}$$

其中，$w(t)$ 为窗函数，本文选择汉明窗；inc 为帧移，设置为 10 ms；帧长设置为 20 ms。

对时频单元提取听觉能量(cochlea gram)，得到第 i 帧中心频率为 f_c 的时频单元的听觉能量。计算公式为：

$$GF_1(i, f_c) = \sum_{t=0}^{L_1-1} y_i^2(t, f_c) \tag{11}$$

其中，$y_i(t, f_c)$ 表示第 i 帧中心频率为 f_c 的子带信号。

(2)幂函数压缩。原始 MRCG 特征使用对数函数对听觉能量进行非线性压缩处理，对数压缩会扩大语音中小信号的影响。当噪声较小时，同样也对噪声进行了放大，所以针对存在噪声的情形，考虑改进非线性压缩方式。通过研究发现，基于强度-响度感知的幂函数可以代替对数函数更好地模拟人耳对各个音强段的感知特性。听觉能量 GF_1 经过幂函数处理，得到耳蜗图 (CG_1)，如式(12)所示：

$$CG_1(i, f_c) = \sqrt[n]{CF_1(i, f_c)} \tag{12}$$

实验表明，$n = 15$ 可以很好地模拟人耳感知音强的非线性特性。

对数函数压缩的函数曲线与幂函数压缩的函数曲线如图 3 所示。由图 3 可以看出，对数函数与幂函数相比，对小信号放大更多，抑制噪声的性能会更差一些。因此本文提取特征时，选择了幂函数代替对数函数对听觉能量进行非线性压缩。

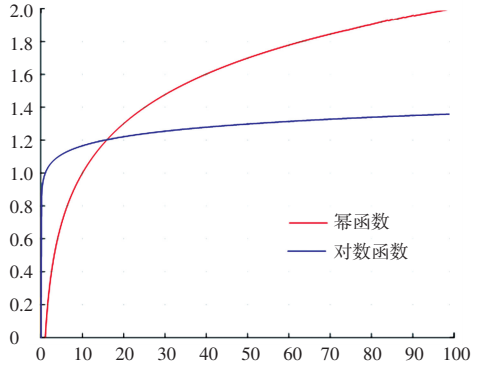

图 3　非线性压缩函数图

Fig. 3　Non-linear compression function graph

(3)多分辨率耳蜗图特征提取。文中，将帧长

改为200 ms,计算时频单元 2 的听觉能量 GF_2,并进行幂函数压缩得到耳蜗图 CG_2,表达式为:

$$CG_2(i,f_c)=\sqrt[15]{\sum_{t=0}^{L_2-1}y_i^2(t,f_c)} \tag{13}$$

取长为 11 帧,宽为 11 个子带的方形窗对 CG_1 进行平滑,得到耳蜗图(CG_3),计算公式为:

$$CG_3(i,f_c)=\sum_{c=c-5}^{c+5}\sum_{i=i-5}^{i+5}\frac{sum(sum(CG_1(i,f_c)))}{11*11} \tag{14}$$

和 CG_3 计算相似,使用长为 23 帧,宽为 23 个子带的方形窗对 CG_1 进行平滑,得到耳蜗图(CG_4),计算公式为:

$$CG_4(i,f_c)=\sum_{c=c-11}^{c+11}\sum_{i=i-11}^{i+11}\frac{sum(sum(CG_1(i,f_c)))}{23*23} \tag{15}$$

将 CG_1、CG_2、CG_3 和 CG_4 合并得到 64 * 4 维的特征向量,其表达式为:

$$MRCG'(i,f_c)=[CG_1(i,f_c);CG_2(i,f_c);CG_3(i,f_c);CG_4(i,f_c)] \tag{16}$$

(4)离散余弦变换(DCT)。对得到的 MRCG 特征进行离散余弦变换的目的是去除相关性,其表达式为:

$$MRACC(i,f_c)=(\frac{2}{M})^{0.5}MRCG'(i,f_c)\cos(\frac{\pi n(2c-1)}{2M}) \tag{17}$$

其中,c 表示频率通道,c 的范围为[0,64];M 为总通道数,本文中 M 取 64。当 $c>32$ 时,$MRACC(i,f_c)$ 的值基本接近于 0,因此选取前 32 维特征,即 32 * 4 维的特征向量。

1.3 神经网络模型

机器学习在语音识别领域取得了可观的研究成果,所以越来越多地将神经网络用于说话人识别中。BP 网络简单实用,但存在网络训练容易陷于局部最优解、无法调整到网络低层参数等问题。近年来提出了各种基于深度学习的深度神经网络,如稀疏编码网络(SA)、卷积神经网络(CNN)等。这些网络拥有更强大的建模和表征能力,能够实现复杂函数的逼近,不过这些网络属于前馈网络,表征时序信号的能力有限。而语音具有时序性,循环神经网络适合处理前后文有明显关系的数据,因此本文的神经网络模型选择 RNN 网络。LSTM 是一种特殊的 RNN 类型,是在 RNN 的基础上增加了输入门、遗忘门和输出门。可以改善 RNN 网络存在的梯度消失、梯度爆炸等问题。网络结构如图 4 所示。

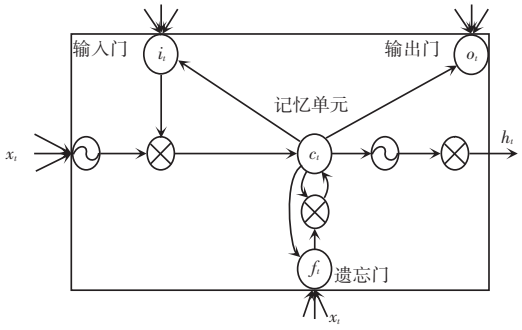

图 4 LSTM 细胞结构图

Fig. 4 LSTM cell structure diagram

在图 4 中,输入门 i_t 决定送入记忆单元的信息以及更新;遗忘门 f_t 根据上一时刻的输出 h_{t-1} 以及此时的输入 x_t 进行信息选择,保留重要信息,遗忘非重要信息;输出门 o_t 控制当前细胞单元的输出,确定哪些信息可作为下一时刻的输入。计算公式如下所示:

$$f_t=\sigma(W_f\cdot[h_{t-1},x_t]+b_f) \tag{18}$$

$$i_t=\sigma(W_i\cdot[h_{t-1},x_t]+b_i) \tag{19}$$

$$o_t=\sigma(W_o\cdot[h_{t-1},x_t]+b_o) \tag{20}$$

其中,W_f、W_i、W_o 为权重参数;b_f、b_i、b_o 为偏置参数;激活函数均为 $sigmoid$。

当前时刻的候选记忆细胞 $\tilde{c}_t$、记忆细胞 c_t 和隐藏单元 h_t 可用式(21)~(23)进行计算:

$$\tilde{c}_t=tanh(W_c\cdot[h_{t-1},x_t]+b_c) \tag{21}$$

$$c_t=f_t*c_{t-1}+i_t*\tilde{c}_t \tag{22}$$

$$h_t=o_t*tanh(c_t) \tag{23}$$

其中,“ * ”为 Hadamard 积。

本文采用 LSTM 神经网络构建说话人识别模型,利用 LSTM 神经网络学习一个由语音特征参数到说话人识别结果的非线性映射。LSTM 网络模型结构如图 5 所示。网络的输入序列层节点数为语音特征的维度;设置 2 层 LSTM 隐藏层,用来传递信息;为了防止过拟合现象,设置 Dropout 层;最后依次为全连接层、Softmax 层和分类层,输出节点数为说话人的数目。训练阶段,将特征序列输入到 LSTM 网络,LSTM 网络将根据序列数据的时间步进行训练,多次训练保存最优模型,由此得到说话人模型。识别阶段,将测试语音的特征序列输入训练好的说话人模型中,得到预测结果,将其进行比对,概率最大的即为预测的说话人身份。

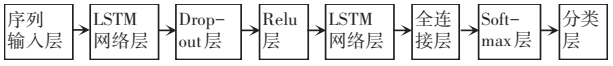

图 5　LSTM 网络模型结构图

Fig. 5　LSTM network model structure diagram

实际应用环境复杂，多数情况下都存在环境噪声，针对噪声环境下说话人识别系统的识别率显著下降问题，本文分别从鲁棒性特征参数和语音预降噪处理角度对识别系统进行了改进。

2　实验和分析

2.1　实验条件

2.1.1　实验数据集

本文使用语音来自 TIMIT 库语音集[12]，语音信号采样率是 16 kHz，量化位数为 16 bit，单通道录音。此语音集由 8 种美国英语方言组成，包含 630 个说话人的录音。每个人共 10 个种类丰富的句子，其中包含方言、紧凑句子以及音素发散句子。实验噪声选取的是来自 noisex-92 噪声库的 write 噪声、pink 噪声以及非平稳 factory 噪声。原始纯净语音分别加入不同信噪比的上述噪声模拟含噪语音。信噪比大小设置为 0 dB、5 dB、10 dB、30 dB。

2.1.2　实验设置

从数据集中选取 50 个说话人(男 30 人，女 20 人)，按照 4∶1 的比例分成训练样本集和测试样本集。每个人的训练模型使用 8 句语音，测试使用 2 句语音。对语音进行预处理后，按帧长 320、帧移 160，逐帧提取语音特征参数。

本文使用的特征包括 MRACC、MRCG、MFCC、LPCC 特征。特征维度设置如下：MACC 特征参数为 128 维；MRCG 特征参数为 256 维；MFCC 参数包含 12 维 MFCC、以及 12 维一阶差分参数和 12 维二阶差分参数，共 36 维；LPCC 参数 12 维。

说话人识别模型采用神经网络模型，分别为 BP 网络、SA 稀疏编码网络、LSTM 网络。网络模型的具体设置如下：LSTM、BP、SA 网络的输入层节点数均为输入特征的维度，输出层节点数为说话人数目；隐藏层设置分别为 LSTM 网络设置 2 层隐藏层，每层节点数均为 400；BP 神经网络隐藏层设置 2 层，每层节点数也都为 400；SA 网络设置 2 层隐藏层，每层节点数也都为 400。

2.2　前端降噪处理对比实验

为了验证本文提出的降噪方法的必要性以及有效性。进行了以下对比试验，本组实验特征参数选择 MRACC，网络模型为 LSTM 网络。实验结果见表 1。

表 1　说话人识别系统是否进行降噪处理的识别率

Tab. 1　The recognition rate of whether the speaker recognition system carries out noise reduction　%

信号噪声	信噪比/dB	未经过降噪处理	前端降噪处理	
			常规谱减	本文方法
white	30	79.39	83.41	**84.92**
	10	38.97	55.75	**62.29**
	5	39.47	40.62	**53.29**
	0	23.32	29.91	**39.46**
pink	30	81.50	83.61	**85.38**
	10	48.02	67.19	**70.42**
	5	36.29	45.90	**53.30**
	0	25.32	34.07	**42.32**
factory	30	81.69	84.94	**86.87**
	10	51.08	69.40	**74.43**
	5	41.58	47.06	**55.30**
	0	32.54	35.15	**46.79**

从表 1 可以看出，在以上噪声环境下，进行前端降噪处理相较于未进行前端降噪处理识别率高，如在 factory 噪声 0 dB 环境下，本文方法和常规谱减法对比未进行降噪处理的识别率分别提高了 14% 和 6%，因此进行前端降噪处理是有必要的，分析后可知对比常规谱减法，本文提出的降噪方法更有效。

2.3　不同特征参数的对比实验

为了验证 MRACC 特征的有效性，选取 MRACC、MRCG、MFCC、LPCC 四种特征参数进行了以下 2 组实验。网络模型选择 LSTM 网络。实验一是验证 MRACC 特征的抗噪性能进行的对比试验，实验结果见表 2。实验二为验证经过前端降噪处理后 MRACC 性能的对比实验，实验结果见表 3。

表 2　未进行前端降噪处理的不同特征参数的识别率

Tab. 2　The recognition rate of different feature parameters without front-end noise reduction　%

噪声信号	信噪比/dB	特征参数			
		MRACC	MRCG	MFCC	LPCC
white	30	79.39	70.95	59.34	43.51
	10	38.97	26.22	21.15	13.50
	5	39.47	22.73	17.57	11.89
	0	23.32	15.76	13.81	11.70
pink	30	81.50	79.30	66.09	50.68
	10	48.02	36.43	32.27	21.94
	5	36.29	26.74	22.99	17.22
	0	25.32	17.70	14.58	13.56
factory	30	81.69	81.77	67.72	51.95
	10	51.08	42.23	34.65	30.98
	5	41.58	32.59	27.54	25.44
	0	32.54	24.25	23.67	17.68

由表 2 得到，在 factory 噪声信噪比 0 dB 时，MRACC 识别率比 MRCG 高出 8%，比传统特征

MFCC 和 LPCC 均高出约 10%。在 10 dB 时，MRACC 识别率比 MRCG、传统特征 MFCC 和 LPCC 分别高出约 10%、15%和 20%。因此 MRACC 特征相较于 MRCG、MFCC、LPCC 特征在噪声环境下的识别率更好，该特征的抗噪性能较好。

表 3 进行前端降噪处理的不同特征参数的识别率

Tab. 3 The recognition rate of different feature parameters for the front-end noise reduction %

噪声信号	信噪比/dB	特征参数			
		MRACC	MRCG	MFCC	LPCC
white	30	**84.92**	76.43	63.45	49.62
	10	**62.29**	41.62	32.32	24.12
	5	**53.29**	28.47	23.21	20.18
	0	**39.46**	26.80	21.42	18.39
pink	30	**85.38**	81.46	68.61	57.69
	10	**70.42**	46.43	35.24	29.84
	5	**53.30**	31.74	29.70	25.90
	0	**42.32**	28.70	23.31	21.76
factory	30	**86.87**	85.80	69.26	59.80
	10	**74.43**	54.87	41.60	37.55
	5	**55.30**	40.98	34.49	30.62
	0	**46.79**	33.78	30.83	24.85

针对复杂的噪声情况，研究中对语音先降噪处理，再提取特征。由表 3 可以看出，在 write 噪声环境下，当信噪比为 30 dB 时，MRACC 识别率比 MRCG 高出约 10%，比传统特征 MFCC 和 LPCC 均高出约 20%。综上，MRACC 特征具有一定的鲁棒性，并且经过前端降噪处理后，MRACC 特征依旧稳定。

2.4 不同网络模型下的对比试验

为了验证 LSTM 网络处理时序信号的性能，选取 LSTM 网络、BP 网络、以及 SA 稀疏自编码网络进行说话人识别对比试验。特征参数选择 MRACC 特征，实验结果见表 4。

表 4 在不同网络模型下 MRACC 特征的识别率

Tab. 4 Recognition rate of MRACC features under different network models %

噪声信号	信噪比/dB	网络模型		
		LSTM	SA	BP
white	30	**84.92**	73.40	69.10
	10	**62.29**	57.90	52.40
	5	**53.29**	51.10	48.40
	0	**39.46**	36.90	32.20
pink	30	**85.38**	74.70	70.00
	10	**70.42**	57.70	54.90
	5	**53.30**	49.50	47.40
	0	**42.32**	38.40	34.20
factory	30	**86.87**	75.50	72.70
	10	**74.43**	62.10	58.90
	5	**55.30**	50.60	47.90
	0	**46.79**	37.50	34.20

可以看出，BP 网络模型对输入特征进行深层抽取的能力不如 SA 和 LSTM 网络，识别效果较差一些；SA 网络属于深层网络，识别效果比 BP 网络好。在 pink 噪声 5 dB 环境下，LSTM 网络较 SA 网络的识别率高出 4%，较 BP 网络高出 6%。综上，无论是在噪声环境下、还是非噪声环境下，LSTM 网络对时序语音信号的处理能力优于 BP 以及 SA 网络。

3 结束语

针对目前噪声环境下说话人识别系统识别率较低的情况，本文提出一种基于 MRACC 特征的说话人识别系统。利用改进型谱减法对语音进行预降噪处理，接着使用幂函数代替对数函数对听觉能量进行非线性压缩，提取语音的 MRACC 特征，最后通过 LSTM 网络完成模型训练与说话人识别。经过实验验证，使用改进型谱减法对语音进行预降噪处理，使说话人系统在低信噪比时的识别效果得到了明显改善。无论是在纯净环境下、还是在噪声环境下，MRACC 特征相比较传统特征 MFCC 和 LPCC 能够得到更好的识别效果。

参考文献

[1] CAMPBELL J P. Speaker recognition: a tutorial[J]. Proceedings of the IEEE, 1997, 85(9): 1437-1462.

[2] DAVIS SB, 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.

[3] 张翠玲，丁盼. 基于 LPC 倒谱特征融合的法庭说话人识别方法[J]. 中国刑警学院学报, 2020(5): 117-121.

[4] CHEN Jitong, WANG Yuxuan, WANG Deliang. A feature study for classification-based speech separation at low signal-to-noise ratios [C] // 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Florence, Italy: IEEE, 2014: 7039-7043.

[5] 杨瑶，陈晓. 基于神经网络的说话人识别实验设计[J]. 实验室研究与探索，2020，39(9)：38-41，50.

[6] 盖晁旭. 基于稀疏编码的鲁棒说话人识别[D]. 哈尔滨：哈尔滨理工大学，2017.

[7] 王华朋. 基于深度双向 LSTM 网络的说话人识别[J]. 计算机工程与设计，2020，41(6)：1768-1772.

[8] LOPZ-MORENO I, GONZALEZ-DOMINGUEZ J, PLCHOT O, et al. Automatic language identification using deep neural networks [C]//Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing. Florence, Italy : IEEE press, 2014: 5337-5341.

[9] HAN Song, KANG Junlong, MAO Huizi, et al. ESE: Efficient speech recognition engine with sparse LSTM on FPGA [C] //Proc. of the 2017 ACM/SIGDA International Symposium on Field - Programmable GateArrays. Monterey, CA, USA: ACM press, 2017: 75-84.

[10] GREFF K, SRIVASTAVA R K, KOUTN J, et al. LSTM: A search space Odyssey [J]. IEEE Transactions on Neural Networks and Learning Systems, 2016, 28 (10): 2222-2232.

(下转第 72 页)