基于特征融合的通信语音干扰效果客观评估

林云[1]，徐怀韬[1]，王森[1]，张思成[1]，庄龙[2]
（1. 哈尔滨工程大学信息与通信工程学院，黑龙江 哈尔滨 150001；2. 安徽大学集成电路学院，安徽 合肥 230039）

摘 要：针对通信语音干扰效果客观评估问题，提出了基于多测度与多模态融合的 2 种评估方法。首先，利用端点检测算法以及动态时间弯折算法对受扰语音数据进行预处理。然后，提取数据中的语音内容并与标准语音进行测度计算得到 5 种测度，将 5 种测度融合后利用随机森林模型进行质量等级评估。最后，结合多模态融合技术，设计了基于残差结构的神经网络模型，融合受扰语音数据的图域、测度域特征并进行质量等级评估。实验结果表明，2 种方法的评估准确率均达到了 90%以上。其中，多模态评估方法与现有的研究方法相比，准确率提升了约 3.269%，证明所提方法具有更优的性能。

关键词：语音质量评估；语音信号处理；多模态融合；深度神经网络
中图分类号：TN912.3
文献标志码：A
DOI: 10.11959/j.issn.1000−436x.2023043

Objective assessment of communication speech interference effect based on feature fusion

LIN Yun[1], XU Huaitao[1], WANG Sen[1], ZHANG Sicheng[1], ZHUANG Long[2]
1. College of Information and Communication Engineering, Harbin Engineering University, Harbin 150001, China
2. School of Integrated Circuits, Anhui University, Hefei 230039, China

Abstract: In view of the objective assessment problem of the effect of communication speech interference, methods based on multi-measurements and multimodal fusion were proposed. First, the interfered speech was preprocessed by the endpoint detection algorithm and time warping algorithm. Then, the content of speech was extracted and performed measurement calculated with the standard speech to obtain five kinds of measure. After the fusion of five measures, random forest model was used to assessed the quality level. Finally, a neural network model based on residual structure was designed combined multimodal fusion technique, which fused the graph domain and measure domain features of the interfered speech data and performed quality level assessment. Experimental results show that the accuracy of two methods have reached more than 90%. Among them, the multimodal assessment method improves the accuracy by about 3.269% compared with the existing research methods, which proves that it has a better performance.

Keywords: speech quality assessment, speech signal processing, multimodal fusion, deep neural network

0 引言

在万物互联的趋势下，频繁出现的各种干扰严重影响了语音通信系统的通信质量，其中包括不可避免的设备热噪声以及由干扰源产生的主动干扰。如何在强干扰环境下对干扰效果进行评估，给出客观的评估指标，以及如何评估通信系统在强干扰环境下的抗干扰性能，对维持高质量语音通信、保障通信系统的有效运行十分关键[1]。一般情况下，语音质量评估方法分为主观评估、客观评估 2 种。主

收稿日期：2022–08–30；**修回日期**：2022–11–30
基金项目：国家自然科学基金资助项目（No.62201172）；中央高校基本科研业务费专项资金资助项目（No.3072022CF0804, No.3072022CF0601）
Foundation Items: The National Natural Science Foundation of China (No.62201172), The Fundamental Research Funds for the Central Universities (No.3072022CF0804, No.3072022CF0601)

观评估方法通常为绝对等级评定（ACR, absolute category rating），通过测听人员对语音打分的高低反映语音质量，打分范围为 1～5，分数越高表明语音质量越好，该指标又被称为均值意见分（MOS, mean opinion score）[2]，已广泛应用于各种质量评估体系。虽然主观评估能够准确地反映人耳感知情况，但这种方法需要消耗大量时间和人力，成本较高，所以利用计算机等非人工手段的客观评估方法逐渐成为主流。

客观评估方法中最典型的代表是分别由 ITU-T P.862、ITU-T P.863 文件提出的感知语音质量评估（PESQ, perceptual evaluation of speech quality）[3]和感知客观听觉质量评估（POLQA, perceptual objective listening quality assessment）[4]。PESQ 主要用于评估语音编解码器性能和端到端网络的语音质量，并添加了 MOS 映射以及宽带语音评估等新功能。作为 PESQ 的改进版，POLQA 实现了技术升级，其涵盖了全新的评估场景，提高了 4G/LTE 和 VoIP（voice over Internet protocol）服务中语音质量评估的准确率，并在 PESQ 的基础上扩展了对超宽带、全频带语音的评估能力。随着数字信号处理技术和人工智能技术的快速发展，越来越多的研究者将深度神经网络与语音信号处理相结合，拓展了客观评估场景。

Affonso 等[5]提出了一种基于深度信念网络的语音质量分类器模型，研究在语音传输过程中丢包率对语音质量的影响，并在公开数据集上与 ITU-T Recommendation P. 563 算法进行对比，验证了将该方法用于协助运营商完成网络管理任务的可行性。Fu[6]等利用双向长短记忆周期结构建立了一种与 PESQ 高度相关的端到端语音质量评估模型，称为“质量网”。随后将其作为目标函数训练语音增强模型，有效缓解了语音增强任务中模型优化准则与质量评估准则之间的失配问题[7]。Rodríguez 等[8]针对无线传输过程中信噪比、多普勒频移等参数对通话质量的影响进行建模，提出了“无线损伤因子”这一概念，并证明了无线信道参数与 MOS 之间存在高度相关性，随后将该参数加入 ITU-T G.107 模型中，提升无线信道场景下质量评估的准确率[9]。Lo 等[10]采用卷积和递归神经网络建立针对语音转换的质量评估模型 MOSNet，并在 2018 年的语音转换挑战赛上取得了优异的成绩，证实了利用该方法对语音转换系统进行评估的可靠性。

上述质量评估研究均采用公开数据集，并且针对电话网络、语音增强、无线传输及语音转换等场景展开，缺乏强干扰（超低信噪比）环境下的质量评估研究。Zhang 等[11]提出了“信息损伤级”这一概念代替 ACR 作为受扰语音数据的主观标签，并在干扰环境下，采集得到 3 种不同场景的受扰语音数据集，进而建立质量评估框架，预测受扰语音数据的质量分数。随后，Wang 等[12]提出了一种深度学习的评估方法，将受扰语音转换为对数梅尔谱图输入 AlexNet 模型中进行评估，并利用实际通信场景下采集的受扰语音数据对该评估方法进行验证。其中，受扰语音质量等级评估的准确率达到了 87.5%。

这两项工作填补了强干扰环境下语音质量评估研究的空白，通过提取受扰语音数据的不同域特征，并分别结合机器学习模型和深度神经网络，在质量分数预测和质量等级分类问题中取得了良好的评估效果。本文在此研究基础上，针对受扰语音特征展开研究。首先，对 5 种特征下受扰语音的测度进行融合，并利用机器学习分类器建立测度值与信息损伤级之间的映射关系，完成多测度融合实验。然后，为进一步研究语音特征对质量等级评估性能的影响，本文结合了多模态特征融合技术，设计了基于 5 层残差结构的网络模型[13]，并利用该模型融合图域、测度域 2 种特征，尝试利用不同域特征的互补性构建多模态特征与语音质量等级之间的映射关系[14]。最后，在数据集上进行实验，并与文献[12]中的评估结果进行对比。实验结果表明，所设计的多模态融合方法提升了受扰语音质量等级评估的准确率。

1 受扰语音数据集

1.1 标准语音库建立

本文针对真实通信干扰场景下的受扰语音数据展开质量评估研究，实验部分所使用的受扰语音数据集由标准语音文件经通信设备传输产生。其中，标准语音文件的生成方式遵循行业标准[15]中的规范，其生成过程如下。

1) 码本内容为数码 0、1、2、3、4、5、6、7、8、9 组成的报文。

2) 10 个数码的读音方式如表 1 语音数码读音标准所示。

表 1　　语音数码读音标准

数码	读音
0	Dong（洞）
1	Yao（幺）
2	Liang（两）
3	San（三）
4	Si（四）
5	Wu（五）
6	Liu（六）
7	Guai（拐）
8	Ba（八）
9	Gou（勾）

3) 将每个数码按表 1 中的读音方式进行录制，如 0（洞）、1（幺）。0～9 共 10 个数码形成 10 个样本文件。

4) 随机排列 10 个样本文件，将每 4 个数码连接成一个数字码组，每 15 个码组形成一个标准语音码本。语音码本中每 2 个数字码组间隔 1 s，数字码组内各数码间隔 0.5 s，每条标准语音码本时长为 75 s[16]。

5) 重复步骤 3)和步骤 4)形成标准语音文件，如图 1 所示。

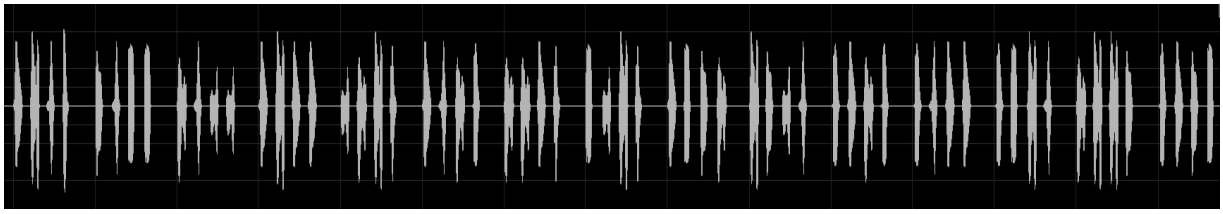


图 1　标准语音文件

1.2　受扰语音数据采集

标准语音文件生成后，进行受扰语音数据采集，采集环境如图 2 所示。其中，收发设备均包含录取接口、控制模块以及平板电脑，用于设备控制以及语音录制。

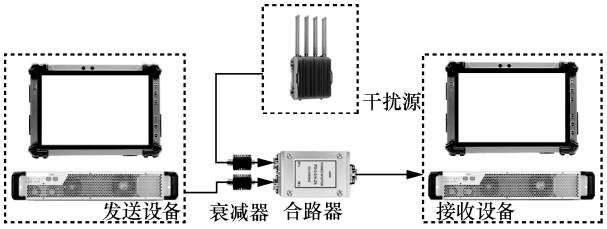


图 2　受扰语音数据采集环境

采集过程中，发送设备与干扰设备发出的信号经衰减器衰减后（降低接收信号功率至接收机动态范围内），利用合路器进行合路，从而使接收设备接收到叠加干扰后的语音数据。采集过程中，干扰源由中控设备控制，干扰参数设置如表 2 所示。

表 2　　干扰参数设置

干扰参数	值
干扰样式	噪声调频
干扰频点/MHz	88.2
干扰带宽/MHz	1
干扰功率/W	1

1.3　受扰语音数据标注

数据采集完毕后，通过人工测听对语音数据的受扰程度进行主观评价。行业标准中提出了信息损伤级这一概念作为受扰等级的主观评价标准。信息损伤级表示人工测听时对受扰语音内容的识别率，可分为 5 个等级，每个等级间可连续打分，信息损伤级划分标准如表 3 所示。

损伤级划分以单字识别率 r 为评估依据，计算方式为

$$r=\frac{W}{H}\times 100\% \tag{1}$$

其中，W 为正确识别的单字个数，H 为每条语音内

表 3　　信息损伤级划分标准

信息损伤级 G_s	r	语音质量描述
1 级	$100\%\geqslant r>95\%$	干扰噪声弱，字音较清晰，个别单字不能准确判别，通信内容可懂
2 级	$95\%\geqslant r>80\%$	干扰噪声较强，部分字音较不清晰，个别单字不能准确判别，通信内容基本可懂
3 级	$80\%\geqslant r>50\%$	干扰噪声强，字音不清晰，大部分单字不能判别，通信内容基本不可懂
4 级	$50\%\geqslant r>10\%$	干扰噪声强，字音模糊，大部分单字不能判别，通信内容不可懂
5 级	$10\%\geqslant r\geqslant 0$	干扰噪声很强，绝大部分单字不能准确判别，通信内容完全不可懂

容的总字数。对全部测听人员给出的信息损伤级结果进行统计，得到信息损伤级打分并取平均值，将该结果作为受扰语音样本的信息损伤级，计算过程为

$$G_s = \frac{1}{t}\sum_{i=1}^{t} G_i \tag{2}$$

其中，G_i 为第 i 名测听人员评估的信息损伤级，t 为测听人员个数。质量等级评估任务中，将 G_s 值四舍五入作为信息损伤级标签。

1.4 受扰语音数据集分布

完成数据采集和标注后，得到 5 个信息损伤级（1～5 级），共 1 000 条受扰语音，其分布如图 3 所示。

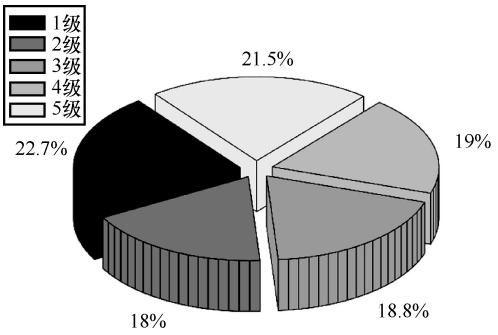


图 3 信息损伤级分布

2 预处理及特征提取

2.1 端点检测

图 4 展示了数据集中某 2 条受扰语音样本的对数梅尔谱图。采集过程中，为保证收发同步，在标准语音文件之前添加由 2 个单频信号组成的同步头，同步头与第一个码组间隔 5 s，每条受扰语音数据的时长由 75 s 变为 85 s。实际传输过程中收发时间无法严格同步，导致受扰语音数据存在时延，使语音样本中第一个码组的起始位置发生变化。而质量等级评估的关键在于分析每条语音内容的受扰情况，起始位置的不同会对语音内容的提取造成较大误差。因此，首先对数据集中的语音样本进行语音端点检测（VED, voice endpoint detection），定位语音内容的起始点。常用的端点检测方法可分为基于阈值、基于机器学习分类器和基于深度学习模型 3 种。后 2 种方法需要估计语音信号和干扰信号的特征参数来进行检测，计算复杂度较高，而在通信对抗场景下干扰信号的先验知识往往难以获得。此外，每条受扰语音数据的语音长度是固定的，只需检测出起始点就可以从数据中提取出完整的语音内容，因此，本文选择基于阈值的端点检测方法。同步头末尾与第一个数码间存在 5 s 的静音段，若存在时延，语音内容的起始点则会出现在 10 s 后，查看样本数据的过程中，发现每条样本语音的时延均不超过 3 s，因此本文对 9～13 s 的受扰语音数据进行检测。

检测时，浊音和静音可以利用短时能量特征判别。由于清音能量较低，在短时能量检测时，可能会因能量小于阈值被误判为静音。此时，利用短时过零率特征对静音和清音加以判别。结合两者即可对浊音、清音和静音段进行区分，计算过程如下。

首先，输入待测信号 x，对 x 进行分帧得到分帧后信号 y，帧数为 μ。然后，计算每帧信号的短时能量 E_μ 及过零率 Z_μ

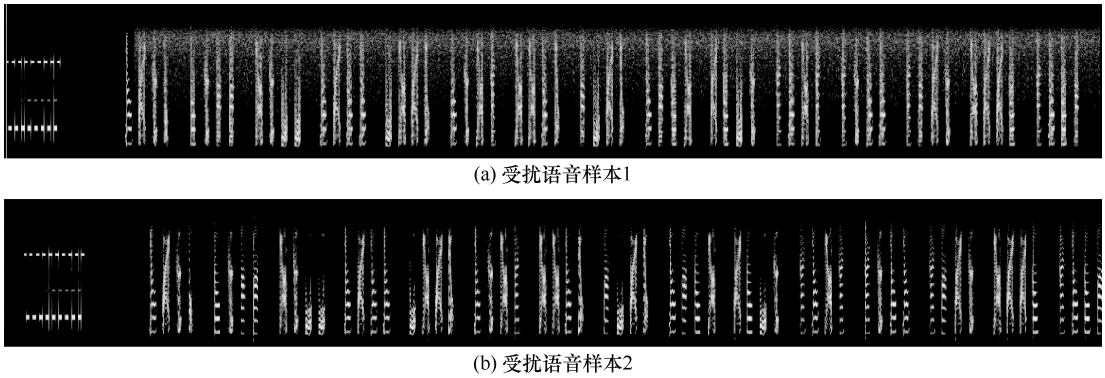
$$E_\mu = \sum_{m=1}^{N} y_\mu^2(m) \tag{3}$$

$$Z_\mu = \frac{1}{2}\sum_{m=1}^{N}\left|\mathrm{sgn}\left[y_\mu(m)\right] - \mathrm{sgn}\left[y_\mu(m-1)\right]\right| \tag{4}$$

其中，N 为信号长度，sgn[·] 为符号函数。

$$\mathrm{sgn}[\varphi] = \begin{cases} -1, \varphi < 0 \\ 1, \varphi \geqslant 0 \end{cases} \tag{5}$$

将分帧后信号的前 115 帧作为静音段，计算该



(a) 受扰语音样本1

(b) 受扰语音样本2

图 4 受扰语音样本的对数梅尔谱图

段的短时能量及短时过零率的平均值 $\overline{E}$、$\overline{Z}$，并设置检测时短时能量及过零率阈值 E_{thr1}、E_{thr2}、Z_{thr} 为

$$\begin{cases} E_{\text{thr1}} = 2\overline{E} \\ E_{\text{thr2}} = 4\overline{E} \\ Z_{\text{thr}} = 20\overline{Z} \end{cases} \tag{6}$$

端点检测算法如算法 1 所示，其中，status 为帧状态标志，其值为 0～3 分别对应 4 种状态，依次表示当前帧为静音段、似语音段、语音段、语音结束段。其中最小、最大静默长度阈值分别设置为 5、15，单位为帧。

算法 1　端点检测算法

输入　待测语音信号 x

输出　检测后第一个数码的起始位置

for　$i=1$　to　n，进行迭代；

switch status

case {0, 1}

if $E_i > E_{\text{thr2}}$，确定进入语音段， status=2，记录当前帧数 x_1

else if status=1

else status=0

end if

case 2

if $E_i > E_{\text{thr1}}$, $Z_i > Z_{\text{thr}}$，保持在语音段

else

if 静音段小于最大静默长度，则语音段未结束

else if 语音段小于最小静默长度，该段为静音或噪声，status=0

else status=3，语音结束

end if

end if

case 3

status=0，初始化各参数，$x_1 = 0$

end case

end for

某条受扰语音样本的端点检测结果如图 5 所示，实线对应语音段的起始位置，虚线对应语音段的结束位置。如图 5 所示，第一个码组内的第一个数码与后 3 个数码所表现出的短时能量及过零率特征存在差异，后 3 个数码由于受到随机噪声干扰，表现出类似清音的性质，其所对应的短时能量较小但短时过零率较高，由于静音段短时能量为零，利用这 2 个参数同时进行判断，能有效避免受扰较严重的语音段因能量较小而被误判为静音。图 5 中左起第一条实线表明利用算法 1 成功检测到语音内容的起始点，根据该起始点位置提取受扰样本的语音内容。

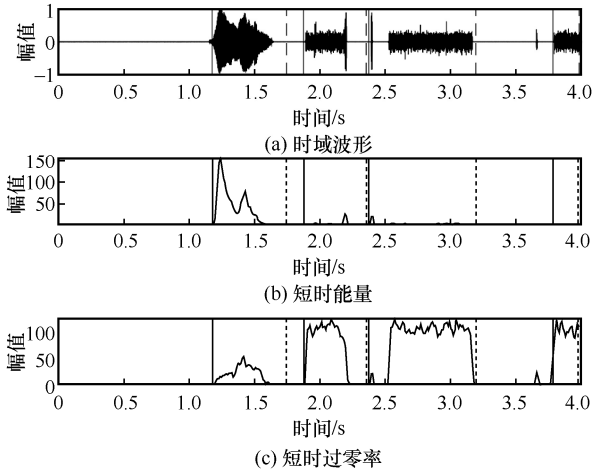


图 5　受扰语音样本的端点检测结果

2.2　特征提取

检测完每条样本的语音内容后，本文首先提取了受扰语音与原始语音的 5 种特征，分别为梅尔频率倒谱系数（MFCC, Mel frequency cepstrum coefficients）、梅尔能量、线性预测系数（LPC, linear prediction coefficient）、线性预测倒谱系数（LPCC, linear prediction cepstrum coefficient）及小波统计特征。

梅尔特征（包括 MFCC、梅尔能量）通过将频率转换为梅尔尺度的方式，更好地匹配人耳的听觉感知效果。其中，MFCC 对噪声及干扰的变化较敏感，能更好地反映受扰语音数据的声学特征。此外，由于 MFCC 的计算过程中使用了离散余弦变换（DCT, discrete cosine transform）进行去相关，其更加适用于各类机器学习算法[10-11,17]。LPC 和 LPCC 表征了发音过程中的声道变化特性，且 LPC 是求解 LPCC 的理论和计算基础，其基本思想是语音信号样点之间存在较强的相关性，可利用过去若干个样值或它们的线性组合对当前或未来时刻的样值进行预测[18]。小波统计特征[19]则是在小波变换的基础上，利用小波基函数对受扰语音信号进行分解，并提取分解后每一层近似系数的统计特征而获得，3 组特征的性质总结如表 4 所示。

表 4　　**特征性质总结**

特征类型	特征性质
梅尔特征（MFCC、梅尔能量）	听觉模型
LPC、LPCC	声道模型
小波统计特征	时频分析模型

1) MFCC 和梅尔能量特征

MFCC 特征是梅尔尺度下的倒谱参数，反映了人耳对频率的感知特性[20]。梅尔尺度 f_{mel} 与频率 f 的对应关系为

$$f_{\text{mel}}(f)=2\,595\log\left(1+\frac{f}{700\ \text{Hz}}\right) \tag{7}$$

MFCC 提取流程如图 6 所示。首先对输入信号进行预加重、分帧及加窗处理，帧长为 25 ms，并采用汉明窗对分帧后信号加窗，加窗后信号 $\boldsymbol{s}(n)$ 为

$$\boldsymbol{s}(n)=\boldsymbol{y}(n)\boldsymbol{w}(n) \tag{8}$$

其中， $\boldsymbol{y}(n)$ 为分帧后信号， $\boldsymbol{w}(n)$ 为窗函数。

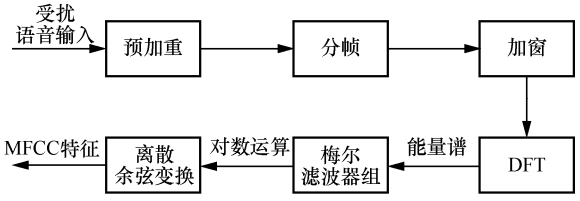


图 6 MFCC 提取流程

对 $\boldsymbol{s}(n)$ 进行离散傅里叶变换（DFT）得到幅度谱 $S_a(k)$ 为

$$S_a(k)=\sum_{n=0}^{N_{\text{DFT}}-1}\boldsymbol{s}(n)\mathrm{e}^{\frac{-\mathrm{j}2\pi k}{N_{\text{DFT}}}},0\leqslant n\leqslant N_{\text{DFT}}-1 \tag{9}$$

其中， N_{DFT} 为 DFT 点数， N_{DFT} =512。对幅度谱取平方得到能量谱，并利用梅尔滤波器组滤波，滤波器频率响应 $H_m(k)$ 为

$$H_m(k)=\begin{cases}0 & ,k<f(m-1)\\ \dfrac{k-f(m-1)}{f(m)-f(m-1)} & ,f(m-1)\leqslant k<f(m)\\ 1 & ,k=f(m)\\ \dfrac{f(m+1)-k}{f(m+1)-f(m)} & ,f(m)<k\leqslant f(m+1)\\ 0 & ,k>f(m+1)\end{cases} \tag{10}$$

其中， m 为滤波器编号， k 为 DFT 后的索引值。然后，计算滤波器系数加权后的对数能量为

$$S_e(m)=\ln\left(\sum_{k=0}^{N_{\text{DFT}}-1}\left|S_a(k)\right|^2\right)H_m(k),0\leqslant m\leqslant M \tag{11}$$

其中， M 为滤波器个数，经 DCT 后最终得到对应的 MFCC 特征为

$$\begin{aligned}\text{MFCC}(n)&=\sum_{m=0}^{N_{\text{DFT}}-1}S_e(m)\cos\left(\frac{\pi d(m-0.5)}{M}\right)\\ &d=1,2,\cdots,L\end{aligned} \tag{12}$$

其中， L 为 MFCC 特征向量的维度。提取过程中，针对维度选取进行实验，在考虑评估精度和资源使用的情况下选取 $L=24$。除 DCT 操作外，梅尔能量特征的提取过程与 MFCC 特征相同，最终选取 40 维的梅尔能量特征。

2) LPC 和 LPCC 特征

LPC 特征的计算过程如下。设受扰语音序列中的第 n 个抽样值为 $x(\lambda)$，其对应的线性预测值 $\hat{x}(\lambda)$ 为

$$\hat{x}(\lambda)=\sum_{i=1}^{D}a_i x(\lambda-i) \tag{13}$$

其中， D 为 LPC 阶数， a_i 为预测系数。预测样值与实际样值之间的误差为

$$\varepsilon(\lambda)=x(\lambda)-\hat{x}(\lambda)=x(\lambda)-\sum_{i=1}^{D}a_i x(\lambda-i) \tag{14}$$

为使误差最小，预测误差的均方值可以表示为

$$E_\lambda=E\{\varepsilon^2(\lambda)\}=E\{[x(\lambda)-\sum_{i=1}^{D}a_i x(\lambda-i)]^2\} \tag{15}$$

对式(15)中的线性预测系数求极值，即

$$\frac{\partial E_\lambda}{\partial a_i}=0,i=1,2,\cdots,D \tag{16}$$

得到

$$x(\lambda-i)x(\lambda)=\sum_{i=1}^{D}\hat{a}_i x(\lambda-i)x(\lambda-k) \tag{17}$$

根据相关函数定义，有

$$R(i,k)=x(\lambda-i)x(\lambda-k) \tag{18}$$

式(13)可改写为

$$R(i,0)=\sum_{i=1}^{D}\hat{a}_i R(i,k) \tag{19}$$

采用 Durbin 算法求解式(15)，得到受扰语音的 12 维 LPC 特征，LPCC 与 LPC 的递推关系如下

$$\begin{cases}b_0=a_1\\ b_c=a_c+\sum_{k=1}^{c-1}\dfrac{k}{c}b_k a_{c-k},1\leqslant c\leqslant D\\ b_c=\sum_{k=1}^{c-1}\dfrac{k}{c}b_k a_{c-k},c>D\end{cases} \tag{20}$$

其中，b_0 表示信号的直流分量，计算求得的系数 b_c 即 LPCC 特征。

3) 小波统计特征

本文采用的小波统计特征是在离散小波分解的基础上，提取分解后每一层近似系数的统计特征得到的。可以将待分解的受扰语音信号看作一段离散时间序列，其中，信号的低频部分通常蕴含序列特征，高频部分通常蕴含序列的变化细节。将分解时所使用的小波基函数构造为 2 个频率互补的滤波器对受扰语音信号进行滤波，即可得到信号中的高低频成分，这一过程称为“半子带滤波”。经过一次半子带滤波和一次 2 倍降采样后，即完成一次小波分解。假设受扰语音信号长度为 N，一次分解后得到高、低频部分的长度都为 $\frac{N}{2}$。分解时，先将受扰语音信号分解为高频和低频两部分，由于低频部分蕴含序列特征，后续的每一次分解都对上一次的低频部分进行再分解得到新的低频和高频部分，分解采用的小波基函数为 DB4（Daubechies）小波，分解层数为 4 层，小波分解流程如图 7 所示。

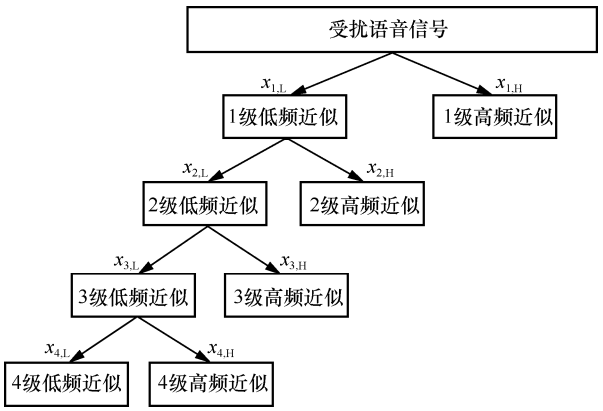


图 7　小波分解流程

分解结果可表示为

$$\begin{aligned} x_{\alpha,\mathrm{L}}[n_\mathrm{f}] &= \sum_{k=0}^{N} x_{\alpha-1,\mathrm{L}}[2n_\mathrm{f}-k_\mathrm{f}]g(k_\mathrm{f}) \\ x_{\alpha,\mathrm{H}}[n_\mathrm{f}] &= \sum_{k=0}^{N} x_{\alpha-1,\mathrm{H}}[2n_\mathrm{f}-k_\mathrm{f}]h(k_\mathrm{f}) \end{aligned} \tag{21}$$

其中，$x_{\alpha,\mathrm{L}}$、$x_{\alpha,\mathrm{H}}$ 分别表示第 α 层受扰语音信号的低频和高频成分，$g(k_\mathrm{f})$、$h(k_\mathrm{f})$ 为 DB4 小波基函数构造出的半子带滤波器。第 α 层中 $x_{\alpha,\mathrm{L}}[n_\mathrm{f}]$、$x_{\alpha,\mathrm{H}}[n_\mathrm{f}]$ 的长度均为 $\frac{N}{2^\alpha}$。

分解后，提取每一层的低频部分并计算其均值、方差、最大值及最大值索引作为受扰语音的小波统计特征。

2.3　测度计算

完成特征提取后，对受扰语音与标准语音间的测度进行计算。测度是指受扰语音与标准语音间的特征距离，它反映了不同等级受扰语音的特征变化趋势。值得注意的是，采用端点检测方法提取到的受扰语音内容与标准语音并不等长，无法直接利用欧氏距离计算测度。而动态时间弯折（DTW, dynamic time warping）算法[21]可用于计算 2 个时间上存在差异的序列间的特征距离。因此，本文采用该算法计算测度。

特征距离计算示意如图 8 所示，黑色粗实线与虚线对应 2 个时间序列。每条细实线将一个序列中的点与另一序列中具有相似值的点相连。

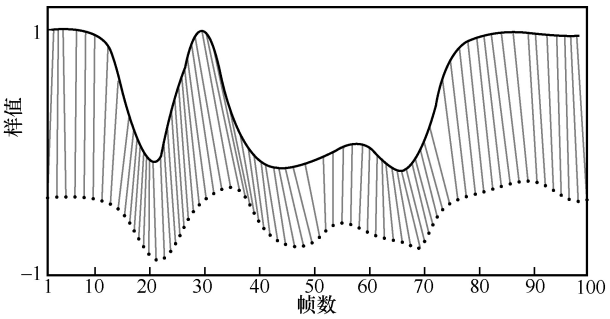


图 8　特征距离计算示意

动态时间弯折的原理是通过比较 2 个序列，并计算它们之间的最小距离来进行相似性度量。假设不同特征下，受扰语音和标准语音的特征序列分别为 p 和 q，所包含帧数分别为 u 和 v，即

$$\begin{cases} p = p_1, p_2, \cdots, p_u \\ q = q_1, q_2, \cdots, q_v \end{cases} \tag{22}$$

为使用 DTW 算法对齐 2 个特征序列，首先构造一个 $u \times v$ 的矩阵，矩阵 (i, j) 位置处的元素对应序列样值间的欧氏距离 $d_\mathrm{e}(p_i, q_j)$。此时，从起始位置 $i=1, j=1$ 至 $i=u, j=v$ 的累计失真距离为

$$\begin{aligned} D_\mathrm{e}(p_i, q_j) = \min\{&D_\mathrm{e}(i-1, j-1), \\ &D_\mathrm{e}(i-1, j), D_\mathrm{e}(i, j-1)\} + d_\mathrm{e}(i, j) \end{aligned} \tag{23}$$

完成测度计算后，可以将评估过程看作已知测度对受扰语音质量等级进行分类的问题，为后续利用机器学习模型进行多测度融合实验提供依据。

3 受扰语音评估方法

3.1 单测度分析

在进行多测度融合实验之前，首先对单测度的分布情况进行分析，单测度值分布如图 9 所示。以 MFCC 特征、小波统计特征 2 种特征测度为例，可以看出即使在 2 种不同特征下，5 个信息损伤级的受扰语音测度也存在不同程度的混淆，其中损伤级 2 和损伤级 3 的混淆程度较严重，这样的测度混淆无疑会给分类器的质量评估性能造成较大影响。因此，仅通过单一测度或某 2 种测度对受扰语音数据进行质量等级评估的难度较大，这也是后续进行多测度融合以及多模态融合实验的主要原因。

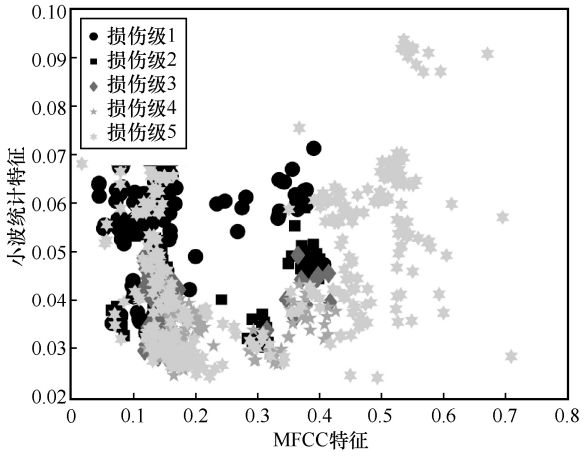


图 9 单测度值分布

3.2 多测度融合方法

根据上述分析，首先对多测度融合实验进行设计，该过程参考了标准文件附录 B“客观评定算法”中的相应内容。考虑到采集过程中，干扰功率的波动会导致同一信息损伤级内的语音数据存在一定的特征差异，这就要求所使用的模型不仅需要适应标注误差，而且可以一定程度地容忍同等级内失真测度的差异。由于随机森林具有准确率高、不易过拟合、特征选择能力强等优点，因此本文选取随机森林作为受扰语音数据的质量等级分类器[22]。其构建过程如下。

步骤 1 对受扰语音数据集进行随机采样，得到多个训练子集。

步骤 2 在各个训练子集上训练得到不同的决策树模型。

步骤 3 将训练得到的多个决策树模型进行组合，并提取最终的输出结果。

根据上述步骤，将每个样本 5 种特征下的测度进行融合，合并为一个 104 维的测度向量，输入随机森林分类器中进行训练，多测度融合模型构建过程如图 10 所示。

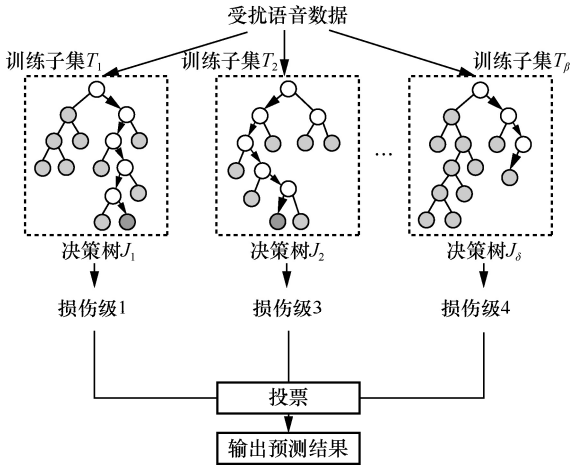


图 10 多测度融合模型构建过程

设 5 种特征对应的测度分别为$V_1,V_2,\cdots,V_5$，并与样本数据的信息损伤级标签合并，组成训练集$T=\{(V_{h1},V_{h2},\cdots,V_{h5}),l_i\}_{h=1}^{K}$，$K$ 为样本数量，l 为数据标签。样本预测步骤如下。

步骤 1 对原训练集进行重采样，随机产生 β 个新训练集$T_1,T_2,\cdots,T_\beta$。

步骤 2 从 5 种测度中任意抽取$\gamma(\gamma\leqslant 5)$个属性作为当前节点的分裂属性。

步骤 3 所有决策树均不进行剪枝。

步骤 4 对于第 h 个样本 X，利用决策进行预测，汇总预测的信息损伤级$J_1(X),J_2(X),\cdots,J_\delta(X)$，$\delta$ 为决策树的数量。

步骤 5 采用投票的方式，将决策树中输出最多的损伤级作为该样本的预测结果。

3.3 多模态融合方法

在语音质量评估领域中，现有方法仅使用图域特征[10,23-24]或测度向量等单一模态特征进行质量评估，并没有考虑不同模态间信息的互补性以及模态融合对语音质量评估的重要性。因此，在利用随机森林模型完成多测度融合评估后，为进一步分析不同域特征对受扰语音评估性能的影响，本文结合了多模态信息融合方法中的“晚期”融合（决策级融合）思想，利用深度学习模型分别对受扰语音数据的图域模态（对数梅尔谱图）、测度模态（5 种测度）进行训练，再通过多模态联合架构中的“加”联合

方式对模型的输出结果进行融合，多模态融合模型如图 11 所示。

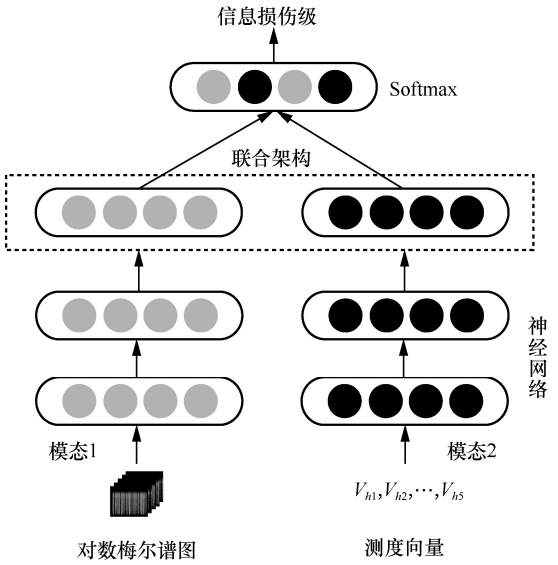


图 11　多模态融合模型

利用卷积神经网络可自动提取高维数据特征的特点[25]，将基于 5 层残差结构的神经网络作为多模态融合模型，残差结构如图 12 所示。

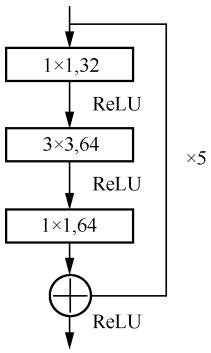


图 12　残差结构

进行特征融合时，将对数梅尔谱图和测度向量并行输入 2 个相同的神经网络模型中，经残差结构中的卷积层提取深层信号特征后，得到受扰语音数据的特征向量，其维度分别为 1×1 024 和 1×128。随后利用 Concatenate 操作将 2 个网络全连接层的输出组合在一起，得到一个维度为 1×1 152 的联合特征向量，利用 Softmax 层进行类别映射，从而完成受扰语音数据的多模态融合质量评估。多模态融合模型评估流程如图 13 所示。

4　实验及分析

本文设计并构建了受扰语音数据的质量等级评估任务，实验所用数据集如第 1 节所述，涵盖 5 个信息损伤级，共 1 000 条样本语音。端点检测后的每条语音样本时长为 73 s。由于端点检测过程中某些通信中断样本（无语音内容）存在检测误差，为避免对后续实验造成影响，将检测失败的样本去除，得到 972 条样本语音，并依据该样本量划分数据集。在多测度融合及多模态融合实验中，训练集、验证集及测试集样本比例均为 6:2:2。

在 Windows 10 操作系统下选择 Python 机器学习框架 Scikit-Learn 完成基于随机森林模型的多测度融合质量等级评估；同时，选择 Keras 深度学习框架完成多模态融合质量等级评估。实验在一台配有 Intel i7-11800H CPU 与 Nvidia 3060 GPU 的电脑上进行，运行内存为 16 GB。

4.1　单测度实验

为定量说明单测度评估的局限性，首先利用随机森林模型进行了单测度质量评估实验，并与支持向量机（SVM, support vector machine）、逻辑回归

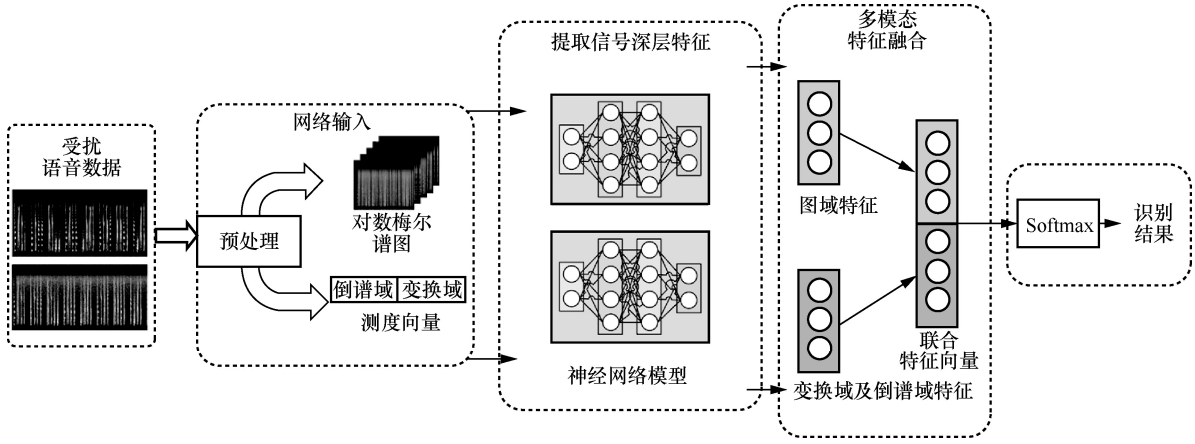


图 13　多模态融合模型评估流程

（LR, logistic regression）这 2 种传统的机器学习分类器进行对比。支持向量机是一种几何方法，其基本思想是求解能正确划分受扰语音数据，且能使不同等级数据几何间隔最大的分类超平面来完成质量等级评估。逻辑回归是一种统计方法，假设 5 个损伤级的语音数据各自服从某种分布，进而采用极大似然估计求解分类概率与输入数据的关系，通过比较概率值判断损伤级。

进行单测度实验时，将每条受扰语音样本的 5 种测度分别输入分类器中，以平均准确率作为衡量指标，实验结果如表 5 所示。通过分析发现，逻辑回归的平均准确率整体较低，支持向量机和随机森林的效果较好，采用梅尔能量作为输入时，3 种分类器得到的平均准确率均较高。此外，将单一测度作为输入时，3 种分类器的质量评估平均准确率均未突破 90%。

表 5　　单测度实验结果

分类器	单一测度	平均准确率
支持向量机	MFCC	87.18%
	梅尔能量	88.21%
	LPC	80.51%
	LPCC	81.54%
	小波统计特征	83.08%
逻辑回归	MFCC	80.00%
	梅尔能量	84.62%
	LPC	75.90%
	LPCC	77.44%
	小波统计特征	79.49%
随机森林	MFCC	85.13%
	梅尔能量	89.74%
	LPC	86.15%
	LPCC	81.03%
	小波统计特征	82.56%

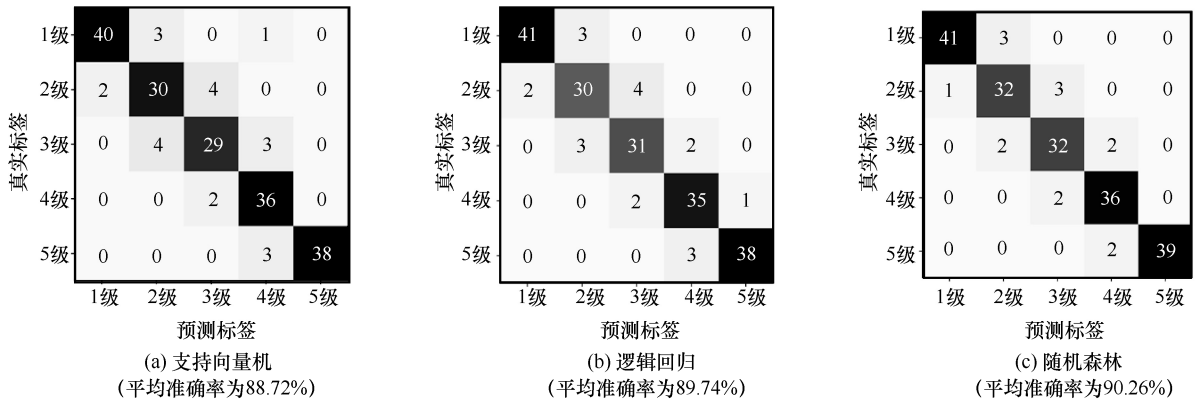
4.2　多测度融合实验

在利用上述 3 种分类器进行多测度融合实验的过程中，每条受扰语音样本对应一个维度为 1×104 的测度向量。与单测度评估结果相比，多测度融合的评估性能有较明显的提升。多测度融合实验结果如图 14 所示。实验结果表明，3 种分类器中随机森林的效果最好，质量等级预测的平均准确率达到 90.26%。由于数据样本量较少，随机森林分类器的抗过拟合能力得以体现，与其他 2 种分类器相比评估过程较稳定，尤其是针对 1、5 这 2 个等级的错分程度相对较低，这也表明随机森林分类器可通过其较强的特征选择能力对受扰程度较高的语音数据进行损伤级的区分。

4.3　多模态融合实验

上述多测度融合实验结果验证了利用随机森林分类器完成质量等级评估的有效性。为进一步分析不同域特征的融合对深度学习模型评估性能的影响，本文设计了 2 个具有 5 层残差结构的神经网络作为特征融合模型，进行多模态融合质量评估实验，并与 VGG16、AlexNet 这 2 种模型进行性能对比。多模态融合实验结果如图 15 所示。

实验结果表明，本文所设计的残差结构的平均准确率达 90.77%，高于另外 2 种模型。与 Wang 等采用单一对数梅尔谱图特征的实验结果相比，评估准确率提高了约 3.269%。此外，相比于其他 2 种模型，残差结构对信息损伤级为 2 级的数据预测错误程度有所降低。值得一提的是，计算模型参数量时发现，本文使用的模型参数缩小至 AlexNet 的约 $\frac{1}{28}$。为进一步验证分类器及模型评估性能，本文分别计算了多测度融合分类器以及多模态融合网络模型对每一类受扰语音数

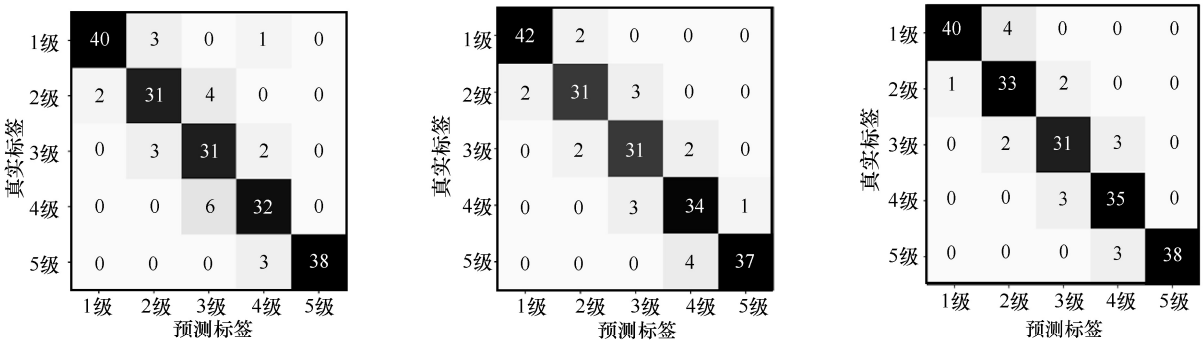


(a) 支持向量机
(平均准确率为88.72%)

(b) 逻辑回归
(平均准确率为89.74%)

(c) 随机森林
(平均准确率为90.26%)

图 14　多测度融合实验结果



(a) AlexNet
（平均准确率为88.21%）

(b) VGG16
（平均准确率为89.74%）

(c) 残差结构
（平均准确率为90.77%）

图 15　多模态融合实验结果

据等级分类的精确度和召回率，并得到 Macro-F1 分数（即每个损伤级 F1 分数的均值），分类器及模型 Macro-F1 分数如表 6 所示。从表 6 可以看出，随机森林分类器、残差结构分别在多测度融合模型、多模态融合模型中具有最高的 Macro-F1 分数，证明其对受扰语音数据具有更好的质量等级评估效果。

表 6　　分类器及模型 Macro-F1 分数

分类器	Macro-F1
支持向量机	0.88
逻辑回归	0.89
随机森林	**0.90**
AlexNet	0.88
VGG16	0.89
残差结构	**0.91**

分析图 14、图 15 中的混淆矩阵发现，错分的受扰语音数据均为相邻信息损伤级混淆，这是由于标注过程中语音受扰程度与损伤级是正相关的，受扰程度越高，损伤级越高，且相邻损伤级逐渐递进，使数据存在一定的相似性，从而导致该现象的产生。综合以上实验结果，本文所提出的多模态质量评估方法具有一定的可行性。

5　结束语

通信语音干扰效果评估技术对于保障语音通信设备的正常运行至关重要。为应对这一问题，本文总结了现有的研究成果，提出了一种基于多测度融合、多模态融合的受扰语音数据质量等级评估方法。首先，对受扰样本数据进行端点检测以提取语音内容，并将受扰语音与标准语音进行比较，计算特征差异，得到 5 种测度。然后，结合随机森林模型在多测度融合实验中验证了该评估方法的有效性。为进一步分析受扰语音数据的不同域特征对评估性能的影响，结合多模态融合技术，将样本的图域及测度域特征进行融合，与现有的研究成果相比较，该方法所得到的质量等级评估准确率提高了约 3.269%。

本文提出了一种较可行的受扰语音数据质量等级评估方法，但该方法仍存在许多值得改进之处。例如，本文的数据预处理及特征提取部分，将受扰语音信号的时域波形转换为图域、测度域等高维特征序列，并与标准语音计算特征距离，虽然一定程度上掌握了受扰数据的先验信息，但整个计算过程较复杂。此外，实验中过少的样本数据容易导致神经网络模型过拟合，且数据采集的实施较困难。因此，接下来的工作将针对评估流程进行改进，寻找更快捷、更有效的预处理方法，从而提高运算效率。同时，针对受扰语音数据采集困难这一问题，已经开展了部分后续工作，依据公开协议搭建了通信仿真系统，并模拟了不同干扰样式、干扰功率、信号调制方式等参数对语音数据的影响，一定程度上还原了真实场景下的采集过程，提升了受扰语音数据的多样性。

参考文献：

[1] 潘志丽, 张宏科, 张思东. 现代电子干扰理论与效能评估的研究[J]. 通信学报, 2003, 24(11): 40-45.
PAN Z L, ZHANG H K, ZHANG S D. Research on modern electronic jamming theory and efficiency evaluation[J]. Journal of China Institute of Communications, 2003, 24(11): 40-45.

[2] ITU-T. Mean opinion score（MOS）terminology[S]. 2003.

[3] RIX A W, BEERENDS J G, HOLLIER M P, et al. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs[C]//Proceedings of 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Piscataway: IEEE Press, 2002: 749-752.

[4] BEERENDS J G, SCHMIDMER C, BERGER J, et al. Perceptual objective listening quality assessment (POLQA), the third generation ITU-T standard for end-to-end speech quality measurement part I—temporal alignment[J]. Journal of the Audio Engineering Society, 2013, 61(6): 366-384.
[5] AFFONSO E T, ROSA R L, RODRÍGUEZ D Z. Speech quality assessment over lossy transmission channels using deep belief networks[J]. IEEE Signal Processing Letters, 2018, 25(1): 70-74.
[6] FU S W, TSAO Y, HWANG H T, et al. Quality-net: an end-to-end non-intrusive speech quality assessment model based on BLSTM[J]. arXiv Preprint, arXiv:1808.05344, 2018.
[7] FU S W, LIAO C F, TSAO Y. Learning with learned loss function: speech enhancement with quality-net to improve perceptual evaluation of speech quality[J]. IEEE Signal Processing Letters, 2020, 27: 26-30.
[8] RODRÍGUEZ D Z, PÍVARO G F, ROSA R L, et al. Improving a parametric model for speech quality assessment in wireless communication systems[C]//Proceedings of 2018 26th International Conference on Software, Telecommunications and Computer Networks. Piscataway: IEEE Press, 2018: 1-5.
[9] ITU-T. The E-model: a computational model for use in transmission planning: G.107[S]. 2002.
[10] LO C C, FU S W, HUANG W C, et al. MOSNet: deep learning based objective assessment for voice conversion[J]. arXiv Preprint, arXiv: 1904.08352, 2019.
[11] ZHANG L, ZHAO X L, LI X. Assessment of extreme communication environment with ultralow SNR: a benchmark[J]. IEEE Access, 2021, 9: 45400-45406.
[12] WANG S, LIN Y, HAO M, et al. Interference quality assessment of speech communication based on deep learning[J]. IEEE Transactions on Reliability, 2022, 71(2): 1011-1021.
[13] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 770-778.
[14] 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.
[15] 张璐琳, 张磊, 赵凌伟. 语音通信干扰效果评定规则: GJB4405B-2017[S]. 2017.
ZHANG L L, ZHANG L, ZHAO L W. Speech communication interference effect assessment rules: GJB4405B-2017 [S]. 2017.
[16] 傅恒丰. 语音通信干扰效果评估方法研究[D]. 哈尔滨: 哈尔滨工程大学, 2021.
FU H F. Research on the evaluation method of speech communication interference effect[D]. Harbin: Harbin Engineering University, 2021.
[17] DUBEY R K, KUMAR A. Non-intrusive objective speech quality assessment using a combination of MFCC, PLP and LSF features[C]//Proceedings of 2013 International Conference on Signal Processing and Communication. Piscataway: IEEE Press, 2014: 297-302.
[18] 张文克. 融合 LPCC 和 MFCC 特征参数的语音识别技术的研究[D]. 湘潭: 湘潭大学, 2016.
ZHANG W K. The research of fusion LPCC and MFCC feature parameters in speech recognition technology[D]. Xiangtan: Xiangtan University, 2016.
[19] KEERTHANA Y M, REDDY M K, RAO K S. CWT-based approach for epoch extraction from telephone quality speech[J]. IEEE Signal Processing Letters, 2019, 26(8): 1107-1111.
[20] 杨路飞, 章新华, 吴秉坤, 等. 基于 MFCC 特征的被动水声目标深度学习分类方法[J]. 舰船科学技术, 2020, 42(19): 129-133.
YANG L F, ZHANG X H, WU B K, et al. Research on the classification method of passive acoustic target depth learning based on MFCC[J]. Ship Science and Technology, 2020, 42(19): 129-133.
[21] MUDA L, BEGAM M, ELAMVAZUTHI I. Voice recognition algorithms using Mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques[J]. arXiv Preprint, arXiv: 1003.4083, 2010.
[22] TIN K H. The random subspace method for constructing decision forests[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(8): 832-844.
[23] MITTAG G, NADERI B, CHEHADI A, et al. NISQA: a deep CNN-self-attention model for multidimensional speech quality prediction with crowdsourced datasets[J]. arXiv Preprint, arXiv: 2104.09494, 2021.
[24] YU M, ZHANG C, XU Y, et al. MetricNet: towards improved modeling for non-intrusive speech quality assessment[J]. arXiv Preprint, arXiv: 2104.01227, 2021.
[25] 张思成, 林云, 涂涯, 等. 基于轻量级深度神经网络的电磁信号调制识别技术[J]. 通信学报, 2020, 41(11): 12-21.
ZHANG S C, LIN Y, TU Y, et al. Electromagnetic signal modulation recognition technology based on lightweight deep neural network[J]. Journal on Communications, 2020, 41(11): 12-21.

[作者简介]

林云（1980− ），男，黑龙江哈尔滨人，哈尔滨工程大学教授、博士生导师，主要研究方向为智能无线电技术、人工智能和机器学习、大数据分析与挖掘、软件和认知无线电、信息安全与对抗、智能信息处理。

徐怀韬（1998− ），男，江西南昌人，哈尔滨工程大学硕士生，主要研究方向为通信干扰语音质量评估。

王森（1994− ），男，吉林四平人，哈尔滨工程大学博士生，主要研究方向为干扰评估、无线通信及信号处理。

张思成（1996− ），男，山东临沂人，哈尔滨工程大学博士生，主要研究方向为基于深度学习的智能电磁信号处理。

庄龙（1998− ），男，江苏徐州人，安徽大学硕士生，主要研究方向为雷达信号处理和计算机视觉。