噪声鲁棒的蒙古语语音数据增广模型结构

马志强[1,2*], 孙佳琦[1], 李晋益[1], 王嘉泰[1]
[1] 内蒙古工业大学数据科学与应用学院, 呼和浩特, 010080
[2] 内蒙古自治区基于大数据的软件服务工程技术研究中心, 呼和浩特, 010080
mzq_bim@imut.edu.cn

摘要

蒙古语语料库中语音多样性匮乏，虽然花费人力和经费收集数据在一定程度上能够增加语音的数量，但整个过程需要耗费大量的时间。数据增广能够解决这种数据匮乏问题，但数据增广模型的训练数据包含的环境噪声无法控制，导致增广语音中存在背景噪声。本文提出一种TTS和语音增强相结合的语音数据增广方法，以语音的频谱图为基础，从频域和时域两个维度进行语音增强。通过多组实验证明，蒙古语增广语音的合格率达到70%，增广语音的CBAK和COVL分别下降了0.66和0.81，WER和SER下降了2.75%和2.05%。

关键词： 语音增强；数据增广；噪声鲁棒性；蒙古语

Noise robust Mongolian speech data augmentation model structure

Ma Zhiqiang[1,2*], Sun Jiaqi[1], Li jinyi[1], Wang Jiatai[1]
[1] College of Data Science and Application Inner Mongolia University of Technology, Huhhot, 010000
[2] Inner Mongolia Autonomous Region Software Service Engineering Technology Research Center Based on Big Data, Huhhot, 010000
mzq_bim@imut.edu.cn

Abstract

There is a lack of phonetic diversity in Mongolian corpus. Although manpower and funds spent on data collection can increase the number of phonetic sounds to some extent, the whole process needs a lot of time. Data augmentation can solve the problem of data scarcity, but the environmental noise contained in the training data of the data augmentation model cannot be controlled, resulting in background noise in the augmentation speech. In this paper, a speech data augmentation method combining TTS and speech enhancement is proposed. Based on the speech spectrum graph, speech enhancement is carried out from two dimensions: frequency domain and time domain. Multiple experiments show that the qualified rate of Mongolian augmented speech reaches 70%, the CBAK and COVL of augmented speech decrease by 0.66 and 0.81, and WER and SER decrease by 2.75% and 2.05%, respectively.

Keywords: Speech enhancement , Data augmentation , Noise robustness , Mongolian

第二十二届中国计算语言学大会论文集，第155页-第163页，哈尔滨，中国，2023年8月3日至5日。

1 引言

蒙古语作为一种低资源语言，在深度学习任务中其训练数据十分有限。为了扩充蒙古语音数据集的规模并提高训练数据的数量，可以采用语音数据增广。语音数据增广可以通过原始语音数据生成新的语音样本，以有效地解决语音数据匮乏和多样性不足的问题。TTS是一种常用的语音数据增广方法，可以提高蒙古语语料库中语音说话人的多样性。然而，与在受控条件下录制蒙古语语音语料库不同，使用TTS语音数据增广方法常需要依赖于域外语音，以引入域外说话人的特征。这可能会导致增广语音受到无法控制的混响和环境噪声的影响。此外，在蒙古语发音中，存在较为复杂的发音现象，比如气音、元音的颤音、非单元音辅音和元音拼接等。这些现象容易导致语音数据增广模型过度拟合域外语音中的噪声，从而导致增广后的语音失真。因此，本文旨在解决语音数据增广过程中的噪音鲁棒性问题，将SE方法应用于基于语音合成(Text to Speech，TTS)的增广框架中，以提高增广语音数据的质量。

语音增强[1](Speech Enhancement，SE)旨在通过减少背景噪声来提高语音质量，是语音数字信号处理中的一个重要环节，其主要目标是提高语音的清晰度，进而提高下游任务对噪声的鲁棒性。语音增强主要分为信号处理和深度神经网络两种方法，标准的信号处理方法以频谱减法和维纳滤波为主，频谱减法[2]通过减去非语音活动期间计算的频谱噪声偏差来抑制语音中的静音噪声，然后衰弱减法后的残余噪声。维纳滤波[3]使用均方误差去估计纯净信号谱，通过估计线性滤波器对不同频段的噪声进行不同程度的抑制。但基于信号处理的SE方法的应用范围有限，如噪声具有一定的平稳性，干净语音和噪声不相关等，这限制了算法的性能和应用场景。基于深度神经网络的语音增强具有更好的性能和更加丰富的应用场景，该方法主要用于直接重建干净语音[4]或从噪声信号中估计掩码[5]，采用监督的方式进行去噪，将倒谱或频谱表示转换为波形表示，相比提取梅尔频率倒谱系数能够保留更加完整的语音信息。受深度神经网络的启发，将SE与语音数据增广模型相结合提高语音增广模型的噪声鲁棒性，消除域外语音中引入的噪音，提高增广语音的质量。

目前，将语音增广作为下游任务的SE研究并不多，本文将SE方法加入到基于TTS的蒙古语语音增广模型中，该方法结合了蒙古语的发音特点和噪声对增广语音的影响，以提高增广模型的噪音鲁棒性。

2 相关工作

基于深度学习的SE方法主要包括DNN、CNN和LSTM等网络。Karjol等人[6]提出基于多个DNN的增强方法，利用一个门控网络提供权重来组合多个输出，该方法有效的提高了语音质量感知评估(Perceptual Evaluation of Speech quality，PESQ)。Y Zhao等人[7]通过在损失函数中加入语音可理解度的度量进行DNN的扩展，在多种信噪比和噪声类型下提高语音清晰度。Bagchi等人[8]也进行DNN模型的扩展，将模拟损失与传统损失结合起来训练语音增强模型。但在基于DNN的SE网络需要大量参数，此外在低信噪比条件下还会导致增强语音恶化。

为了有效的学习语音时间信息，SE方法开始由DNN向循环神经网络(Recurrent Neural Network，RNN)和卷积神经网络(Convolutional Neural Network，CNN)转变。Maas等人[9]使用RNN增强带噪声语音的MFCC特征，利用立体噪声和干净音频特征进行训练，预测噪声语音的干净特征。Gao等人[10]提出了一种LSTM的渐进式学习框架，将输入和中间目标的估计进行拼接后学习下一个目标，这种方法可以充分利用多个学习目标的信息，缓解信息丢失，提高低信噪比环境下的SE性能。此外，采用RNN还能够实现非平稳加性噪声[11]、混响[12]和多通道噪声语音[13]的语音去噪。基于CNN的语音增强能够处理语音的局部时间信息，可以有效地分离噪声信号中的语音和噪声信息，在频谱域和波形域都能完成语音增强。Kinoshita等人[14]受语音分离的时域卷积网络[15]启发，采用CNN在时域上进行掩蔽估计去噪，完成语音增强。P Plantinga等人[16]使用残差网络去除残余噪声重构输入信号，借助ResNet频谱映射架构来提高语音增强性能，将该方法与下游模型联合训练，实现下游任务中模型的噪声鲁棒性。

通过RNN和CNN神经网络进行语音增强可以捕获语音信号的时序特征，通过多次迭代学习到更多的特征以此来提高语音增强的准确性，但RNN和CNN网络模型存在某些参数不具备可解释性，因此网络结构上的优化比较困难，导致模型训练时可控性较差。将SE方法加入到基于TTS的蒙古语语音增广模型中可以在模型训练时对所需的性能进行调整和优化，从而更好地控制优化模型得到输出语音质量更好的语音数据。

第二十二届中国计算语言学大会论文集，第155页-第163页，哈尔滨，中国，2023年8月3日至5日。

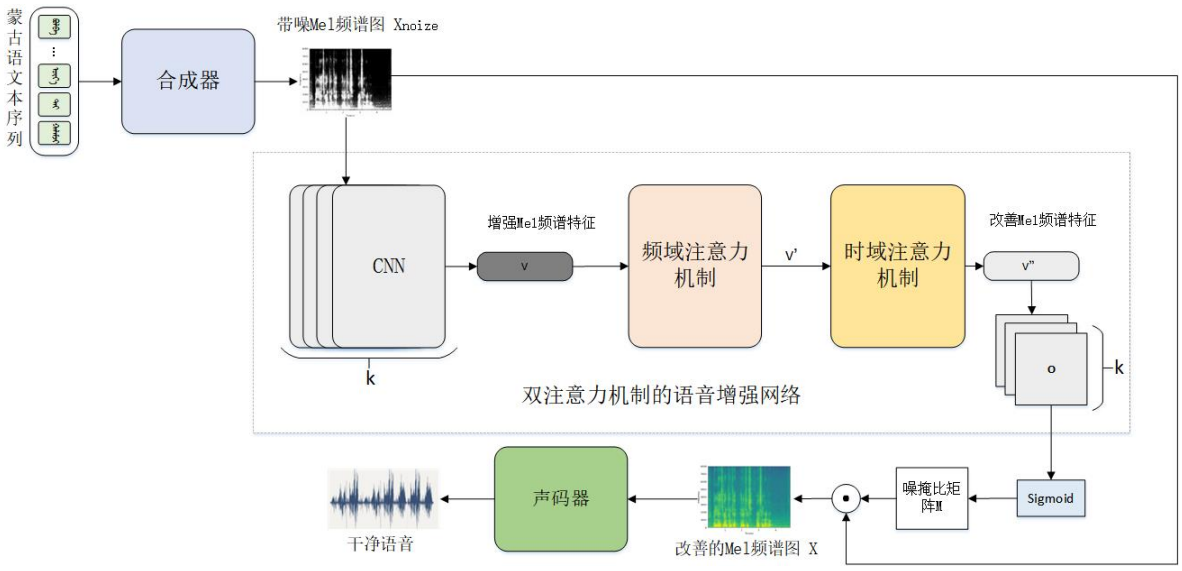

Figure 1: 噪声鲁棒的语音生成器架构图

3 方法

3.1 噪声鲁棒的语音生成器

为了提高基于TTS增广模型对噪声的鲁棒性，在语音生成器中加入SE网络，噪声鲁棒的语音生成器架构如图1所示，合成器将蒙古语文本转换为Mel频谱图，因为Mel频谱图能够同时表示语音时域和频域的信息，SE网络利用这一特性使用双注意力机制进行增强，图1中虚线部分为双注意力机制的语音增强网络(Double Attention Speech Enhancement，D-Attention-SE)。$X_{noize} = R^{T*F}$表示域外语音的Mel频谱向量，T表示时间维度，F表示频率维度。使用k层CNN对X_{noise}增强，输出增强后的Mel频谱特征V，然后利用频域注意力机制从频域维度增强得到特征V'；将V'输入到时域注意力机制完成时域增强，得到改善后的增强频谱特征V"。双注意力机制语音增强网络将改善后的特征V"细化为$O = R^{T_k F_k}$,其中k表示k层CNN改善的频谱特征。最后，细化特征O通过Sigmoid计算输出噪掩比矩阵M，将该矩阵与原始频谱图进行点积，过滤掉被噪声破坏的Mel频谱图，输出最终改善的Mel频谱图X，声码器利用X将Mel频谱图转换为语音，实现频域和时域维度的语音增强。

3.2 双注意力机制网络

双注意力机制包括频域注意力机制和时域注意力机制，其网络结构如图2所示，图中上半部分为频域注意力机制，下半部分为时域注意力机制。该网络结构的计算流程为$V' = a_F \odot V$，$V'' = a_T \odot V'$，其中a_F和a_T分别表示频率注意和时域注意的权重，增强Mel频谱特征V与a_F点积，得到频域增强特征V'；V'与a_T点乘，得到时域增强特征V"，实现两个维度的增强改善。在频域注意力机制中，首先从输入特征的信道维度进行最大池化和平均池化，两者的池化结果进行拼接后，通过时间池化进行最大池化和平均池化。最后通过卷积运算和Sigmoid激活函数计算得到频域注意力权重a_F，将a_F与V进行点积，得到频率增强特征V'。时域注意力机制与频域注意力机制类似，对频域注意力机制增强的特征V'进行信道池化和频域池化，利用卷积层和Sigmoid激活函数计算得到时域注意力权重a_t，通过广播后扩展维度，与V'点积得到时域增强特征，实现频域和时域两个维度的Mel频谱增强。

3.3 模型训练

两阶段的语音增强网络使用均方差(Mean Square Error，MSE)作为的损失函数，其损失函数如公式(1)所示。

$$Loss = \frac{\sum \|X_{noise} \odot M - X_{clean}\|^2}{n} \quad (1)$$

其中,X_{noise}表示带噪Mel频谱图，X_{clean}表去噪的Mel频谱图，n表示Mel的频率转换尺度，M表示Mel频谱的噪掩比矩阵。双注意力机制的SE网络在训练过程主要更新频域注意力

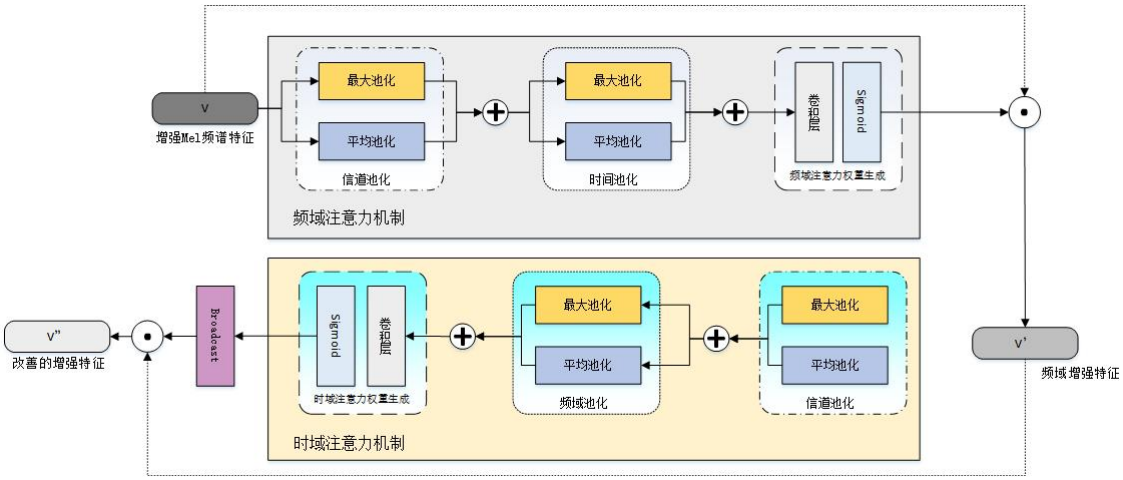

Figure 2: 两阶段的语音增强网络结构图

机制和时域注意力机制的权重，通过输入增强Mel频谱特征，计算出频域注意力权重a_F和时域注意力权重a_T，利用a_F和a_T求出改善后的Mel频谱特征V"，该训练算法如表1所列。

Table 1: 模型训练算法

输入： 带噪语音Mel频谱图X_{noize}；干净语音Mel频谱图X_{clean}；训练轮数$epoch$；
输出： 噪掩比M；改善的Mel频谱图X；

For $i = 0$; $epoch$ **do**
  $V = Train_{cnn}\left(X_{noise}, X_{clean}\right)$ //增强网络CNN训练
  $a_F = F\_Attention(V)$ // 为频域注意力权重
  $V' = V \bigodot a_F$
  $a_T = T\_Attention\left(V'\right)$ // 为时域注意力权重
  $V'' = V' \odot a_T$
  $\mathrm{H} = Sigmoid\left(V''\right)$
  If $X_{noise} \odot M \neq X_{clean}$ **:**
    $a_F = backpropogation(Loss)$
    $a_T = backpropogation(Loss)$
  Else:
    Return X

4 实验

4.1 实验设置

在语音增强网络的训练过程中，将Audio Set[17]数据集中的噪声加入到蒙古语语音数据集IMUT-MC1中，得到干净语音所对应不同环境噪音下的语音，用于训练语音增强网络。Audio Set中包含五类真实世界中的噪声环境，包括人类噪声、动物噪声、自然噪声、音乐噪声和事物噪声，训练数据如表2所列。

Table 2: 语音增强网络训练数据

噪声类别	文本句子	词数	人数	性别	平均词个数	平均时长	总时长
5类	1255条	2237	1	女	6	12秒	4.1h

在两阶段的语音增强网络的训练过程中，语音的采样率为16kHz，将带噪语音和干净语音归一化到[-1,1]，从归一化的结果中提取帧。使用Adam优化器对随机梯度下降进行优化，批量

处理大小为4条语音，为了匹配最长语音的大小，将批处理中较短的语音用0进行填充。此外，训练轮数为500轮，学习率为2e-4。

4.2 评价指标

在噪声鲁棒的语音增广模型实验中，从增广语音的质量、自然度和消除噪声的效果三个角度进行模型评价。使用平均意见得分(Mean Opinion Score，MOS)和生成合格率(Generated Pass Rate，GPR)对语音质量进行评价。该方法的指标包括缺字句数(Missing Words，MW)、多字句数(Insert Words，IW)、字序错误(Word Sequence Error，WSE)、发音错误(Pronunciation Error，PE)、噪音句数(Noise Sentence，NS)和无域外说话人句数(Without Foreign Speakers，WFS)。其中，MW表示增广语音存在字符缺失的句数；IW表示增广语音插入了非对应文本词的句数；WSE表示增广语音发音顺序错误的句数；PE表示增广语音发音错误的句数；NS表示增广语音存在噪音的句数；WFS表示生成语音中没有加入域外说话人的语音。生成合格率(Generated Pass Rate，GPR)：GPR指人工评价中没有错误的句子数占总句数(Total Sentences，TS)的百分比，见式(2)。

$$GPR = \left(1 - \frac{MW+IW+WSE+PE+NS+WFS}{TS}\right) \times 100\% \quad (2)$$

使用梅尔频谱失真(Mel Cepstral Distortion，MCD)对语音的自然度进行评价。使用背景干扰综合测度(The Composite Measure for Background Interferences，CBAK)和总体综合测度[18](The Overall Composite Measure，COVL)评价语音中噪声的抑制程度和效果，其取值范围为1到5。此外，使用增广说话人的语音训练语音识别模型，通过语音识别的词错误率（Word Error Rate，WER）和句错误率(Sentence Error Rate, SER)来评价提出方法的有效性。

4.3 实验结果与分析

4.3.1 有效性实验

为了验证语音增广模型的噪声鲁棒性，使用五种不同类型的噪声作为域外语音进行增广，从每个噪声类型中选择10个不同说话人，每个说话人生成相同的200条语音，增广数据集详情如表3所列。IMUT-MC1-N1、IMUT-MC1-N2、IMUT-MC1-N3、IMUT-MC1-N4和IMUT-MC1-N5表示域外语音中噪声类型数不同的增广数据集，其中IMUT-MC1-N1 ⊆ IMUT-MC1-N2 ⊆ IMUT-MC1-N3 ⊆ IMUT-MC1-N4 ⊆ IMUT-MC1-N5，IMUT-MC1-N5中包含了人类噪声、动物噪声、自然噪声、音乐噪声和事物噪声五类噪声，不同噪声语音增广所对应的文本标签相同，没有相同说话人。

Table 3: 地区数不同的增广数据集

名称	噪声类别数	说话人数	总句数（万）	总时长（h）
IMUT-MC1-N1	1	10	0.2	2.4
IMUT-MC1-N2	2	20	0.4	4.8
IMUT-MC1-N3	3	30	0.6	5.2
IMUT-MC1-N4	4	40	0.8	7.6
IMUT-MC1-N5	5	50	1.0	10.0

使用带有不同噪声的域外语音进行语音增广，对IMUT-MC1-N5中五类噪声背景下的增广语音进行评价，每类噪声背景选择相同文本对应的200条增广语音，从语音的质量、自然度和噪声抑制程度来验证提出方法的有效性，增广语音的评价结果如表4所列。分析表3可知，五种噪声环境下增广语音的GPR都达到65%，增广语音的整体合格率达到70%；MOS均达到4.0以上，与静音环境下增广语音的MOS值接近，这表明在增广模型中引入语音增强，有效的提高了噪声环境下增广语音的质量。MCD的评价结果均小于19.20，Mel频谱失真低于与静音场景下增广语音的19.61，表明语音增强网络有效的提高了增广语音的自然度。此外，CBAK和COVL的值都达到2.6以上，相比域外语音中的1.5，CBAK和COVL有了明显的提示，表明语音增强网络显著的抑制了不同类型的背景噪声。

为了验证Double-Attention-SE针对噪声数据增广的鲁棒性，使用包含10位说话人的语音作为测试集，分别对不同噪声环境下增广数据集训练的声学模型进行测试，测试集从五类噪声中挑选文本相同的200条语音。使用不同噪声环境下的增广数据集构建的声学模型测试结果如

Table 4: 噪声鲁棒的增广语音评价

噪声类型	质量评价		自然度评价	噪声抑制度评价	
	GPR	MOS	MCD	CBAK	COVL
人类噪声	65%	4.10	19.05	2.77	2.71
动物噪声	75%	4.20	18.67	2.82	2.79
自然噪声	70%	4.13	18.23	2.69	2.64
音乐噪声	65%	4.08	19.11	2.66	2.61
事物噪声	75%	4.22	18.58	2.88	2.85

表5所列，在训练数据中加入不同噪声环境下的增广语音，声学模型的I、D和S均呈现下降趋势，而且下降幅度逐渐增大，其中I、D和S对比原始训练数据分别下降2.36%、3.51%和1.33%，这表明加入Double-Attention-SE的增广语音数据能够有效降低声学模型的插入错误和删除错误。此外，WER和SER分别下降7.20%和6.88%，当训练集中包含五类不同噪声环境的增广语音时，IMUT-MC1-N5数据集训练的声学模型在测试数据上的WER和SER最低，这表明加入不同噪声环境下的增广语音作为训练数据没有降低语音识别的准确率，反而因为域外说话人的增加提高了语音识别的识别准确率。

Table 5: 语音识别实验结果

噪声类型	I	D	S	WER	SER
IMUT-MC1	7.29%	12.26%	6.99%	26.54%	31.04%
IMUT-MC1-N1	7.17%	12.14%	6.56%	25.87%	29.91%
IMUT-MC1-N2	6.93%	11.89%	6.42%	25.24%	29.72%
IMUT-MC1-N3	6.55%	11.01%	6.10%	23.66%	28.15%
IMUT-MC1-N4	5.32%	10.16%	5.91%	21.39%	25.88%
IMUT-MC1-N5	4.93%	8.75%	5.66%	19.34%	24.16

4.3.2 消融实验

在基于TTS语音增广模型中加入语音增强网络消除域外语音引入的噪声，为了避免增强过程中破坏域外说话人特征，语音增强网络从频域和时域进行增强，准确消除噪声的同时不破坏说话人特征。增强网络以单个频域注意力机制Single-Attention-SE为基础，加入频域注意力机制，完成两阶段语音增强网络模型Double-Attention-SE的构建。消融实验从增广语音的质量、自然度、噪声抑制程度和语音识别四个方面进行评估，验证两阶段语音增强的去噪效果。该实验选取不同方法增广的语音进行评价，挑选与训练集中相同说话人的干净语音作为测试集进行语音识别测试，消融实验结果如表6所列。分析表6可知，Single-Attention-SE方法与None-SE相比，GPR和MOS分别提高了2%和0.05；MCD降低了0.13；CBAK和COVL分别下降了0.42和0.25，各项指标的改善较小，表明Single-Attention-SE的有效性不足。虽然加入语音增强网络提高了增广语音的质量，但语音识别评价的WER和SER升高了1.78%和1.77%，分析表明语音增强网络破坏了域外语音中的说话人特征，导致语音识别准确率下降。Double-Attention-SE方法与None-SE相比，GPR和MOS分别提高了5%和0.10；MCD降低了1.95；CBAK和COVL分别下降了0.66和0.81。此外，WER和SER下降了2.75%和2.05%，这表明Double-Attention-SE语音增强网络在提高增广语音质量的过程同时没有破坏域外说话人特征。

Table 6: 消融实验结果

方法	GPR	MOS	MCD	CBAK	COVL	WER	SER
None-SE	65%	4.01%	20.68%	3.54%	3.66%	22.09%	26.21%
Single-Attention-SE	67%	4.06%	20.55%	3.12%	3.41%	23.87%	27.98%
Double-Attention-SE	70%	4.11%	18.73%	2.88%	2.85%	19.34%	24.16

4.3.3 对比实验

对比实验将提出的Double-Attention-SE增强方法与基于深度学习的语音增强方法进行对比，包括基于DNN、RNN、LSTM和ResNet的语音增强方法，同时与时单通道域语音增强方法Time-Domain-SE进行对比，将上述方法加入到基于TTS的语音增广模型中，采用相同的文本和说话人相同的噪声语音进行增广，对增广语音进行评价验证所提方法的先进性，对比结果如表7所列。分析表7可知，所有方法对应增广语音的GPR均达到60%以上，其中Double-Attention-SE最高为70%，表明提高方法有效的提高了增广语音的合格率。此外，Double-Attention-SE取得了该组对比实验最低的MOS、MCD、CBAK和COVL，均优于其它方法。通过对比发现，加入Double-Attention-SE 方法的增广语音CBAK和COVL降低幅度较小，均保持在1.00之内，但WER和SER的有较为明显的下降，与Multiple-DNN-SE方法的差距最大，分别为10.98%和9.86%。表明不同的增强方法都可以有效的抑制增广语音的噪声，但在一定程度上会破坏说话人特征，导致语音识别的准确率降低，Double-Attention-SE在增强过程中对说话人特征的破坏最低，相比其它方法取得最低WER和SER。

Table 7: 语音增强对比实验结果

方法	GPR	MOS	MCD	CBAK	COVL	WER	SER
Single-Attention-SE	60%	3.98%	21.12%	3.72%	3.81%	30.32%	34.02%
Double-Attention-SE	62%	4.01%	21.05%	3.46%	3.59%	25.65%	29.88%
Single-Attention-SE	65%	4.05%	20.88%	3.41%	3.53%	24.87%	28.35%
Double-Attention-SE	68%	4.07%	20.19%	3.32%	3.45%	23.92%	28.09%
Single-Attention-SE	68%	4.09%	19.78%	3.11%	3.24%	21.66%	25.79%
Double-Attention-SE	70%	4.11%	18.73%	2.88%	2.85%	19.34%	24.16%

4.3.4 分析实验

为了验证Double-Attention-SE增强网络的适应性，分析实验分别使用Aishell-1中文数据集，Librispeech-clean英文数据集，IMUT-MC1蒙古语数据集，JSUT日语数据集和Ruslan俄语数据集五类不同语言进行增广，将五类噪声分别加入到三种数据集中，从各类增广语音中选取200条语音数据进行评价。使用GPR和MOS评价不同语言增广语言的质量，使用MCD评价增广语音的自然度，CBAK和COVL验证增广模型的噪声鲁棒性。汉语、英语、蒙古语、日语和俄语的分析实验结果如表7所列。分析表7可知，在语音增广模型中加入Double-Attention-SE进行汉语、英语、蒙古语、日语和俄语增广，增广语音的合格率GPR均达到65%以上；MOS均达到4.0以上，这表明该模型针对不同语言进行增广时，语音的质量均能达标。此外，分析表8可知汉语、英语、日语、俄语和蒙古语增广语音的MCD、CBAK和COVL没有明显差距，蒙古语与汉语增广语音的差距最大，为0.29、0.13和0.27，这表明使用Double-Attention-SE进行噪声鲁棒的语音增广方法同样适用于汉语、英语、日语和俄语，从而证明该方法对不同语言具有适应性。

Table 8: 多语言实验结果

评价指标	汉语 Aishell-1	英语 Librispeech-clean	蒙古语 IMUT-MC1	日语 JSUT	俄语 Ruslan
GPR	67%	68%	70%	66%	69%
MOS	4.01	4.09	4.11	4.02	4.07
MCD	19.02	18.98	18.73	18.68	18.94
CBAK	3.01	2.91	2.88	2.93	2.91
COVL	3.12	2.98	2.85	2.74	2.88

5 结论

本文针对基于TTS蒙古语语音数据增广模型的噪声鲁棒性展开研究，提出基于频谱特征的语音增强单元，该单元从时域和频域两个维度进行去噪，消除域外语音中引入的噪声，降低语

音增强对说话人特征的影响。实验结果表明加入语音增强单元后，蒙古语增广语音的合格率达到70%，增广语音的CBAK和COVL分别下降了0.66和0.81，WER和SER下降了2.75%和2.%，这表明基于频谱特征的语音增强网络消除了域外语音中的噪声，提高了蒙古语语音数据增广模型的噪声鲁棒性，提升了增广语音的合格率。

References

[1] Asri Rizki Yuliani et al. "Speech Enhancement Using Deep Learning Methods: A Review". In: *Jurnal Elektronika dan Telekomunikasi* 21.1 (2021), pp. 19–26.

[2] Steven Boll. "Suppression of acoustic noise in speech using spectral subtraction". In: *IEEE Transactions on acoustics, speech, and signal processing* 27.2 (1979), pp. 113–120.

[3] Tim Van den Bogaert et al. "Speech enhancement with multichannel Wiener filter techniques in multimicrophone binaural hearing aids". In: *The Journal of the Acoustical Society of America* 125.1 (2009), pp. 360–371.

[4] Yong Xu et al. "An experimental study on speech enhancement based on deep neural networks". In: *IEEE Signal processing letters* 21.1 (2013), pp. 65–68.

[5] Arun Narayanan and DeLiang Wang. "Ideal ratio mask estimation using deep neural networks for robust speech recognition". In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE. 2013, pp. 7092–7096.

[6] Pavan Karjol, M Ajay Kumar, and Prasanta Kumar Ghosh. "Speech enhancement using multiple deep neural networks". In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2018, pp. 5049–5052.

[7] Yan Zhao et al. "Perceptually guided speech enhancement using deep neural networks". In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2018, pp. 5074–5078.

[8] Deblin Bagchi et al. "Spectral feature mapping with mimic loss for robust speech recognition". In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2018, pp. 5609–5613.

[9] Andrew Maas et al. "Recurrent neural networks for noise reduction in robust ASR". In: (2012).

[10] Tian Gao et al. "Densely connected progressive learning for lstm-based speech enhancement". In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2018, pp. 5054–5058.

[11] Martin Wöllmer et al. "Feature enhancement by bidirectional LSTM networks for conversational speech recognition in highly non-stationary noise". In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE. 2013, pp. 6822–6826.

[12] Felix Weninger et al. "The Munich feature enhancement approach to the 2nd CHiME challenge using BLSTM recurrent neural networks". In: *Proceedings of the 2nd CHiME workshop on machine listening in multisource environments*. 2013, pp. 86–90.

[13] Xiaofei Li and Radu Horaud. "Multichannel speech enhancement based on time-frequency masking using subband long short-term memory". In: *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE. 2019, pp. 298–302.

[14] Keisuke Kinoshita et al. "Improving noise robust automatic speech recognition with single-channel time-domain enhancement network". In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2020, pp. 7009–7013.

[15] Yi Luo and Nima Mesgarani. "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation". In: *IEEE/ACM transactions on audio, speech, and language processing* 27.8 (2019), pp. 1256–1266.

[16] Peter Plantinga, Deblin Bagchi, and Eric Fosler-Lussier. “An exploration of mimic architectures for residual network based spectral mapping”. In: *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE. 2018, pp. 550–557.

[17] Jort F Gemmeke et al. “Audio set: An ontology and human-labeled dataset for audio events”. In: *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE. 2017, pp. 776–780.

[18] Neil Shah, Hemant A Patil, and Meet H Soni. “Time-frequency mask-based speech enhancement using convolutional generative adversarial network”. In: *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE. 2018, pp. 1246–1251.