基于Transformer的单通道语音增强模型综述

范君怡[1],杨吉斌[1],张雄伟[1],郑昌艳[2]
1.陆军工程大学 指挥控制工程学院,南京 210007
2.火箭军士官学校 测试控制系,山东 潍坊 262500

摘 要:深度学习可以有效地解决带噪语音信号与干净语音信号之间复杂的映射问题,改善单通道语音增强的质量,但是增强语音的质量依然不理想。Transformer在语音信号处理领域中已得到了广泛应用,由于集成了多头注意力机制,可以更好地关注语音的长时相关性,该模型可以进一步改善语音增强效果。基于此,回顾了基于深度学习的语音增强模型,归纳了Transformer模型及其内部结构,从不同实现结构出发对基于Transformer的语音增强模型分类,详细分析了几种实例模型。并在常用数据集上对比了Transformer单通道语音增强的性能,分析了它们的优缺点。对相关研究工作的不足进行了总结,并对未来发展进行展望。

关键词:语音增强;深度学习;Transformer;单通道;多头注意力机制
文献标志码:A **中图分类号:**TP183;TN912.35 doi:10.3778/j.issn.1002-8331.2201-0371

Research on Transformer-Based Single-Channel Speech Enhancement

FAN Junyi[1], YANG Jibin[1], ZHANG Xiongwei[1], ZHENG Changyan[2]
1.College of Command and Control Engineering, Army Engineering University, Nanjing 210007, China
2.Department of Test Control, High-Tech Institute, Weifang, Shandong 262500, China

Abstract: Deep learning can effectively solve the complex mapping problem between noisy speech signals and clean speech signals to improve the quality of single-channel speech enhancement, but the enhancement effect based on network models is not satisfactory. Transformer has been widely used in the field of speech signal processing due to the fact that it integrates multi-headed attention mechanism and can focus on the long-term correlation existing in speech. Based on this, deep learning-based speech enhancement models are reviewed, the Transformer model and its internal structure are summarized, Transformer-based speech enhancement models are classified in terms of different implementation structures, and several example models are analyzed in detail. Furthermore, the performance of Transformer-based single-channel speech enhancement is compared on the public datasets, and their advantages and disadvantages are analyzed. The shortcomings of the related research work are summarized and future developments are envisaged.

Key words: speech enhancement; deep learning; Transformer; single-channel; multi-attention mechanism

语音增强技术是指从带噪语音信号中恢复出尽可能干净的语音信号,提高噪声条件下语音的质量和可懂度,在学术界和工业领域中得到了广泛的研究和应用[1-3]。和多通道语音增强相比,单通道语音增强具有硬件成本低,能耗小的优势,但由于缺失声源信息和噪声的空间信息,研究更具挑战性。

和传统的单通道语音增强技术[4-10]相比,基于深度神经网络的语音增强技术,不需要对数据设置额外假设条件。通过挖掘大规模数据的内在关联,能够准确实现语音和噪声的估计,在平稳噪声环境下取得了较大的进展。目前,各种网络模型都得到了应用,如深度神经网络(deep neural network,DNN)[11-13]、递归神经网络(recurrent neural network,RNN)[14-15]、卷积神经网络(convolutional neural network,CNN)[16-18]、U-net神经网络[19-22]等。Wang等人在文献[11-12]中率先将深度学习用于语音增强任务的研究。他们使用DNN估计出理想二值掩模(ideal binary mask,IBM)值,将带噪语音信号直接映射到干净语音信号,但DNN存在参数量大、无法利用上下

基金项目:国家自然科学基金(62071484)。
作者简介:范君怡(1997—),女,硕士研究生,研究方向为智能语音处理;杨吉斌(1978—),通信作者,男,博士,副教授,研究方向为智能语音处理、信息内容安全,E-mail:yjbice@sina.com;张雄伟(1965—),男,博士,教授,研究方向为智能语音处理、信息内容安全;郑昌艳(1990—),女,博士,讲师,研究方向为智能语音处理。
收稿日期:2022-01-24 **修回日期:**2022-03-18 **文章编号:**1002-8331(2022)12-0025-12

信息等问题。Weninger等人在文献[14]中利用RNN对上下文的特征信息进行建模,在文献[15]中进一步采用长短期记忆人工神经网络(long short-term memory,LSTM)对语音信号进行近似估计,但RNN存在训练时间长、网络规模大、难以实现并行化处理等问题。Park等人在文献[16]中提出了基于CNN的增强模型,通过输入前几帧的带噪语音信号来预测当前干净语音信号,很好地利用了时间相关性。与RNN相比,这种基于时域的卷积网络具有更小的网络规模、更短的训练时间,但CNN存在感受野受限,上下文建模能力弱等问题。为了缓解传统CNN模型的问题,Rethage等人在文献[17]中采用扩张卷积神经网络来提高语音增强的性能。张天骐等人在文献[18]中采用门控机制和扩张卷积神经网络,在增加感受野的基础上,门控机制可以较好地处理上下文特征信息。

近两年,Transformer因可并行、能处理长时间依赖的优势,在语音识别、自然语言处理、图像分割等领域取得了很大成功[23-25],然而由于其采用的解码器(decoder)结构需要同时使用上下文特征信息,不适用于实时处理,在语音增强方面的工作相对较少[26-29]。为了更好地利用Transformer模型提升单通道语音增强的性能,挖掘其在单通道语音增强方面的应用潜力,本文在归纳基于深度学习的语音增强框架基础之上,对基于Transformer的语音增强模型进行了系统梳理,根据Transformer集成的结构不同分类介绍了基于Transformer的语音增强模型,并综合对比了它们的性能。最后,对基于Transformer的语音增强发展方向进行了展望。

1 基于深度学习的单通道语音增强模型

1.1 单通道语音增强数学模型

单通道语音增强中,带噪语音信号可由公式(1)给出:

$$y(n)=x(n)+d(n) \tag{1}$$

其中,$x(n)$表示干净语音信号,$d(n)$表示加性噪声信号,$y(n)$表示带噪语音信号。加性噪声是对语音信号质量影响最为严重的噪声之一[30],混响噪声等非加性噪声可以通过某些方式将其转换为加性噪声。

语音增强需要从带噪语音信号$y(n)$中估计出干净语音信号$\hat{x}(n)$,使得$\hat{x}(n)$和$x(n)$差异尽可能小,如式(2)所示:

$$\min\ dis(\hat{x}(n),x(n))\ \text{s.t.}\ \hat{x}(n)=f(y(n)) \tag{2}$$

其中,$dis(\cdot)$度量了$\hat{x}(n)$和$x(n)$之间的差异,常见的度量方法包括均方误差(MSE)、平均绝对误差(MAE)等。

由于语音中叠加的噪声存在不同类型、不同信噪比的变化,语音增强模型需要对噪声拥有很好的泛化性能,即需要拥有去除不同类型和不同信噪比噪声的能力。

1.2 基于深度学习的增强模型

传统的单通道语音增强模型为了建模和求解的方便,在公式(2)的基础上加入了其他一些约束或者假设条件,然后根据假设的先验知识来直接估计$\hat{x}(n)$。这些约束和假设不满足时,语音增强性能难以提升。而基于深度学习的语音增强模型不再直接求解模型(2)中的信号估计问题,而是依据设定的目标函数在数据集上获得最优化解的参数,从而隐式地挖掘出带噪语音信号和干净语音信号之间的非线性映射关系$f(\cdot)$,实现由带噪语音信号$y(n)$到干净语音信号$\hat{x}(n)$的映射。

基于深度学习的语音增强模型如图1所示,其中神经网络可以采用DNN、RNN、CNN、LSTM、U-net、Transformer等不同的网络结构。形式化描述可以用如下函数表示:

$$\hat{x}=F(y,\theta) \tag{3}$$

其中,y和$\hat{x}$分别表示网络的输入和输出,它们既可以是时域波形,也可以是时频域变换特征。$\hat{x}$还可以是时频域掩码估计值。此时,利用$\hat{x}$对y进行掩模操作得到干净语音的估计。$F(\cdot,\theta)$表示参数为θ的网络模型。基于深度学习的语音增强模型将语音增强的问题转换为求参数θ最优解的问题,如公式(4)所示:

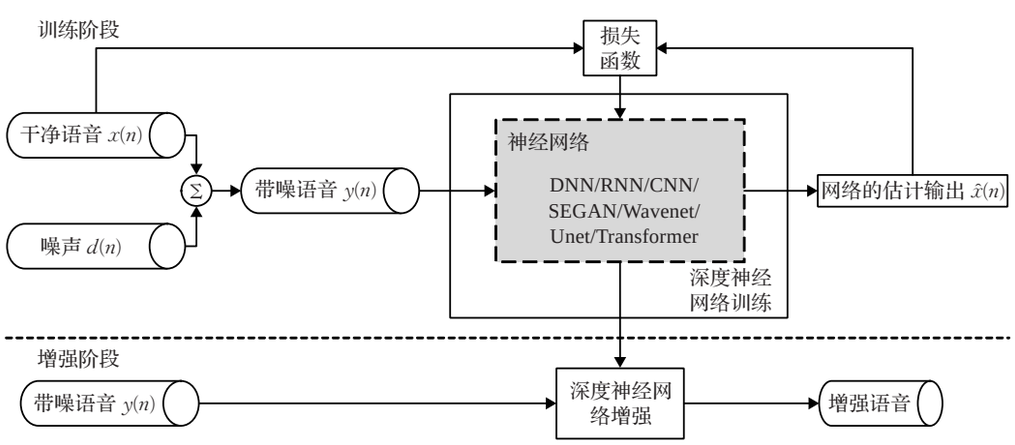

图1 基于深度学习的语音增强模型框图

Fig.1 Block diagram of deep learning-based speech enhancement model

$$\theta^* = \arg\min J(F(y,\theta),x) \tag{4}$$

其中，J 表示目标函数，可以采用MSE等度量形式。

2　基于Transformer单通道语音增强模型

Transformer神经网络能够以并行方式处理输入数据，有效地解决长时依赖问题，显著减少训练时间和推理时间，已在许多自然语言处理任务上展现了突出的性能[31]。然而，语音增强对上下文特征信息的使用不同于机器翻译等自然语言处理任务，因此传统的Transformer神经网络在语音增强方面表现并不佳。为此，需要使用经过改进的Transformer神经网络[26-29]，才能在语音增强中有效发挥Transformer模型的优势。

2.1　传统Transformer模型

传统的Transformer模型如图2所示，由位置编码模块、多头注意力机制模块和前馈网络模块组成。

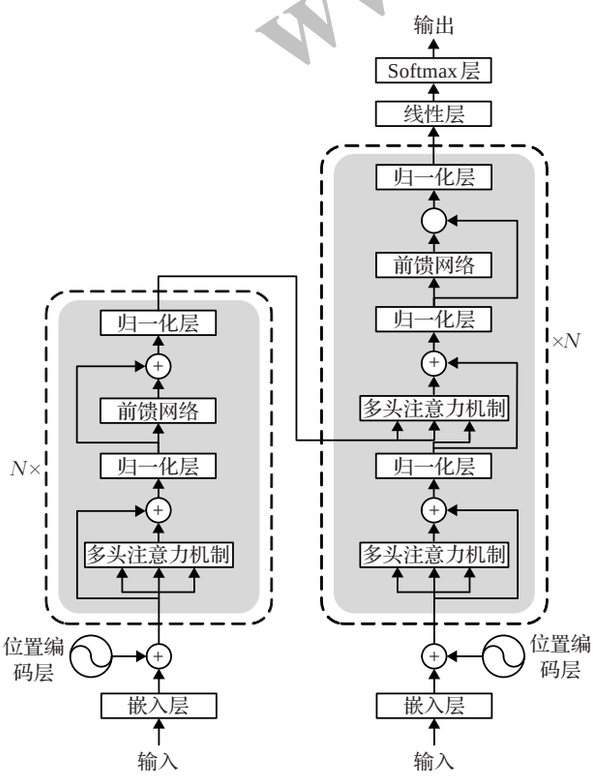

图2　传统的Transformer模型示意图

Fig.2　Schematic diagram of conventional Transformer

多头注意力机制模块的核心是自注意力机制。在自注意力机制的实现中，利用一组键 K 和值 V 记录已学习的信息，通过查询 Q 来得到注意力输出，如图3所示。首先将 Q 和 K 进行相似度计算获得权重，缩放层除以参数 d_k（k 表示维度）起到缩放调节作用，控制内积不至于太大，然后使用softmax函数对相似度权重进行归一化，最后将归一化的权重和相应的 V 进行加权求和得到注意力输出。

计算自注意力机制输出向量的公式如下：

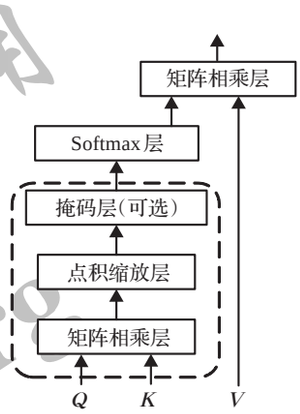

图3　自注意力机制示意图

Fig.3　Schematic diagram of self-attention mechanism

$$attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = softmax(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}})\boldsymbol{V} \tag{5}$$

自注意力机制可以“动态”地生成不同连接的权重，从而得以处理变长的信息序列，在一定程度上解决长时依赖问题。

多头注意力机制模块如图4所示，本质是 h 个自注意力机制的集成，其模块结构并不是很复杂。其中，所有自注意力机制都关注相同的 $\boldsymbol{Q}$、$\boldsymbol{K}$ 和 $\boldsymbol{V}$，但每个模块只对应最终输出序列中的一个子空间，并且输出序列互相独立，这就使得多头注意力机制模块能够对不同位置表征子空间中的不同信息实现同时关注。而在自注意力机制情况下，归一化会抑制这种信息

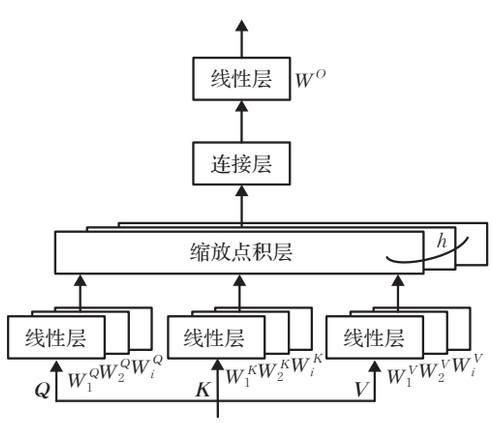

图4　多头注意力机制模块示意图

Fig.4　Schematic diagram of multi-head self-attention mechanism

在实现时，首先初始化 h 组 $\boldsymbol{Q}$、$\boldsymbol{K}$ 和 $\boldsymbol{V}$ 向量，每组 $\boldsymbol{Q}$、$\boldsymbol{K}$ 和 $\boldsymbol{V}$ 的权重参数 $\boldsymbol{W}$ 都不一样，如式(6)所示，通过引入不同的权重可以允许多头注意力机制模块在表征子空间里学习到更多的信息。然后对每组进行自注意力机制的计算，将得到的自注意力机制输出结果连接起来，再乘以一个权重向量 $\boldsymbol{W}^O$ 就可以得到最终多头注意力机制模块的输出向量。

多头注意力机制模块的计算公式如下所示：

$$head_i = attention(\boldsymbol{Q}\boldsymbol{W}_i^Q, \boldsymbol{K}\boldsymbol{W}_i^K, \boldsymbol{V}\boldsymbol{W}_i^V) \tag{6}$$

计算机工程与应用 www.ceaj.org

$$multihead(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=concat(head_1,head_2,\cdots,head_h)\boldsymbol{W}^O \quad (7)$$

Transformer模型完全避免了循环结构[31],采用多头注意力机制实现了输入输出的全局依赖估计。由于每个注意头可以学会执行不同的任务,多头注意力机制可以产生更具解释性的模型。

前馈网络模块如图5所示,由两个线性变换和一个ReLU激活组成。虽然不同位置的线性变换是相同的,但它们在层与层之间使用不同的参数。受RNN对序列信息跟踪有效性的启发,用于语音增强的前馈网络模块通常会把第一个线性变换层替换为GRU层,来学习位置信息[32]。

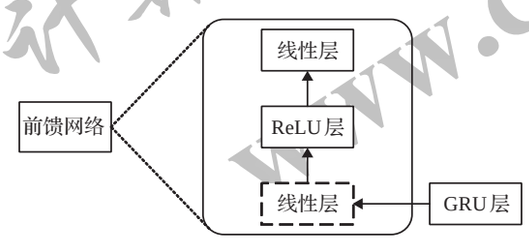

图5 前馈网络模块示意图

Fig.5 Schematic diagram of feedforward network module

2.2 改进Transformer模型

已有研究表明,位置编码模块不适合用于声学建模[33]。为了适应语音处理的应用需求,研究者们已提出多种改进的Transformer模型。

文献[26]在Transformer模型中引入高斯加权矩阵,提出了带有高斯加权的自注意力机制,可以实现以高注意力关注较近的上下文帧,以低注意力关注较远的上下文帧。文献[27]利用局部LSTM对语音信号的位置结构进行建模,替换了原模型中的位置编码模块。文献[29]通过引入两个级联的Transformer模块实现双路径Transformer,用于能同时学习语音信号的局部和全局上下文信息。

为适应流式语音处理的需要,需要对自注意力机制或Transformer模型进行修改,避免其全序列上下文信息进行建模。Transformer-Transducer[34]利用两个VGGnet实现位置信息编码,然后将特征送入Transformer进行编码,在Transformer中利用截断的自注意力限制上下文窗口,降低了处理延迟。文献[35]提出了Conv-Transformer Transducer模型,将自注意力限制为只获取上文信息,实现了流式的处理。文献[36]提出了Chunk自注意力编码器,在利用Transformer编码时只需要使用一个Chunk的上下文信息,不再依赖于整段音频输入。Conformer也是一种改进的Transformer模型[37],其包含一维深度卷积,以实现更有效的上下文特征信息建模。文献[38]提出了DF-Conformer模型,使用线性复杂度的注意力和堆叠的扩张卷积来扩展Conformer,通过减少相邻时间帧的建模范围和观察实时因子(real-time feedback,RTF)可知,该模型可以完成实时任务。

2.3 基于Transformer的增强模型

人类在处理复杂听觉场景时,既能注意到关注的语音内容,又能注意到场景中的背景变化。实际上听觉存在多个注意的焦点。同时,由于关注的语音发音通常由同一人发出,和噪声内容相比,语音在较长时间尺度上特征分布的相似性较强。Transformer所具有的多头注意力机制、长时依赖关系估计能力强的优势可以很好地与人类听觉感知的这些特点相吻合。表1给出了目前已有的多种集成Transformer的语音增强工作,这些模型不同程度地改善了原有系统的增强性能。

表1 集成Transformer的语音增强模型分析

Table 1 Analysis of speech enhancement model integrated with Transformer

模型	优点	缺点
T-GSA[26]	根据目标帧和上下文帧确定注意力权重,合理分配注意力权重	模型参数量大,不能用于实时语音增强
SETransformer[27]	在多头注意力机制模块前加入局部LSTM,学习局部位置结构	模型参数量较大,不能用于实时语音增强
TST-NN[28]	模型参数量较小,模型复杂度较低,提取局部和全局的上下文特征信息	局限于时域特征,不能用于实时语音增强
DPT-FSNET[29]	模型参数量较小,提取子频带和全频带特征信息	不能用于实时语音增强

根据Transformer模块在网络中的不同位置,可将已有工作分为嵌入式结构和组合式结构两类。采用嵌入式结构的模型在网络的编码层或者解码层中加入Transformer,主要用于改善编码层或者解码层的学习效果(如图6所示)。采用组合式结构的模型则在编码器和解码器之间加入Transformer,主要用于计算掩码(Mask)值,以改善解码器的输入(如图7所示)。

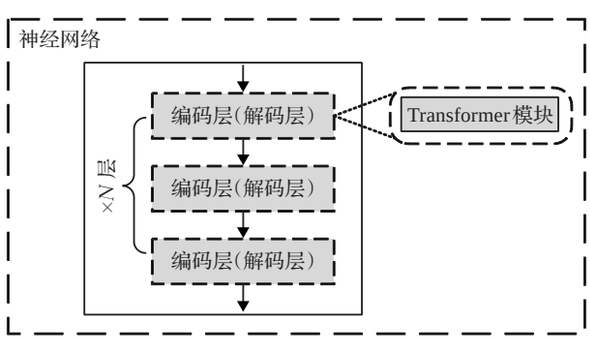

图6 嵌入式结构的Transformer语音增强模型

Fig.6 Embedded structure in Transformer speech enhancement framework

2.3.1 嵌入式结构的Transformer增强模型

带有高斯加权自注意力机制的Transformer(Trans-

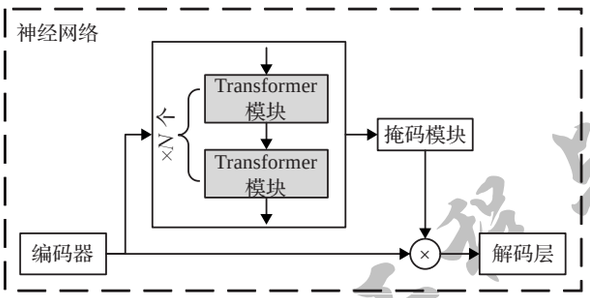

图7　组合式结构的Transformer语音增强模型

Fig.7　Combined structure in Transformer-based speech enhancement framework

former with Gaussian-weighted self-attention，T-GSA）模型和语音增强Transformer（speech enhancement transformer，SETransformer）模型是两个采用嵌入式结构的Transformer增强模型。

T-GSA模型。和传统的Transformer不同，T-GSA模型通过部署一个高斯加权矩阵来修改分数矩阵，使注意力权重可以随着目标帧和上下文帧之间距离的增大而减弱，符合语音信号之间的相关性关系。该模型的自注意力机制输出向量的计算公式如下：

$$attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=softmax\left(\left|G\cdot\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\right|\boldsymbol{V}\right) \tag{8}$$

其中，G表示高斯加权矩阵，$\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)$表示分数矩阵。所提出的高斯加权矩阵$G$计算如下：

$$G=\begin{bmatrix} g_{1,1} & g_{1,2} & \cdots & g_{1,T} \\ g_{2,1} & g_{2,2} & \cdots & g_{2,T} \\ \vdots & \vdots & & \vdots \\ g_{T,1} & g_{T,2} & \cdots & g_{T,T} \end{bmatrix} \tag{9}$$

其中，$g_{i,j}$表示$\mathrm{e}^{\frac{-|i-j|^2}{\sigma^2}}$，$i$表示目标帧的索引，$j$表示上下文帧的索引，$\sigma$表示一个可训练的参数，用于确定权重方差的大小。在语音增强中，组$\boldsymbol{Q}$、$\boldsymbol{K}$和$\boldsymbol{V}$按帧提取，因此$g_{i,j}$与目标帧和上下文帧之间的距离成反比，以便对较远的上下文帧提供较大的注意力衰减，对较近的上下文帧提供较小的注意力衰减。由于σ是一个可训练的参数，因此上下文帧的位置可以通过学习干净语音信号和带噪语音信号组成的声学训练数据来获得。

在语音增强过程中，为了避免实施输入和输出序列的对齐操作，T-GSA单独使用编码网络或解码网络。图8是在编码网络上进行语音增强的Transformer模型图。每个编码网络层由多头注意力机制模块、全连接层模块和归一化层模块组成。网络的输出是一个时频掩码。将掩码和带噪语音信号幅度谱相乘，得到干净语音信号幅度谱的估计。然后结合原始带噪语音信号的相位重构干净语音信号。

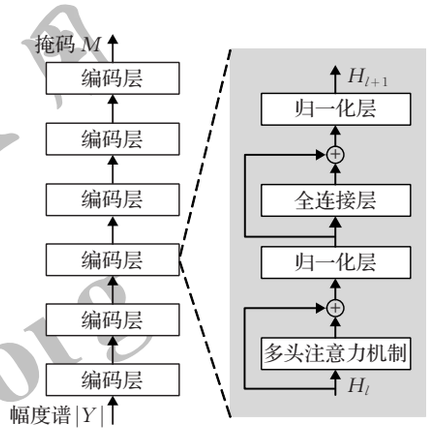

图8　基于Transformer的编码网络模型图

Fig.8　Transformer-based coding network model diagram

上述模型只对幅度谱进行处理，缺乏对相位信息的处理，利用复数网络可以同时保留幅度信息和相位信息。复数网络上的Transformer模型需要两个输入和两个输出，分别是带噪语音信号频谱的实部和虚部、时频掩码的实部和虚部。编码网络中只有一个多头注意力机制模块，不能混合提取实部和虚部的隐藏特征，所以需要使用具有两个多头注意力机制模块的解码网络。图9是在解码网络上进行语音增强的Transformer模型图。

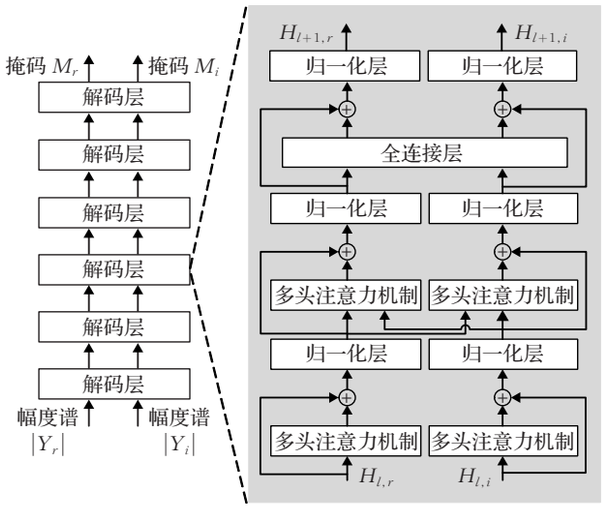

图9　基于Transformer的解码网络模型图

Fig.9　Transformer-based decoding network model diagram

和编码网络中的Transformer模块不同，图9右侧的第一层多头注意力机制模块各自关注上层实部和虚部的输出。第二层多头注意力机制模块关注实部和虚部混合路径的输入，利用它们之间的交叉关系来获取更多的隐藏特征。

SETransformer模型。该模型由局部LSTM、多头注意力机制模块和一维卷积网络模块组成，如图10所示。

SETransformer与T-GSA有两个不同之处。第一个不同是在多头注意力机制模块之前加入了局部LSTM来描述语音信号的位置结构。局部LSTM能够充分捕捉每个窗口内的顺序特征，再通过逐一地滑动操作就能

www.ceaj.org

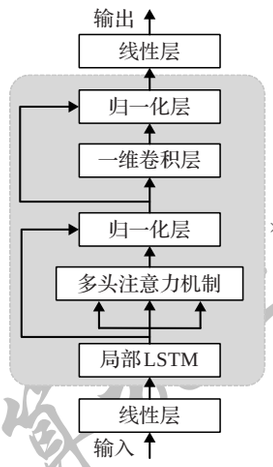

图10　SETransformer整体框架图

Fig.10　Framework diagram of SETransformer

包含全局的顺序特征。另外,局部LSTM只关注本地的短期依赖,不考虑任何长期依赖,因此可以以并行方式来独立处理短序列,降低计算复杂度。第二个不同是把传统Transformer中的前馈网络模块替换成一维卷积网络模块。采用具有ReLU激活的两层一维卷积网络,能够使顺序特征之间的关系更加密切,对提升语音增强效果十分有利。

2.3.2　组合式结构的Transformer增强模型

基于双阶段Transformer的神经网络(two-stage transformer based neural network,TST-NN)模型和基于双路径Transformer的全频带/子频带融合网络(dual-path transformer based full-band and sub-band fusion network,DPT-FSNET)都是组合式结构的Transformer增强模型。

TST-NN模型。该增强模型采用了时域端到端的增强结构,由分割模块、编码器模块、双阶段Transformer模块(two-stage Transformer module,TSTM)、掩码模块、解码器模块和重叠添加模块组成。其中采用的双路径Transformer模型位于编码器和解码器之间,用于估计带噪语音的掩码。图11给出了TST-NN模型图,图中C、N、F分别表示通道、帧的数量帧的大小。

TSTM模块由四个堆叠的双阶段Transformer块组成。双阶段Transformer块由一个局部Transformer和一个全局Transformer组成,可以同时提取局部和全局的上下文特征信息。局部Transformer模块对输入的局部特征信息进行平行化处理,全局Transformer模块用来融合局部Transformer模块的输出信息,以学习全局特征信息,它们都包含了多头注意力机制模块和前馈网络模块。

掩码模块利用TSTM模块的输出来计算用于增强的掩码。该模块首先将TSTM模块的输出通过PReLU运算和卷积对通道维度进行加倍,然后,通过双路二维卷积和sigmoid/tanh非线性运算,将两者的输出相乘,再一次经过二维卷积和PReLU运算后得到掩码。

TST-NN模型直接对时域波形进行处理,避免了频域变换可能带来的失真。

DPT-FSNET模型　该模型在频域上进行语音增强,考虑到语音的频带分布特性,利用双路径Transformer来分别处理全频带和子频带融合网络模型。图12给出了DPT-FSNET模型图,图中C、T、F分别表示通道、帧的数量、频域带数。

DPT-FSNET的编码器模块和解码器模块结构与TST-NN模型的相同,不同之处仅在于输入输出的特征不同。DPT-FSNET模型的编码器模块输入的是高维时频特征(通道数×帧数×频域带数),解码器模块输出频谱用于恢复增强波形。与TST-NN模型相比,这种处理具有更强的可解释性和更多的特征信息。

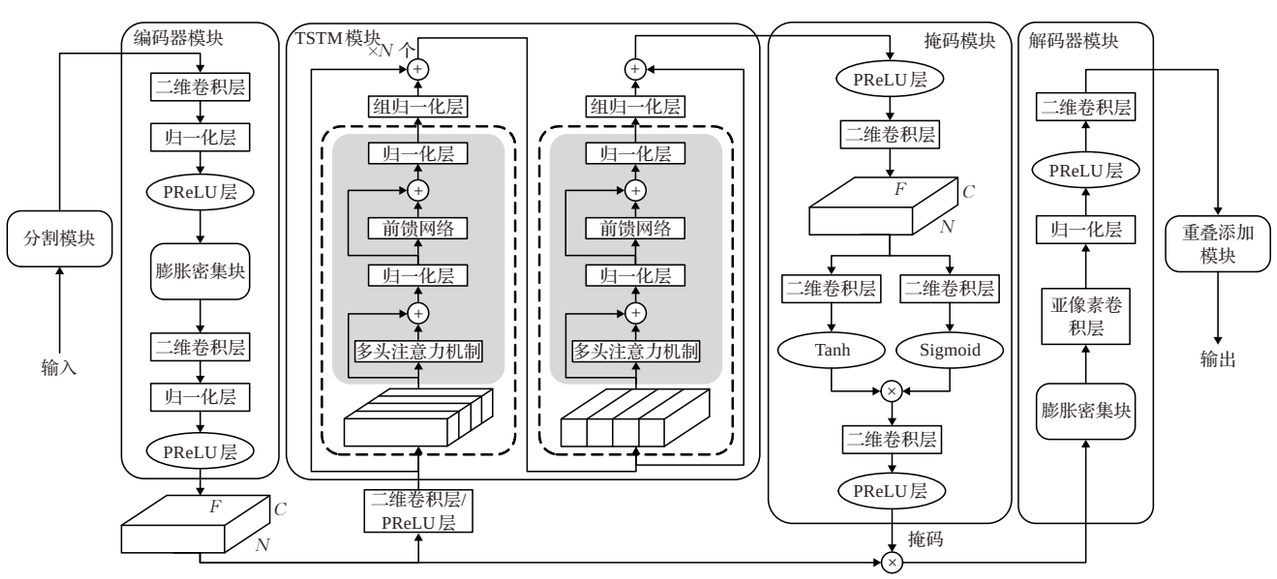

图11　TST-NN模型图

Fig.11　TST-NN model diagram

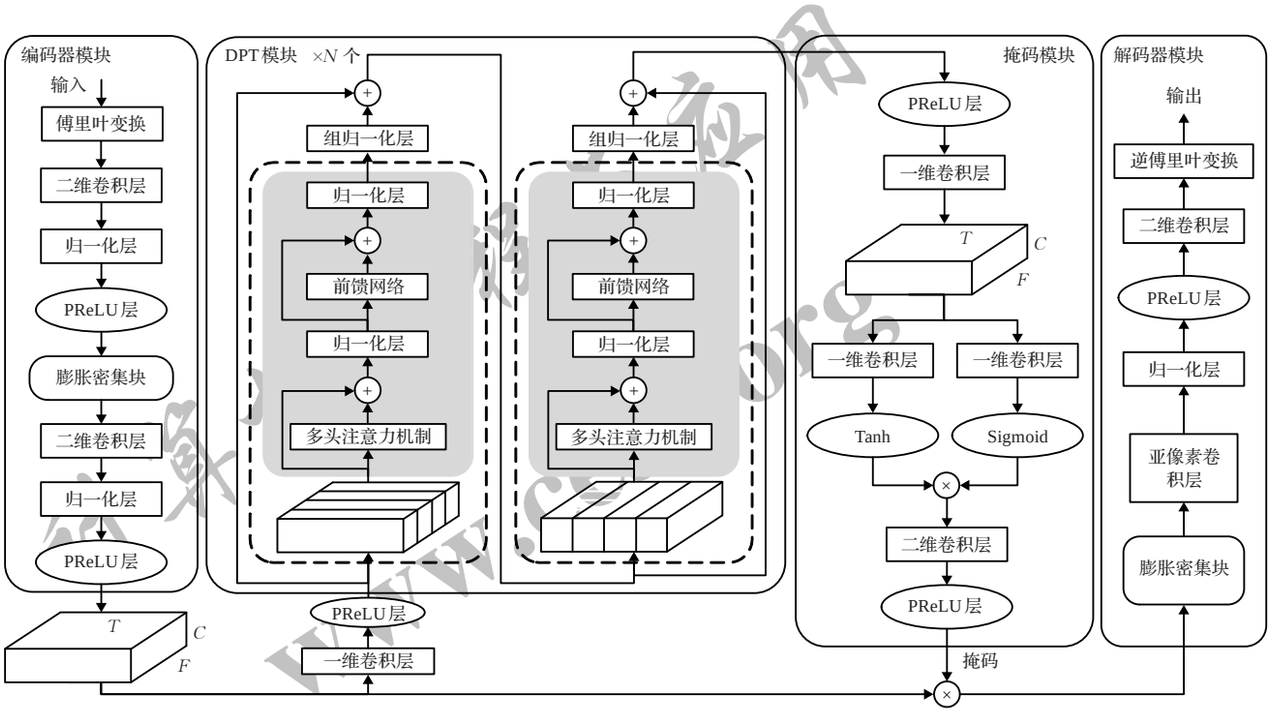

图12 DPT-FSNET模型图

Fig.12 DPT-FSNET model diagram

虽然DPT-FSNET模型的双路径Transformer(dual-path transformer,DPT)模块和TST-NN模型的TSTM模块中的网络模型都是对局部和全局的上下文特征信息进行建模,但是它们的物理意义不同。DPT中的局部Transformer模块是对输入的局部特征信息进行建模,即对语音信号每个子带的所有时间步长的特征进行建模;全局Transformer模块用于汇总局部Transformer模块输出的每个子带特征,即对语音信号所有子带的特征进行建模以学习语音信号的全局特征信息。和TSTM模块相比,DPT具有更好的解释性。

2.3.3 一种融合U-net的组合式语音增强模型

和CNN不同,U-net采用了具有跳跃连接的U型网络结构,可以实现多尺度特征融合处理[39-40]。文献[19]提出了一种端到端的Wave-U-net语音增强模型,不再需要预处理或后处理,为语音增强任务提供了新的解决方案。文献[21]提出了一个融合LSTM的U-net网络,可以实现基于时域的端到端语音增强。该网络目前能够在用户级别CPU上实现实时语音增强。

基于U-net的特征分析能力,在U-net的框架中引入Transformer,设计了一种新的组合式结构语音增强模型(TU-NET),该模型实现了基于时域的端到端单通道语音增强。如图13给出了TU-NET模型图,图中C、N、F分别表示通道、帧的数量、帧的大小。

TU-NET模型与TST-NN模型、DPT-FSNET模型仅在编码器模块和解码器模块上有不同。TU-NET编解码器模块采用U-net网络的编解码层结构,能够实现不同尺度下的特征融合。

TU-NET的编码器模块如图14所示,该模块包括一个上采样层和多个卷积编码层。该模块直接输入语音时域波形。编码模块首先对语音信号进行上采样,然后通过多层编码层分别进行卷积编码。每个编码层都由一维卷积、ReLU函数激活层、一维卷积层和GLU函数激活层级联而成。

TU-NET的解码器模块如图15所示,该模块包括多个解码层和一个下采样层。每个解码层由一维卷积层、GLU函数激活层、一维转置卷积层级联而成。同时,每个解码层的输入都由上一个解码层的输出和同级编码层的输出拼接而成。在最后一层,通过下采样将语音信号的采样频率还原为原始输入频率。

3 Transformer单通道语音增强模型的性能分析

为分析不同Transformer模块对提升单通道语音增强模型的效果,本章选择几种典型Transformer语音增强模型,在不同测试集上进行对比分析。

3.1 测试数据集

对比实验共采用了两个语音增强中常用的测试数据集。一个数据集是VoiceBank-DEMAND数据集[41],其包含干净语音信号和对应预混合的带噪语音信号。干净语音信号选自Voice Bank corpus数据集[42],噪声信号选自DEMAND数据集[43]和2种人工合成的噪声数据集。带噪语音信号使用ITU-T P.56方法将干净语音信号和噪声信号按不同信噪比加性合成。

另一个数据集包含有由TIMIT干净语音数据集和Musan噪声信号数据集[44]生成的带噪语音数据。TIMIT

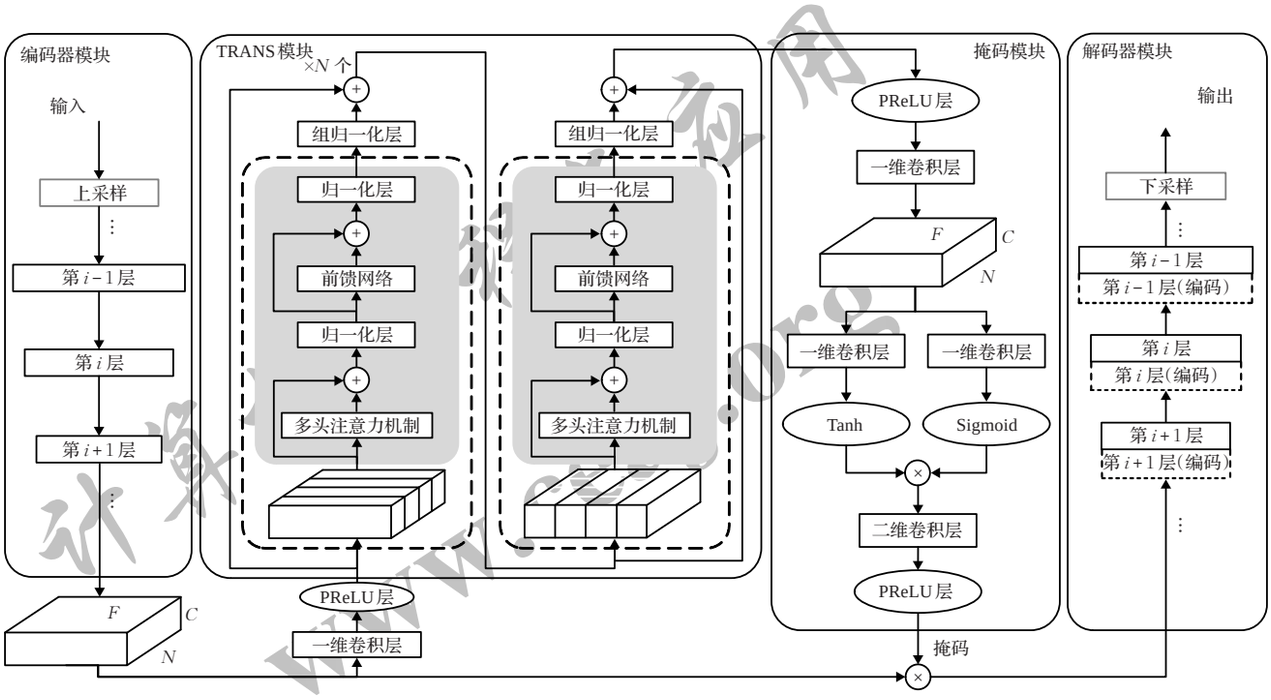

图13　TU-NET模型图

Fig.13　TU-NET model diagram

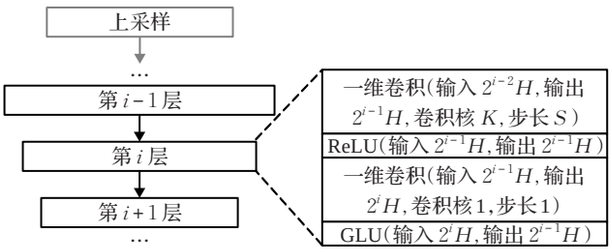

图14　TU-NET编码器模块示意图

Fig.14　Schematic diagram of TU-NET encoder module

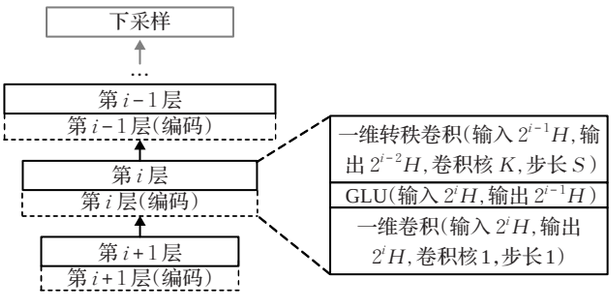

图15　TU-NET解码器模块示意图

Fig.15　Schematic diagram of TU-NET decoder module

数据集一共有6 300条语音信号,包括了630个说话人,每人10条语句的发音。带噪语音信号是将Musan数据集中不同类型的噪声信号按照不同信噪比添加到TIMIT干净语音信号上形成的。

3.2　评价指标

语音增强的性能评价指标主要有两大类,一类是客观质量指标,一类是主观测试指标。客观指标主要包括PESQ(perceptual evaluation of speech quality)[45]和STOI(short-time objective intelligibility)[46],主观指标[47]主要包括MOS(mean opinion score)评估方法中的CSIG(MOS predictor of speech distortion)、CBAK(MOS predictor of intrusiveness of background noise)和COVL(MOS predictor of overall processed speech quality)。

PESQ方法侧重于评估处理语音的总体质量。其分值范围为−0.5~4.5,分值越高,语音的总体质量越好。STOI方法是短时客观可懂度得分方法,侧重于评估处理语音的可懂度。其得分范围为0~1,得分越高,语音的可懂度越高。

MOS分由一组测试者试听原始语音和测试语音,并按照评分标准进行主观打分得到。由于MOS评估成本较高,多用CSIG、CBAK和COVL等客观计算方法来拟合。CSIG是用于计算语音失真度的MOS值,CBAK是用于计算背景噪声干扰的MOS值,COVL是用于计算整体语音质量的MOS值。它们的评分范围都为[1,2,3,4,5],共5个等级,1表示语音质量很差,5表示语音质量非常好,且评分越高,语音质量越好。本文对测试模型分别计算PESQ、STOI、CSIG、CBAK和COVL的评估值(用于计算评估值的代码:HTTPS://GITHUB.COM/IMLHF/SPEECHENHANCEMENTMEASURES),来综合评估增强语音的客观质量和主观质量。

3.3　结果及分析

3.3.1　Voice Bank corpus数据集上增强效果对比

为综合对比基于Transformer的语音增强模型性能,引入SEGAN模型[48]、Wave U-net模型[19]、DCUNet-16模型[20]、PHASE模型[49]、DEMUCS模型[21]作为对照。表2

给出了不同增强模型在Voice Bank corpus数据集上五种评价指标的结果,前八种增强模型的实验结果源自原始论文。

表2 在Voice Bank corpus上的语音评价得分

Table 2 Speech evaluation scoreson Voice Bank corpus

增强模型	T/F	PESQ	STOI	CSIG	CBAK	COVL
SEGAN[48]	T	2.16	0.93	3.48	2.94	2.80
Wave U-net[19]	T	2.40	—	3.52	3.24	2.96
DCUNet-16[20]	F	2.93	—	4.10	3.77	3.53
DEMUCS[21]	T	2.93	0.95	4.22	3.25	3.52
PHASE[49]	F	2.99	—	4.21	3.55	3.62
T-GSA[26]	F	3.06	—	4.18	3.59	3.62
TST-NN[28]	T	2.96	0.95	4.33	3.53	3.67
DPT-FSNET[29]	F	3.30	0.95	4.51	3.69	3.94
TU-NET	T	3.19	0.95	4.45	3.60	3.83

从表2中可以分析得出:

第一,与表1的前五种增强模型相比,后四种使用Transformer的增强模型可以显著提升PESQ、CSIG、CBAK和COVL指标得分,这说明使用Transformer模型可以对语音内在关联信息的学习更为充分,因此增强后的语音音质有了较大改善。

第二,综合比较后四种基于Transformer增强模型的各项指标得分,T-GSA和DPT-FSNET的效果要优于TST-NN和TU-NET。T-GSA和DPT-FSNET都是基于频域处理的增强模型,这说明通过Transformer提取时频域特征所包含的注意力信息比原始的时域特征更加有效。由于自注意力机制采用了并行计算,无需将输入语音的特征拉平,避免了因使用全连接网络而产生对语音信号时频结构的破坏。因此,对时频域特征采用自注意力机制,可以更好地区分带噪语音信号中的干净语音和噪声。

3.3.2 TIMIT数据集上增强效果对比

在TIMIT数据集上,对SETransformer和TU-NET的增强效果进行对比,给出了不同信噪比条件下PESQ、STOI的指标得分,并绘制了不同信号的波形图和语谱图。SETransformer所有实验结果源自原始文献[27]。

由表3可知TU-NET的PESQ、STOI指标得分都优于SETransformer,且在不同信噪比条件下语音的增强效果都有了一定的改善。这说明,同时利用U-net不同尺度特征融合的优势和Transformer多头注意力的优势,能够有效提升不同信噪比条件下语音的增强性能。

表3 在TIMIT上的语音评价得分

Table 3 Speech evaluation scores on TIMIT

信噪比/dB	Noisy		SETransformer		TU-NET	
	PESQ	STOI	PESQ	STOI	PESQ	STOI
−5	1.25	0.58	1.96	0.71	1.99	0.74
0	1.57	0.68	2.36	0.79	2.41	0.83
5	1.94	0.79	2.75	0.86	2.89	0.88
10	2.30	0.88	3.10	0.90	3.13	0.91
15	2.66	0.93	3.42	0.94	3.51	0.96

为了进一步对比SETransformer和TU-NET的增强效果,图16展示了带噪语音信号、干净语音信号、SETransformer增强语音信号、TU-NET增强语音信号的波形图和语谱图。图中的带噪信号被信噪比为0 dB的手机来电噪声所干扰。红色方框表示语音波形的变化,黑色圆框表示语音信号的谐波结构。对比红色方框中的语音波形可以发现,经过SETransformer增强后的语音波形图存在严重的失真问题且后半段语音信号的波形包络不太完整,而经过TU-NET增强后的语音波形包络依然可以完整的保存。对比黑色椭圆框中的频谱分量可以发现,SETransformer增强模型能够抑制更多的噪声,但会导致语音谐波结构的不清晰,语音信号已明显失真,在听觉感知上该增强后的语音信号音量较小且存在明显的机械声音,而TU-NET增强模型能够保存相对完整的谐波结构;通过观察语谱图可以发现图16(h)中的背景色调偏暖,这是因为舒适噪声的存在,在听觉感知上并不影响人耳的感受,该增强后的语音信号音量正常且较为清晰。

4 发展趋势

由于Transformer模型具有可并行、长时预测的性能,在语音增强的研究中逐渐受到越来越多的关注。在现有模型的基础上,还将在以下几个方面有所发展。

(1)结合优化的网络结构和损失函数。深度网络结构和使用的损失函数对网络性能有着重要的影响。TU-net采用了U-net结构,性能在SETransfomer基础上有所提高。SETransformer增强模型采用均方误差(MSE)作为损失函数。而TST-NN同时结合了时域和时频域的损失函数,时域中采用MSE损失函数,时频域中采用平均绝对误差(MAE)损失函数。这些结果表明,采用表征学习能力更强的网络框架,以及更准确反映语音音质和噪音抑制效果的损失函数,将进一步提升语音增强质量。

(2)结合人类听觉感知。在低信噪比、混响等条件下,即使采用了深度神经网络仍难以较好地提升语音增强的质量[47]。若从人类听觉感知机理出发,研究基于听觉感知的语音增强模型或者基于听觉感知的损失函数,将对语音增强模型有很好的推动作用。

(3)引入噪声自适应机制。利用自适应字典学习算法为深度模型构造噪声字典,或者利用强化学习算法自适应学习不同噪声环境下的奖励,可以对网络模型的增强结果作进一步优化,有利于提升增强模型在不同噪声条件下的适应能力。

(4)设计因果Transformer模型。现有基于Transformer的语音增强模型,需要同时使用上下文特征信息,这种

计算机工程与应用 www.ceaj.com

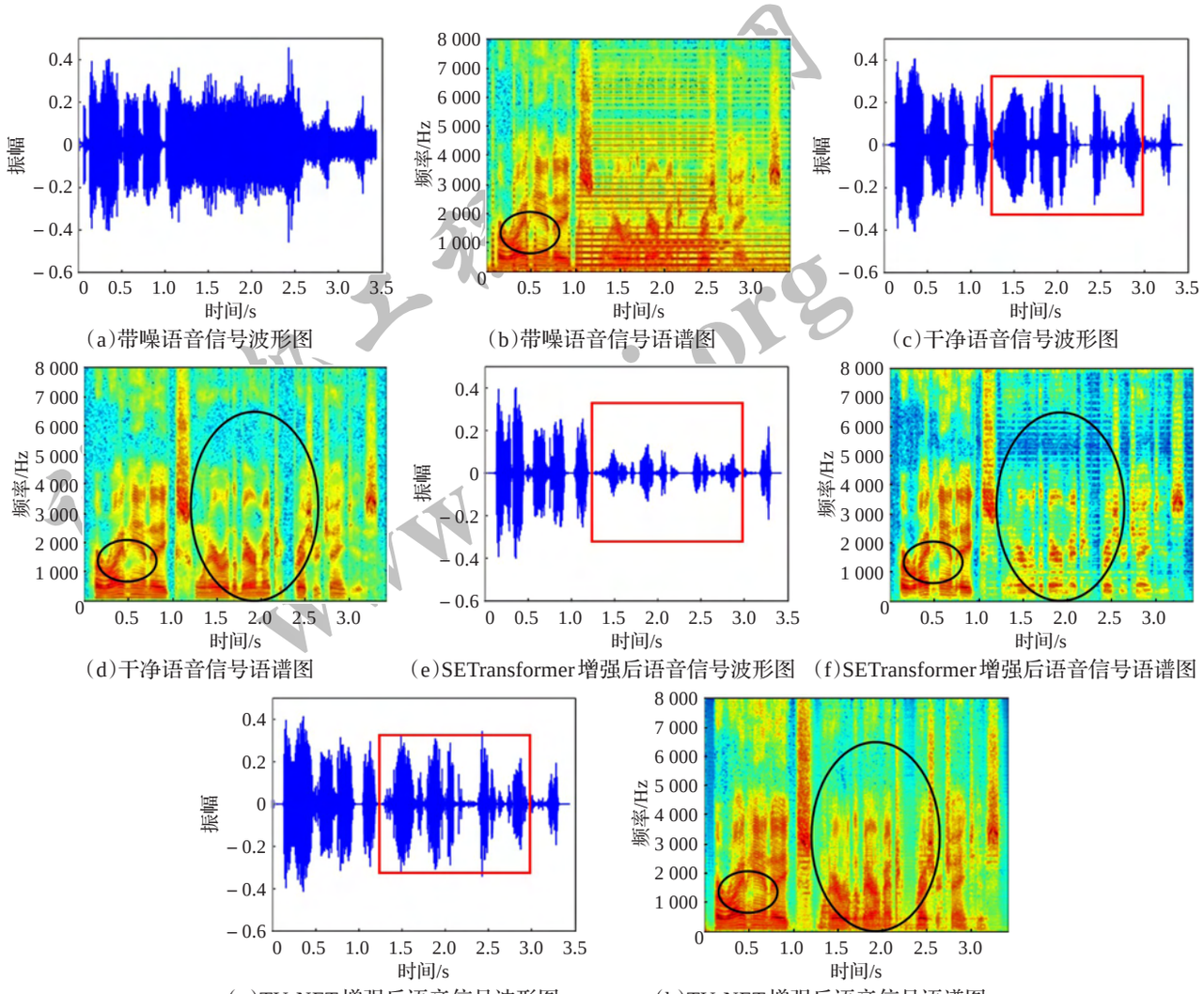

图16　语音质量的对比图

Fig.16　Comparison chart of voice quality

对下文信息的依赖使得Transformer模型并不具有因果性,实现的语音增强模型也无法适应即时通信场合。由于Transformer具有参数量少的优势,TST-NN[28]和DPT-FSNET[29]的参数规模已低于4 MB,通过因果Transformer模型的优化设计,结合现有硬件发展成果,提高实时比,有望实现嵌入式的实时语音增强系统。

(5)设计多通道增强模型。本文所研究的内容都是基于单通道的语音增强模型。现在很多终端设备都具有多个麦克风,如果能够合理利用不同的通道信息,实现通道注意力机制或者空间注意力机制,从理论上来说有助于恢复干净语音信号。

5　结束语

本文系统介绍了基于Transformer的单通道语音增强模型,通过对网络结构的研究与分类,详细地阐述了T-GSA、SETransformer、DPT-FSNET、TST-NN和TU-NET等网络模型结构,对比分析了这些模型的各自优缺点。文中介绍的五种语音增强模型的原始网络都是基于编码-解码网络模型的,无论采用了何种集成方式,Transformer模块均可以发挥模型自身的优势,很好地提高语音的质量和可懂度。下一步,可根据发展趋势进一步探索Transformer模块在语音增强的深度应用,以更少的网络参数,更快的处理速度为最终的目标,从而更好地实现高质量的单通道语音增强。

参考文献:

[1] WANG W,XING C,WANG D,et al.A robust audio-visual speech enhancement model[C]//2020 IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP),Barcelona,Spain,2020:7529-7533.

[2] GOGATE M,DASHTIPOUR K,BELL P,et al.Deep neural network driven binaural audio visual speech separation[C]//2020 International Joint Conference on Neural Networks(IJCNN),Glasgow,UK,2020:1-7.

[3] LI L,WANG D,ZHENG T F.Neural discriminant analysis

for deep speaker embedding[J].arXiv:2005.11905,2020.
[4] 陶智,赵鹤鸣,龚呈卉.基于听觉掩蔽效应和bark子波变换的语音增强[J].声学学报,2005,30(4):367-372.
TAO Z,ZHAO H M,GONG C H.Speech enhancement based on masking properties of human auditory system and bark wavelet transform[J].Acta Acustica,2005,30(4):367-372.
[5] ERKELENS J S,HENDRIKS R C,HEUSDENS R,et al. Minimum mean-square error estimation of discrete fourier coefficients with generalized gamma priors[J].IEEE Transactions on Audio Speech and Language Processing,2007,15(6):1741-1752.
[6] BORGSTROM B J,ALWAN A.Log-spectral amplitude estimation with generalized gamma distributions for speech enhancement[C]//2011 IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP),Prague,Czech Republic,2011:10.
[7] JU G H,LEE L S.A perceptually constrained gsvd-based approach for enhancing speech corrupted by colored noise[J]. IEEE Transactions on Audio Speech and Language Processing,2007,15(1):119-134.
[8] BOROWICZ A.A signal subspace approach to spatio-temporal prediction for multichannel speech enhancement[J].Eurasip Journal on Audio Speech and Music Processing,2015(1):1-12.
[9] YAN Q,VASEGHI S,ZAVAREHEI E,et al.Kalman tracking of linear predictor and harmonic noise models for noisy speech enhancement[J].Computer Speech and Language,2008,22(1):69-83.
[10] CHEN R F,CHAN C F,SO H C.Model-based speech enhancement with improved spectral envelope estimation via dynamics tracking[J].IEEE Transactions on Audio Speech and Language Processing,2012,20(4):1324-1336.
[11] WANG Y X,WANG D L.Towards scaling up classification-based speech separation[J].IEEE Transactions on Audio,Speech,and Language Processing,2013,21(7):1381-1390.
[12] HEALY E W,YOHO S E,WANG Y X,et al.An algorithm to improve speech recognition in noise for hearing-impaired listeners[J].Journal of the Acoustical Society of America,2013,134(4):3029-3038.
[13] XU Y,DU J,DAI L,et al.A regression approach to speech enhancement based on deep neural networks[J].IEEE/ACM Transactions on Audio,Speech,and Language Processing,2015,23(1):7-19.
[14] WENINGER F,HERSHEY J R,ROUX J L,et al.Discriminatively trained recurrent neural networks for single-channel speech separation[C]//2014 IEEE Global Conference on Signal and Information Processing(GlobalSIP),Atlanta,GA,USA,2014:577-581.
[15] WENINGER F,ERDOGAN H,WATANABE S,et al.Speech enhancement with LSTM recurrent neural networks and its application to noise-robust asr[C]//Latent Variable Analysis and Signal Separation,Cham,2015:91-99.
[16] PARK S R,LEE J.A fully convolutional neural network for speech enhancement[J].arXiv:1609.07132,2016.
[17] RETHAGE D,PONS J,SERRA X.A wavenet for speech denoising[C]//2018 IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP),Calgary,AB,Canada,2018:5069-5073.
[18] 张天骐,柏浩钧,叶绍鹏,等.基于门控残差卷积编解码网络的单通道语音增强方法[J].信号处理,2021,37(10):1986-1995.
ZHANG T Q,BAI H J,YE S P,et al.Single-channel speech enhancement method based on gated residual convolution encoder-and-decoder network[J].Journal of Signal Processing,2021,37(10):1986-1995.
[19] STOLLER D,EWERT S,DIXON S.Wave-u-net:a multi-scale neural network for end-to-end audio source separation[J].arXiv:1806.03185,2018.
[20] CHOI H S,KIM J H,HUH J,et al.Phase-aware speech enhancement with deep complex u-net[J].arXiv:1903.03107,2019.
[21] DEFOSSEZ A,SYNNAEVE G,ADI Y.Real time speech enhancement in the waveform domain[J].arXiv:2006.12847,2020.
[22] 徐峰,李平.DVUGAN:基于STDCT的DDSP集成变分U-Net的语音增强[J].信号处理,2022,38(3):582-589.
XU F,LI P.DVUGAN:DDSP integrated variational u-net speech enhancement based on STDCT[J].Journal of Signal Processing,2022,38(3):582-589.
[23] ZHOU S,DONG L,XU S,et al.A comparison of modeling units in sequence-to-sequence speech recognition with the transformer on mandarin chinese[J].arXiv:1805.06239,2018.
[24] DAI Z,YANG Z,YANG Y,et al.Transformer-xl:attentive language models beyond a fixed-length context[J].arXiv:1901.02860,2019.
[25] CHEN J,LU Y,YU Q,et al.Transunet:transformers make strong encoders for medical image segmentation[J].arXiv:2102.04306,2021.
[26] KIM J,EL-KHAMY M,LEE J.T-gsa:transformer with Gaussian-weighted self-attention for speech enhancement[C]//2020 IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP),Barcelona,Spain,2020:6649-6653.
[27] YU W,ZHOU J,WANG H,et al.Setransformer:speech enhancement transformer[J].Cognitive Computation,2021.
[28] WANG K,HE B,ZHU W P J A E P.Tstnn:two-stage transformer based neural network for speech enhancement in the time domain[J].arXiv:2103.09963,2021.
[29] DANG F,CHEN H,ZHANG P.Dpt-fsnet:dual-path trans-

计算机工程与应用 www.ceaj.org

former based full-band and sub-band fusion network for speech enhancement[J].arXiv:2104.13002,2021.

[30] 李斌.基于深度神经网络的单通道语音增强方法研究[D].杭州:浙江大学,2020.
LI B.Research on single channel speech enhancement based on deep neural network[D].Hangzhou:Zhejiang University,2020.

[31] VASWANI A,SHAZEER N,PARMAR N,et al.Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems,Long Beach,California,USA,2017:6000-6010.

[32] SPERBER M,NIEHUES J,NEUBIG G,et al.Self-attentional acoustic models[J].arXiv:1803.09519,2018.

[33] CHEN J,MAO Q,LIU D.Dual-path transformer network:direct context-aware modeling for end-to-end monaural speech separation[J].arXiv:2007.13975,2020.

[34] YEH C F,MAHADEOKAR J,KALGAONKAR K,et al.Transformer-transducer:end-to-end speech recognition with self-attention[J].arXiv:1910.12977,2019.

[35] HUANG W,HU W,YEUNG Y T,et al.Conv-transformer transducer:low latency,low frame rate,streamable end-to-end speech recognition[J].arXiv:2008.05750,2020.

[36] MIAO H,CHENG G,GAO C,et al.Transformer-based online ctc/attention end-to-end speech recognition architecture[C]//2020 IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP),Barcelona,Spain,2020:6084-6088.

[37] O'MALLEY T,NARAYANAN A,WANG Q,et al.A conformer-based asr frontend for joint acoustic echo cancellation,speech enhancement and speech separation[J].arXiv:2111.09935,2021.

[38] KOIZUMI Y,KARITA S,WISDOM S,et al.Df-conformer:integrated architecture of conv-tasnet and conformer using linear complexity self-attention for speech enhancement[C]//2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics(WASPAA),New Paltz,NY,USA,2021:161-165.

[39] DONG H,PAN J,XIANG L,et al.Multi-scale boosted dehazing network with dense feature fusion[J].arXiv:2004.13388,2020.

[40] 吕佳,马超,程超.改进的U-Net网络用于视网膜血管分割[J/OL].计算机科学与探索:1-12[2021-12-01].https://kns.cnki.net/kcms/detail/11.5602.TP.20210825.1945.007.html.
LYU J,MA C,CHENG C.Improved U-Net network for retinal vascular segmentation[J].Journal of Frontiers of Computer Science and Technology:1-12[2021-12-01].https://kns.cnki.net/kcms/detail/11.5602.TP.20210825.1945.007.html.

[41] VALENTINI-BOTINHAO C,WANG X,TAKAKI S,et al.Investigating RNN-based speech enhancement methods for noise-robust text-to-speech[C]//9th ISCA Speech Synthesis Workshop,2016:146-152.

[42] VEAUX C,YAMAGISHI J,KING S.The voice bank corpus:design,collection and data analysis of a large regional accent speech database[C]//2013 International Conference Oriental COCOSDA Held Jointly with 2013 Conference on Asian Spoken Language Research and Evaluation(O-COCOSDA/CASLRE),Gurgaon,India,2013:1-4.

[43] THIEMANN J,ITO N,VINCENT E.The diverse environments multi-channel acoustic noise database(demand):a database of multichannel environmental noise recordings[J].Journal of the Acoustical Society of America,2013,133(5):3591.

[44] SNYDER D,CHEN G,POVEY D.Musan:a music,speech,and noise corpus[J].arXiv:1510.08484,2015.

[45] BOTCHEV V J C R.Speech enhancement:theory and practice(2nd ed.)[J].Computing Reviews,2013,54(10):604-605.

[46] TAAL C H,HENDRIKS R C,HEUSDENS R,et al.An evaluation of objective measures for intelligibility prediction of time-frequency weighted noisy speech[J].The Journal of the Acoustical Society of America,2011,130(5):3013-3027.

[47] HU Y,LOIZOU P C.Evaluation of objective quality measures for speech enhancement[J].IEEE Transactions on Audio Speech Language Process,2008,16(1):229-238.

[48] PASCUAL S,BONAFONTE A,SERRà J.Segan:speech enhancement generative adversarial network[J].arXiv:1703.09452,2017.

[49] YIN D,LUO C,XIONG Z,et al.Phasen:a phase-and-harmonics-aware speech enhancement network[C]//Proceedings of the AAAI Conference on Artificial Intelligence,2020:9458-9465.

计算机工程与应用 www.ceaj.org