低信噪比下多级特征深度融合的视听语音增强

张天骐，沈夕文，唐娟，谭霜
（重庆邮电大学通信与信息工程学院，重庆 400065）

摘　要：为解决视听语音增强中特征提取受限、模态间的特征融合度低等问题，提出一种在低信噪比下的多级特征深度融合的视听语音增强方法。该方法采用视、听编码网络-视听融合网络-听觉解码网络的结构，在听觉编码网络中设计一种多路协作单元（MCU）；在每层的视觉和听觉编码网络间设计一种视听注意力融合模块（AVAFM）；在视听融合网络中设计一种融合加权模块（FWB），将每级输出进行特征优化、动态加权得到更具判别性的特征。最终在TMSV、LGRID视听数据集上的多种低信噪比的实验结果表明，LGRID视听数据集下的平均PESQ、STOI分别提升52.30%~74.06%、46.74%~67.15%，且相比纯音频语音增强，在−5 dB、−2 dB、1 dB低信噪比下的平均PESQ和STOI分别提升38.95%和33.92%，表现出所提网络的高降噪性能和添加视觉信息的有效性。

关键词：视听语音增强；低信噪比；多级特征融合；融合加权；视听注意力
中图分类号：TP18；TN912.35
文献标志码：A
DOI：10.11959/j.issn.1000–436x.2025075

Audio-visual speech enhancement with multi-level feature deep fusion under low signal-to-noise ratio

ZHANG Tianqi, SHEN Xiwen, TANG Juan, TAN Shuang
School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China

Abstract: To address the limitations in feature extraction and cross-modal fusion in audio-visual speech enhancement, a multistage deep fusion method was proposed for low signal-to-noise ratio (SNR) conditions. The method consisted of an audio-visual encoding network, a fusion network, and an auditory decoding network. A multi-branch collaborative unit (MCU) was introduced in the auditory encoder, along with an audio-visual attention fusion module (AVAFM) between each visual and auditory layer. A fusion weighting block (FWB) was also designed to optimize and dynamically weight features at each stage. Experiments on TMSV and LGRID datasets showed that the proposed method significantly improved PESQ and STOI scores under various low-SNR conditions. Compared to audio-only enhancement, average gains of 38.95% in PESQ and 33.92% in STOI were achieved at −5 dB, −2 dB, and 1 dB. These results demonstrate the method's strong denoising ability and the effectiveness of visual information.

Keywords: audio-visual speech enhancement, low signal-to-noise ratio, multi-level feature fusion, fusion weighted, audio-visual attention

0　引言

日常生活中的高强度噪声干扰往往会严重降低语音质量和可懂度，语音增强[1]技术旨在从含噪语音中分离出目标语音，以实现降噪效果。语音增强

收稿日期：2025–02–08；修回日期：2025–04–14
通信作者：沈夕文，674051078@qq.com
基金项目：重庆市自然科学基金资助项目（No.cstc2021jcyj-msxmX0836）
Foundation Item：The Natural Science Foundation of Chongqing (No.cstc2021jcyj-msxmX0836)

可以应用于日常各方面，包括语音识别、助听器、耳机、视频会议降噪等。近年来，研究人员提出很多基于神经网络的方法来恢复噪声环境下目标说话人的纯净语音，这些方法可以分为2类，即纯音频语音增强（AOSE, audio-only speech enhancement）和视听语音增强（AVSE, audio-video speech enhancement）。

纯音频语音增强只是对单一音频信号进行特征建模，近年来得到了较大的发展。Park等[2]使用全卷积代替全连接，在降低参数量的同时提升网络性能。Tan等[3]在此基础上设计了卷积递归网络，编解码层使用全卷积模块，中间层使用长短期记忆（LSTM, long short term memory）网络以捕捉长短期特征。Wang等[4]提出基于编解码的双路径Transformer结构，对有声段语音的时间和频率维度特征进行长期建模，但在语音沉默段对噪声消除能力较差。与纯音频语音增强相比，视听语音增强被证实有更强的降噪性能[5]，一些学者推测，由于视觉信息传输到大脑的速度比声音快，视觉对听觉系统具有预测辅助作用，提高了大脑对声音的注意力[6]。与此同时，视听语音增强可以更好地学习视觉和听觉间的相关性，加强语音信号和环境信息之间的联系，以更好地区分特征差异，还原纯净语音。

视听语音增强是对音频和视频信号进行多模态的特征建模，文献[7]采用卷积神经网络分别处理音频和视频（嘴唇）特征，拼接融合后经过多层全连接层和音频解码层，最后输出音频特征。文献[8-9]在文献[7]的基础上将视觉特征进行压缩编码处理，并且将2个网络的输出拼接融合成一个视听联合网络输入LSTM中学习长期记忆特征，最后通过线性层分别输出视觉和听觉特征。文献[10]将文献[9]中的嘴唇特征进行了进一步特殊处理，在保护隐私的同时减少了视频流的数据。文献[11]提出了一种基于U形结构的视听网络，通过综合音频和视频信息，以提升噪声语音信号的可懂度和整体感知质量。文献[12]将视觉和听觉特征图在生成器中多次进行融合、下采样和卷积，以处理长时间上下文信息，并在不同层次提取高级和低级特征。文献[13]设计了一个视听Transformer模块，通过学习模态内和模态间的长期依赖性，以更好地整合音频和视频流的长期跨度信息。文献[14]侧重于提取嘴唇的三维结构，并进行多级卷积处理以消除更多的视觉冗余信息和干扰因素。文献[15]抛弃了单模态的纯音频语音增强方法，首先利用纯净语音生成虚拟的人脸视频，然后截取生成的数字人脸视频的嘴唇特征，将其作为视觉信息以辅助语音信号建模，取得了较好的降噪效果。这些多模态的新方法为语音增强领域带来了全新的研究方向。多模态语音增强相对于单模态语音增强具有更强的适应性和更全面的环境感知，为语音技术在各个领域的发展带来了更多可能性，将推动语音技术朝着更智能化、人性化的方向发展。

然而，上述纯音频语音增强模型在低信噪比下的降噪能力较差，尤其在语音沉默段。视听语音增强模型中对于音频处理只是在单支路上对特征进行提取，存在特征单一、对关键信息捕捉不强等缺点，且上述模型对视觉和听觉信息的简易融合过程中会丢失较多的细节特征，导致对细节特征的处理不足，不能充分利用2种感官信息的互补性，会造成特征不匹配和信息丢失等问题，影响最终的特征融合效果。

针对上述问题，本文提出一种在低信噪比下的视听语音增强模型，并在此模型中设计了多级特征深度融合的方法。该方法采用视、听编码网络-视听融合网络-听觉解码网络的结构，在视、听编码网络部分，听觉编码网络设计一种多路协作单元（MCU），通过交互学习多支路特征以提取更丰富的视觉特征；听觉编码网络设计一种视觉卷积单元（VCU, visual convolutional unit）以提取深层视觉特征；在视、听编码模块之间设计一种视听注意力融合模块（AVAFM）以更好地学习模态间的相关性；在视听融合网络中设计了一种融合加权模块（FWB），动态选择重要的低精度特征和丢弃冗余特征。最后，在大、小视听数据集上验证本文网络的降噪效果和网络泛化能力。

1 相关工作

1.1 基于神经网络的视听语音增强流程

视听语音增强与纯音频语音增强相比，输入特征不只是单一的音频信号，还包含了视频信号，视频信号可以辅助语音信号更好地学习含噪语音到干净语音特征空间的映射函数，视听语音增强流程如图1所示。

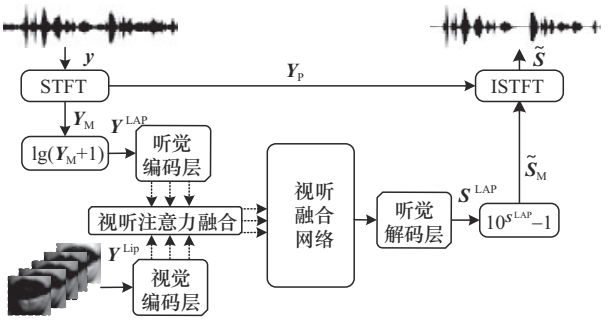


图1　视听语音增强流程

设含噪语音信号为

$$\boldsymbol{y} = \boldsymbol{s} + \boldsymbol{n} \tag{1}$$

其中，$\boldsymbol{y}$、$\boldsymbol{s}$、$\boldsymbol{n}\in\mathbb{R}^{1\times L}$分别为含噪语音、纯净语音、噪声语音的时域表示，$L$为时域信号的长度。首先将式(1)中含噪语音$\boldsymbol{y}$进行短时傅里变换（STFT, short time fourier transform）处理得到幅度谱$\boldsymbol{Y}_{\mathrm{M}}$和相位谱$\boldsymbol{Y}_{\mathrm{P}}$，再求其对数幅度谱$\boldsymbol{Y}^{\mathrm{LAP}}$

$$\boldsymbol{Y} = \mathrm{STFT}(\boldsymbol{y}) = \boldsymbol{Y}_{\mathrm{r}} + \mathrm{j}\boldsymbol{Y}_{\mathrm{i}} \tag{2}$$

$$\boldsymbol{Y}_{\mathrm{M}} = \sqrt{\boldsymbol{Y}_{\mathrm{r}}^{\,2} + \boldsymbol{Y}_{\mathrm{i}}^{\,2}}, \boldsymbol{Y}_{\mathrm{P}} = \arctan 2(\boldsymbol{Y}_{\mathrm{i}}, \boldsymbol{Y}_{\mathrm{r}}) \tag{3}$$

$$\boldsymbol{Y}^{\mathrm{LAP}} = \lg\left(\boldsymbol{Y}_{\mathrm{M}} + 1\right) \tag{4}$$

其中，$\boldsymbol{Y}_{\mathrm{r}}$和$\boldsymbol{Y}_{\mathrm{i}}$分别表示含噪语音的实部和虚部，lg (·)为取对数函数。令$\boldsymbol{Y}^{\mathrm{LAP}}$为听觉输入特征，$\boldsymbol{Y}^{\mathrm{Lip}}$为视觉输入特征，将视觉特征和听觉特征输入网络模型中，通过模型映射得到增强后的对数幅度谱$\boldsymbol{S}^{\mathrm{LAP}}$，增强后的语音幅度谱$\tilde{\boldsymbol{S}}_{\mathrm{M}}$为

$$\tilde{\boldsymbol{S}}_{\mathrm{M}} = 10^{\boldsymbol{S}^{\mathrm{LAP}}} - 1 \tag{5}$$

最后联合增强后幅度谱$\tilde{\boldsymbol{S}}_{\mathrm{M}}$和噪声相位谱$\boldsymbol{Y}_{\mathrm{P}}$进行短时逆傅里叶变换（ISTFT, inverse short time fourier transform），得到增强的时域波形$\tilde{\boldsymbol{s}}$

$$\tilde{\boldsymbol{s}} = \mathrm{ISTFT}\left(\tilde{\boldsymbol{S}}_{\mathrm{M}} \exp(\mathrm{j}\boldsymbol{Y}_{\mathrm{P}})\right) \tag{6}$$

1.2　音频特征预处理

首先将所有音频处理为单声道，且重采样为16 kHz，然后进行短时傅里叶变换以产生1 × 256 × F × 2的听觉复频谱特征，每秒处理50帧，与视频特征同步。其中，帧长为31.937 5 ms，帧移为20 ms，汉明窗长为31.937 5 ms。对于每个语音帧，首先计算其对数幅度谱，并减去平均值再除以标准偏差以进行归一化处理，其次将±2帧连接到中心帧作为上下帧窗口，最后在每个时间步长上产生1 × 256 × 5的听觉特征。

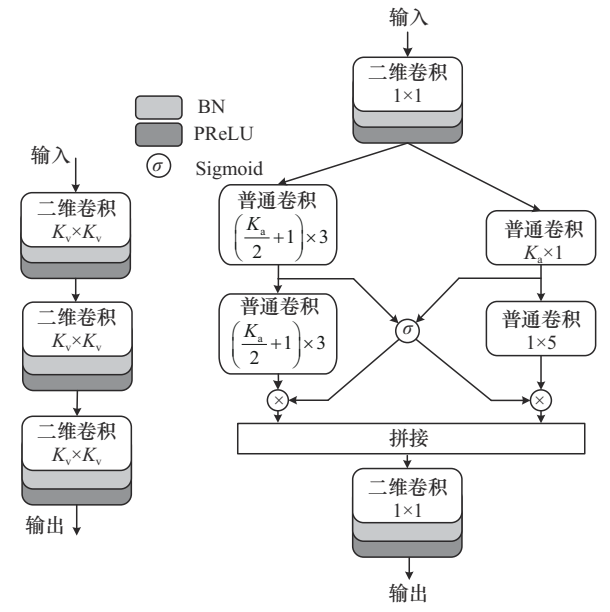
1.3　视频特征预处理

首先以每秒50帧的速率将每个音频所对应的视频转换成图像序列，以保持语音帧和图像帧的同步。接下来使用Dlib库函数检测出嘴巴区域，裁剪32像素 × 32像素，将所有彩色图像变为灰白图像，并使通道数变为1以降低数据量。最后在−1到1的范围内进行归一化处理以重新调整图像的像素强度。此外，将±2帧连接到中心帧，从而在每个时间步上产生尺寸为1 × 32 × 32 × 5的视觉特征。

2　网络结构

2.1　视觉卷积单元

VCU遵循Gabbay等[7]提出的视听网络视觉编码结构，如图2(a)所示，输入特征$\boldsymbol{Y}^{\mathrm{Lip}}\in\mathbb{R}^{B\times C\times H\times W\times F}$，其中，$B$、$C$、$H$、$W$、$F$分别为批次数、通道数、图像高度、图像宽度、帧数。第一个VCU使用三层卷积模块，第二个和第三个VCU均使用二层卷积模块，3个VCU中的卷积核尺寸K_{v}依次为5、5、5，5、5，3、3，且每层的卷积模块步长以2 × 1、1 × 2交替进行来调整维度变化，以适应每层视听编码网络的特征融合，每层卷积模块均包含批次归一化和参数化修正线性单元（PReLU, parametric rectification linear unit）激活函数。经过N层VCU模块后得到输出特征$\boldsymbol{Y}_N^{\mathrm{Lip}}\in\mathbb{R}^{(B\times F)\times C\times\left(\frac{H\times W}{2^{N+2}}\right)}$，其中$N$为层数，且本文设置为3。



(a) 视觉卷积单元　　(b) 多路协作单元

图2　视觉和听觉编码模块

2.2 多路协作单元

目前的视听语音增强方法中对于音频特征的提取是单支路卷积（SC, single-branch convolution），而单支路卷积往往因其固定的卷积核尺寸和单分支信息提取的限制性，只能捕获单一尺度的特征。受双支路卷积[16]和特征交互[17]启发，本文在听觉编码网络中设计一种MCU，通过将多个支路不同尺寸的卷积核堆叠在一起，可以更好地捕捉大感受野信息和不同尺度信息；支路间的信息交互、共享可以减少冗余特征的提取，更好地捕捉特征间的依赖性，从而提取更加丰富的特征。

MCU如图2(b)所示，图中⊗表示按位相乘，输入的听觉特征 $\boldsymbol{Y}^{\mathrm{LAP}} \in \mathbb{R}^{B \times C \times T \times F}$，其中，$B$、$C$、$T$、$F$分别为批次数、通道数、时间帧数、频率帧数。首先通过卷积核尺寸为1 × 1、步长为2 × 1的卷积来调整通道数和降低频率帧的大小以减小参数量；然后左支路输入2个卷积核尺寸为 $\left(\frac{K_{\mathrm{a}}}{2}+1\right)\times 3$ 的堆叠卷积中，实际感受野为 $K_{\mathrm{a}} \times 5$，右支路输入卷积核尺寸为 $K_{\mathrm{a}} \times 1$ 和 1×5 的堆叠卷积中，实际感受野也为 $K_{\mathrm{a}} \times 5$，左右分支通过不同卷积核尺寸提取相同感受野的不同尺度信息，使网络可以更好地理解多样化的特征；随后中间分支通过Sigmoid激活函数与两分支相乘以交互学习不同层次特征和相互补偿信息，并将两分支的输出在通道维度上拼接后再输入卷积核尺寸和步长均为1 × 1的二维卷积中以调整输出的通道数；最后经过批次归一化（BN, batch normalization）和PReLU激活函数以加速网络的收敛速度。经过N层MCU模块后得到输出特征 $\boldsymbol{Y}_N^{\mathrm{LAP}} \in \mathbb{R}^{(B \times F) \times C \times \left(\frac{T}{2^N}\right)}$，其中$N$为层数且设置为3，卷积核尺寸$K_{\mathrm{a}}$分别设置为13、9、5。

2.3 视听注意力融合模块

受Transformer[18-19]中的多头自注意力机制和Yu等[20]提出的深度特征融合启发，本文提出一种AVAFM，不再只是处理单纯的语音信号，而是处理经过编码网络的视觉、听觉多模态信号。

如图3所示，左右框内分别是对听觉和视觉的特征进行处理，不同于纯音频的多头注意力，作为听觉映射特征，$\boldsymbol{Q}$、$\boldsymbol{K}$相乘后的输出特征与$\boldsymbol{V}$相乘后经过尺度归一化、Softmax得到视听关联性权重特征值$\mathbf{Atten}_{\mathrm{av}}$，$\boldsymbol{V}$作为视觉映射特征，最后输出视听注意力特征$\mathbf{Atten}'_{\mathrm{av}}$。

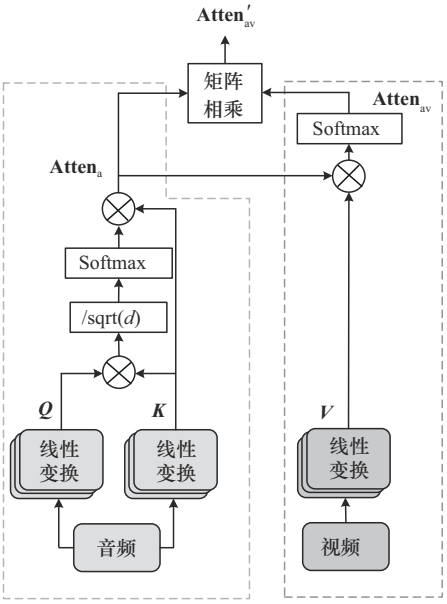


图3 视听注意力融合模块

具体来说，首先将输入的视觉特征$\boldsymbol{X}_{\mathrm{a}}$和听觉特征$\boldsymbol{X}_{\mathrm{v}}$调整为三维并通过线性层映射出查询$\boldsymbol{Q}_{\mathrm{a}}$、键$\boldsymbol{K}_{\mathrm{a}}$、值$\boldsymbol{V}_{\mathrm{v}}$参数矩阵，再将听觉特征映射出的$\boldsymbol{Q}_{\mathrm{a}}$、$\boldsymbol{K}_{\mathrm{a}}$进行矩阵相乘和尺度归一化，并将经过Softmax函数后的特征与键$\boldsymbol{K}_{\mathrm{a}}$相乘得到听觉注意力权重$\mathbf{Atten}_{\mathrm{a}}$，表达式如下

$$\left[\boldsymbol{Q}_{\mathrm{a}},\boldsymbol{K}_{\mathrm{a}}\right]=\left[\boldsymbol{W}^{\boldsymbol{Q}},\boldsymbol{W}^{\boldsymbol{K}}\right]\boldsymbol{X}_{\mathrm{a}} \tag{7}$$

$$\left[\boldsymbol{V}_{\mathrm{v}}\right]=\left[\boldsymbol{W}^{\boldsymbol{V}}\right]\boldsymbol{X}_{\mathrm{v}} \tag{8}$$

$$\mathbf{Atten}_{\mathrm{a}}=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}_{\mathrm{a}}\boldsymbol{K}_{\mathrm{a}}^{\mathrm{T}}}{\sqrt{d}}\right)\boldsymbol{K}_{\mathrm{a}} \tag{9}$$

其中，$\boldsymbol{X}_{\mathrm{a}}$和$\boldsymbol{X}_{\mathrm{v}}$分别为输入的听觉和视觉特征，$\boldsymbol{W}^{\boldsymbol{Q}}$、$\boldsymbol{W}^{\boldsymbol{K}}$、$\boldsymbol{W}^{\boldsymbol{V}}$分别为$\boldsymbol{Q}_{\mathrm{a}}$、$\boldsymbol{K}_{\mathrm{a}}$、$\boldsymbol{V}_{\mathrm{v}}$的参数矩阵，且维度为$B \times C \times C$，再将听觉注意力权重$\mathbf{Atten}_{\mathrm{a}}$与视觉特征值$\boldsymbol{V}_{\mathrm{v}}$进行点乘和Softmax后得到视听觉注意力权重$\mathbf{Atten}_{\mathrm{av}}$，根据特征的贡献度以调整模态间的信息权重，最后将视觉和听觉注意力相乘以学习视听模态间的潜在联系，得到视听注意力$\mathbf{Atten}'_{\mathrm{av}}$，操作如下

$$\mathbf{Atten}_{\mathrm{av}}=\mathrm{Softmax}\left(\mathbf{Atten}_{\mathrm{a}}\cdot\boldsymbol{V}_{\mathrm{v}}\right) \tag{10}$$

$$\mathbf{Atten}'_{\mathrm{av}} = \mathbf{Atten}_{\mathrm{v}} \cdot \mathbf{Atten}_{\mathrm{a}} \tag{11}$$

语音与唇形同步信息的互补，可为2种模态特征建立多层次关联性，在其中一种模态信息模糊时补充上下文信息，增强模型应对复杂环境下的鲁棒性。

2.4　融合加权模块

考虑到仅处理单级AVAFM会丢失前几级的低精度特征，因此在输入LSTM前设计一种可学习的FWB，对多级AVAFM输出特征进行关键信息提取。如图4所示，第一级特征$\boldsymbol{A}_1$经过两层一维卷积，卷积核尺寸均为13，且每层卷积核的步长为2 × 1；第二级特征$\boldsymbol{A}_2$经过卷积核尺寸为9、步长为2 × 1的单层一维卷积，随后将两分支信息相加并输入频率注意力（FA, frequent attention）中，频率注意力的输出$\boldsymbol{A}_{\mathrm{f}}$再与第三级特征$\boldsymbol{A}_3$动态加权得到最后的融合加权输出，其中可学习因子$\alpha$、$\beta \in [0,1]$，将两分支的输出特征加权融合得到网络输出$\boldsymbol{A}$

$$\boldsymbol{A} = \alpha \boldsymbol{A}_{\mathrm{f}} + \beta \boldsymbol{A}_3 \tag{12}$$

其中，α初始值设为0，β初始值设为1。对于FA，首先将输入特征通过卷积核尺寸为3、步长为1 × 1的一维卷积，再经过全局平均池化以聚合频率维度的重要特征和降低过拟合风险，随后经过卷积核尺寸为3、步长为1 × 1的一维卷积和Sigmoid激活函数来学习频率注意力系数。FWB作为视听信息的补偿网络，能够学习到前几级被忽略的关键信息，更有效地感知视听多模态细粒度特征。

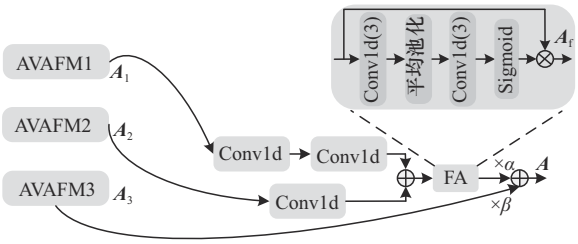


图4　融合加权模块

2.5　网络总体结构

网络总体结构由4个部分组成，即数据预处理、视听编码网络、视听融合网络和听觉解码网络，如图5所示。

首先将语音波形和嘴唇图像进行数据预处理得到语谱图和嘴唇多维数据特征，随后将预处理后的视觉和听觉特征分别输入视觉和听觉编码网络中。听觉特征经过3层MCU来逐级压缩听觉特征的尺寸以降低模型参数量，视觉特征经过3层VCU以逐级压缩视觉特征来与听觉特征尺寸相匹配，其中卷积核的通道数均设置为32。每层视听编码输出再输入AVAFM中学习模态间的相关性信息，随后将三级AVAFM输出特征输入融合加权网络中以补充前两级缺失的低层次信息，再输入单层的LSTM中以学习视听特征的长序列相关性，其中LSTM的隐藏层单元为128，最后经过两层线性层输出增强后的语谱图，经过短时逆傅里叶变换等操作得到增强后的语音波形。

网络总体模型的参数设置如表1所示，其中，K_{v}、K_{a}、C、B、F分别为视觉编码网络的卷积核尺寸、听觉编码网络的卷积核、尺寸通道数、批次数、帧数，F设置为5。

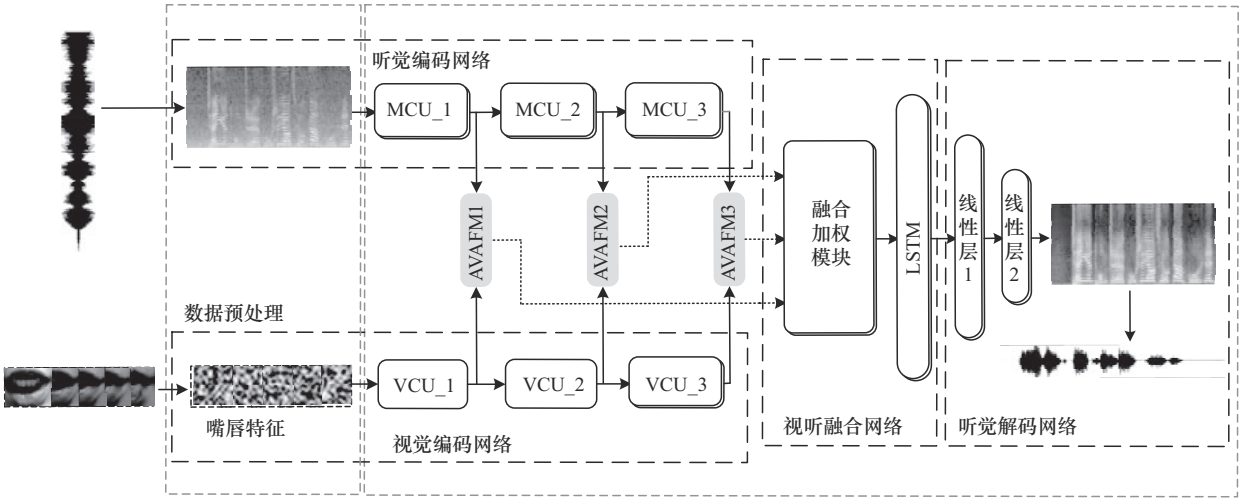


图5　网络总体结构

表1 网络总体模型的参数设置

网络结构	模块	参数设置	输入维度	输出维度
听觉编码网络	MCU_1	$K_a=13$，$C=32$	$B\times1\times256\times F$	$(B\times F)\times32\times128$
	MCU_2	$K_a=9$，$C=32$	$(B\times F)\times32\times128$	$(B\times F)\times32\times64$
	MCU_3	$K_a=5$，$C=32$	$(B\times F)\times32\times64$	$(B\times F)\times32\times32$
视觉编码网络	VCU_1	$K_v=5$，$C=32$	$B\times1\times32\times32\times F$	$(B\times F)\times32\times128$
	VCU_2	$K_v=5$，$C=32$	$(B\times F)\times32\times128$	$(B\times F)\times32\times64$
	VCU_3	$K_v=3$，$C=32$	$(B\times F)\times32\times64$	$(B\times F)\times32\times32$
视听注意力融合模块	AVAFM1	$C=32$	$(B\times F)\times32\times128$	$(B\times F)\times32\times128$
	AVAFM2	$C=32$	$(B\times F)\times32\times64$	$(B\times F)\times32\times64$
	AVAFM3	$C=32$	$(B\times F)\times32\times32$	$(B\times F)\times32\times32$
视听融合网络	融合加权模块	$C=32$	—	$(B\times F)\times32\times32$
	LSTM	Hidden=128	$B\times F\times1024$	$B\times F\times256$
听觉解码网络	线性层1	512	$B\times(F\times256)$	$B\times512$
	线性层2	256	$B\times512$	$B\times1\times256$

3 实验及结果分析

3.1 数据集设置

为验证本文网络的泛化能力、性能提升效果和在真实场景下的性能表现，本文设置了大、小2种规模的数据集分别对本文网络进行训练和测试，并选用多个噪声数据集的多种日常生活噪声以验证真实场景下的语音增强效果。具体数据集划分如下。

1) TMSV视听小数据集

该数据集选自TMHINT（中国台湾普通话噪声环境中听力测试）的TMSV[8]音视频数据集，其中包括18名说话者（13名男性和5名女性）。视频片段是在充足光线下的录音棚内录制的，说话者的音频以48 kHz录制，视频以1080P分辨率的正面视角拍摄，每秒50帧。语音的训练集和验证集由9名说话者（6名男性和3名女性）的第1条至第300条语音构成，总共2 700条纯净语音，噪声部分选自100种非语言噪声库、Noise-92噪声库、Isolated Urban Sound噪声库、Demand噪声库，共涉及32种场景噪声，选择其中23种作为训练集噪声，9种作为测试集噪声，具体噪声统计如表2所示。

纯净语音与23种训练集噪声在−10~10 dB范围内，以1 dB为步长随机选取信噪比混合制成含噪语音，选取训练集的十分之一作为验证集，最后训练集和验证集共产生32 400条含噪语音。为了保持性别比例的平衡，语音的测试集由3名说话者（2名男性和1名女性）的第1条至第250条语音构成，总共750条纯净语音。测试集选用Noise-92噪声库中的Factory2、Leopard等噪声以测试网络在强噪声下的泛化能力，选用Isolated urban sound和Demand噪声库中的噪声，来模拟真实视频通话场景（街道步行、公园散步、公交乘车等），纯净语音与噪声在−10 dB、−7 dB、−4 dB、−1 dB信噪比下混合以制成3 000对纯净−噪声语音。

2) LGRID视听大数据集

该数据集选自LGRID[21]数据集，由30名女性和24名男性说话者构成，每个说话者提供了1 000个

表2 噪声统计

噪声类型	噪声名称
TMSV训练集噪声	Babble、Buccaneer1、F16、Factory1、Pink、Volvo、M109、Meeting、Metro、crowd、schoolyard、Train、Rain、Wind、River、Traffic、Animal、Laugh、Bell、Machine、Cry、Alarm、shower
TMSV测试集噪声	Factory2、Leopard、Hfchannel、White、Ventilation、Park、Plane、Street、Bus
LGRID训练集和测试集噪声	Air conditioners、Car horns、Children playing、Dog barking、Drilling、Engine idling、Gunshots、Jackhammer、Siren、Street music

语句，音频采样率为48 kHz。选取其中的12名男性和12名女性用于训练、验证和测试，随机选取每位说话者的前300条语音并重采样到16 kHz，共计7 200条；为了使男女比例平衡，训练集和验证集选用10名男性和10名女性，测试集选用2名男性和2名女性。噪声数据集选自UrbanSound8K[22]数据集，其由8 732个噪声信号组成丰富的噪声源，包括10种常见的城市噪声，如空调噪声、汽车喇叭声、儿童玩闹声等，随机选取其中80%作为训练集噪声，余下的作为测试集噪声，将所有噪声和纯净语音均截取前3 s，随机选取纯净语音和噪声在信噪比为−10~10 dB、步长为1 dB进行混合以生成训练集和测试集，选取训练集的15%作为验证集，最后训练集产生51 000条语音，验证集产生9 000条语音；测试集中的1 200条纯净语音分别在−5 dB、−2 dB、1 dB、4 dB信噪比下混合以产生4 800对测试集纯净−噪声语音。噪声数据集的详细信息如表2所示。

3.2　损失函数

本文使用的损失函数为计算每一帧上的最小均方误差，具体计算式为

$$\mathrm{Loss}=\frac{1}{M}\frac{1}{N}\sum_{m=1}^{M}\sum_{n=1}^{N}\left(\boldsymbol{Y}_{mn}^{\mathrm{LAP}}-\boldsymbol{S}_{mn}^{\mathrm{LAP}}\right)^{2} \tag{13}$$

其中，M表示一个训练批次中样本迭代的次数($m=1,2,\cdots,M$)，N表示语音帧数($n=1,2,\cdots,N$)，$\boldsymbol{Y}_{mn}^{\mathrm{LAP}}$和$\boldsymbol{S}_{mn}^{\mathrm{LAP}}$分别表示含噪声对数幅度谱和增强后对数幅度谱。

3.3　实验参数设置及评价指标

实验参数设置如下：批次数batch_size为64，轮次数epoch为100。使用Adam优化器进行训练，初始学习率为0.000 5，若验证集损失连续3个训练轮次不降低，则学习率减半，如果验证集的损失连续10个轮次没有降低，则停止训练。实验在RTX3090的GPU上进行，使用PyTorch1.9、CUDA11.4和Python3.7等环境。

评估语音质量时，采用客观指标和主观评价指标，客观指标包括语音质量感知评估（PESQ）和短时客观可懂度（STOI），得分区间分别为[−0.5,4.5]和[0,1]，PESQ的分数值越高表示语音听感越好，STOI的分数值越高表示越容易被理解；主观评价指标包括CSIG（用于信号失真评估）、CBAK（用于噪声失真评估）和COVL（用于总体质量评估），3种主观评价指标的得分均在1~5范围内，分数越高代表语音质量越好。

3.4　实验结果及性能分析

1) TMSV数据集下的对比实验

首先，在小数据集下横向对比不同视听网络模型和本文网络的纯音频语音增强性能，对比的网络为近年来添加视觉信息的视听网络模型，如VSE[7]、LAVSE[9]、L2L[23]等。VSE为基于编解码的多层视听网络，视觉和听觉编码器均使用多层卷积模块，中间层使用多层线性层，听觉解码器使用多层反卷积模块以还原出增强的语音；L2L是应用于语音分离的经典视听网络，视听编码网络使用深层卷积模块，中间层使用LSTM，解码层使用多层线性层来还原语音信号；LAVSE借鉴了L2L结构，在视听编解码网络中使用多层卷积，中间层网络使用LSTM结构，且输出解码层使用视听解码网络以输出增强语音和重建后的嘴唇特征。所有对比网络的视觉和听觉特征参数、网络结构参数、模型参数均与原论文保持一致，且所有对比网络的最优模型均达到收敛状态。图6为本文提出的视听语音增强（AVSE）与纯音频语音增强（AOSE）的训练和验证的最小均方误差（MSE）损失曲线。从图6可以看出，视听语音增强相比纯音频语音增强可以有效地最小化损失误差并达到网络收敛。

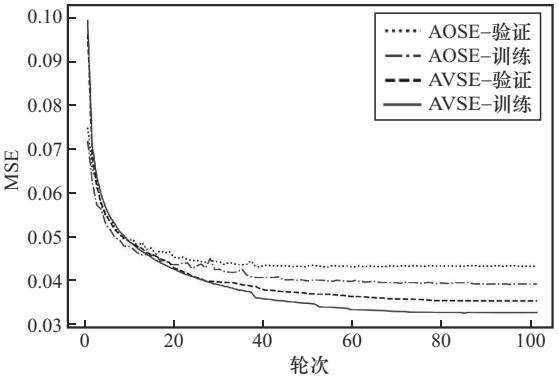


图6　AVSE和AOSE的训练误差和验证误差

图7和图8为TMSV数据集下不同网络模型的PESQ和STOI性能对比。可以看出，在TMSV数据集下，本文网络的性能均优于其他网络模型及纯音频的方法，说明本文提出的视听多模态特征相比单模态特征能够学习到更丰富的信息，对抗噪声的能力较强，且语音增强性能相比其他的网络模型有更好的降噪效果。

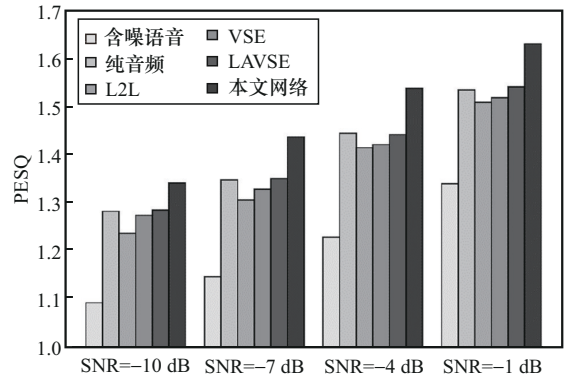


图7　TMSV数据集下不同网络模型的PESQ对比

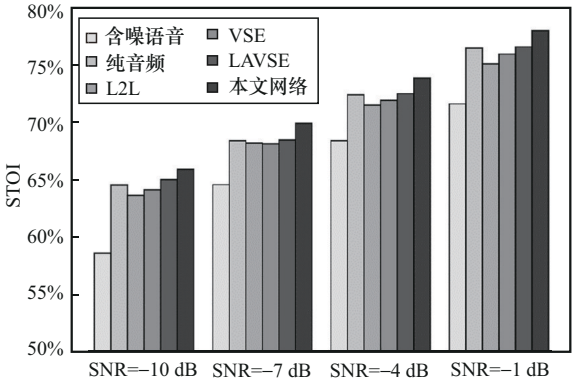


图8　TMSV数据集下不同网络模型的STOI对比

2) LGRID数据集下的消融实验

其次，在大数据集上考察不同输入特征、不同网络模块组合对网络性能的影响。表3和表4对比了不同输入特征、不同信噪比下的PESQ和STOI性能。

对表3和表4中多种信噪比下的性能分析后可以得出，加入不同尺寸的脸部特征比嘴唇特征的性能要差，原因可能是脸部除嘴巴以外的信息会额外地增加噪声干扰，产生较多的冗余信息，从而较难提取到关键特征；64像素×64像素的嘴唇特征与32像素×32像素的嘴唇特征相比性能也稍差，原因可能是大尺寸图像数据的加入使模型更多地关注视觉特征的处理，忽略了听觉特征；16像素×16像素的嘴唇特征性能稍差于32像素×32像素的嘴唇特征的原因可能在于小尺寸特征会丢失部分嘴唇细节特征，无法捕捉到图像细微变化，导致特征提取不准确。此外，通过对比表3和表4中的纯音频特征与不同尺寸的脸部特征可以看出，在低信噪比下的视觉信息是有辅助作用的，但是高信噪比下的性能不如纯音频特征，原因可能在于脸部的冗余信息干扰和高信噪比对于视觉特征的依赖性降低。但对比表3和表4中的纯音频特征与嘴唇特征的性能可以看出，加入嘴唇视觉特征后的性能均要优于只有纯音频的模型性能，说明视觉信息的选择很重要，过多的冗余信息会干扰模型的学习能力。最后，综合上述的实验对比分析可以得出，视听语音增强在低信噪比下可以更好地学习多模态间的特征关联性以提升网络的降噪能力，而在高信噪比下，由于噪声的减弱，视觉信息的贡献相对降低。

表5为不同编码层数下的PESQ和PTOI性能对比。从表5可以看出，在编码层数为3时，PESQ性能最优，STOI稍低于编码层数为4时的性能，但参数量减少0.07×10^6且计算量降低了约10%，因此将编解码层数为3层设置为本文的最优配置，所有对比实验均使用此配置。

表6为消融实验下不同网络模块的编号设置，其中，M1为视觉编码器使用单层卷积模块，听觉编码网络为单支路卷积（SC），SC表示只使用MCU的单个分支；M2表示视觉编码网络使用VCU且视听特征融合方式为普通相加融合；M3表示将单支路卷积替

表3　不同输入特征、不同信噪比下的PESQ性能对比

输入特征	PESQ					
	SNR=−5 dB	SNR=−2 dB	SNR=1 dB	SNR=4 dB	SNR=7 dB	均值
含噪语音	1.141	1.168	1.208	1.267	1.351	1.227
纯音频	1.357	1.448	1.536	1.732	1.892	1.593
脸部特征（16像素×16像素）	1.365	1.466	1.562	1.736	1.859	1.598
脸部特征（32像素×32像素）	1.372	1.475	1.574	1.753	1.883	1.611
脸部特征（64像素×64像素）	1.368	1.472	1.566	1.748	1.865	1.604
嘴唇特征（16像素×16像素）	1.379	1.487	1.593	1.757	1.898	1.623
嘴唇特征（32像素×32像素）	**1.431**	**1.551**	**1.68**	**1.821**	**1.965**	**1.690**
嘴唇特征（64像素×64像素）	1.392	1.493	1.601	1.768	1.908	1.632

表 4　不同输入特征、不同信噪比下的 STOI 性能对比

输入特征	STOI					
	SNR=−5 dB	SNR=−2 dB	SNR=1 dB	SNR=4 dB	SNR=7 dB	均值
含噪语音	57.15%	61.65%	66.04%	70.16%	73.91%	65.78%
纯音频	63.46%	68.19%	71.84%	75.41%	77.71%	71.32%
脸部特征（16 像素×16 像素）	63.99%	69.05%	72.04%	75.12%	77.33%	71.51%
脸部特征（32 像素×32 像素）	64.56%	69.37%	72.49%	75.28%	77.43%	71.83%
脸部特征（64 像素×64 像素）	64.44%	69.23%	72.2%	75.18%	77.39%	71.69%
嘴唇特征（16 像素×16 像素）	65.69%	69.41%	72.81%	75.43%	77.55%	72.18%
嘴唇特征（32 像素×32 像素）	**66.27**%	**70.21**%	**73.35**%	**75.93**%	**77.97**%	**72.75**%
嘴唇特征（64 像素×64 像素）	65.87%	69.62%	72.88%	75.51%	77.59%	72.29%

换为 MCU；M4 为在 M3 基础上添加单层的 AVAFM；M5 为在 M4 基础上编码层使用多层视听注意力融合且使用融合加权模块（无 FA），以补充丢失的初级特征；M6 为在 M5 中的 FWB 中添加 FA。

表 5　不同编码层数下的平均 PESQ 和 STOI 性能对比

编码层数	PESQ	STOI	参数量	计算量
2	1.668	72.42%	2.26×10^6	128.64×10^6
3	**1.690**	72.75%	2.31×10^6	146.85×10^6
4	1.682	**72.78**%	2.38×10^6	163.62×10^6
5	1.673	72.56%	2.46×10^6	181.35×10^6

表 6　消融实验下不同网络模块的编号设置

模型	SC	VCU	MCU	AVAFM	FWB	FA
M1	√					
M2	√	√				
M3		√	√			
M4		√	√	√		
M5		√	√	√	√	
M6		√	√	√	√	√

不同网络模型的消融实验对比如表 7 所示。从表 7 实验结果中可以看出，本文提出的 VCU、MCU、AVAFM、FWB 等模块都能够有效地提升网络的性能。VCU 相比单层卷积可以更深地提取到嘴唇的边缘特征，识别唇形的运动状态；MCU 通过多分支间的交互学习相比单支路可以更好捕捉多角度特征；AVAFM 通过融合视听模态间的特征，有助于学习关联性信息；FWB 可以从多级视听注意力模块中提取低层次特征，以作为信息补充。此外，在 FWB 中加入 FA 注意力后，能够进一步提取低精度特征中的关键信息，从而提升模型的识别和选择能力。

表 7　不同网络模型的消融实验对比

模型	PESQ	STOI	CSIG	CBAK	COVL
M1	1.551	70.11%	2.698	2.227	2.086
M2	1.559	70.43%	2.789	2.236	2.130
M3	1.667	72.15%	2.924	2.328	2.255
M4	1.680	72.62%	2.950	2.342	2.276
M5	1.686	72.69%	2.972	2.364	2.303
M6	**1.690**	**72.75**%	**2.983**	**2.372**	**2.314**

3) LGRID 数据集下的对比实验

最后，在大数据集下横向对比不同网络模型的主、客观评价指标的性能表现，对比的网络模型与小数据集上的网络模型一致。表 8 为不同网络模型的 3 种主观评价指标（CSIG、CBAK、COVL）得分对比。实验结果表明，本文网络的主观评价指标均优于其他网络模型，可知本文网络对纯净语音的还原度最高，抑制噪声的效果、整体听觉质量较好。

表 8　不同网络模型的 3 种主观评价指标得分对比

网络模型	CSIG	CBAK	COVL
含噪语音	1.695	1.586	1.373
L2L[23]	2.528	1.982	1.935
VSE[7]	2.583	2.021	1.965
LAVSE[9]	2.738	2.146	2.081
本文网络	**2.983**	**2.372**	**2.314**

不同信噪比的 PESQ 和 STOI 性能对比如表 9 和表 10 所示。从表 9 和表 10 可以看出，本文网络相比其

表9 不同信噪比的PESQ性能对比

网络模型	PESQ					
	SNR=−5 dB	SNR=−2 dB	SNR=1 dB	SNR=4 dB	SNR=7 dB	均值
含噪语音	1.141	1.168	1.208	1.267	1.351	1.227
纯音频	1.357	1.448	1.536	1.732	1.892	1.593
L2L	1.298	1.386	1.482	1.594	1.703	1.493
VSE	1.306	1.395	1.489	1.601	1.712	1.501
LAVSE	1.338	1.427	1.525	1.629	1.738	1.531
本文网络-L	1.416	1.531	1.669	1.795	1.938	1.670
本文网络	**1.431**	**1.551**	**1.680**	**1.821**	**1.965**	**1.690**

表10 不同信噪比的STOI性能对比

网络模型	STOI					
	SNR=−5 dB	SNR=−2 dB	SNR=1 dB	SNR=4 dB	SNR=7 dB	均值
含噪语音	57.15%	61.65%	66.04%	70.16%	73.91%	65.78%
纯音频	63.46%	68.19%	71.84%	75.41%	77.71%	71.32%
L2L	63.26%	67.76%	70.83%	72.76%	75.12%	69.95%
VSE	63.41%	67.91%	70.98%	72.82%	75.24%	70.07%
LAVSE	64.18%	68.01%	71.11%	73.66%	75.71%	70.53%
本文网络-L	66.02%	69.96%	73.12%	75.65%	77.69%	72.49%
本文网络	**66.27%**	**70.21%**	**73.35%**	**75.93%**	**77.97%**	**72.75%**

他网络，PESQ提升了52.30%~74.06%，STOI提升了46.74%~67.15%，说明本文网络能够更好地抑制噪声，同时性能得到了较好的平衡。随着信噪比的增加，视听语音增强的性能与纯音频语音增强性能的差距在缩小，即在−5~7 dB范围内的PESQ、STOI分别提升26.50%和25.81%，在低信噪比下的−5~1 dB范围内PESQ、STOI分别提升38.95%和33.92%，随着噪声对纯净语音的干扰逐渐变小，视觉特征对网络性能的改善也在慢慢减弱，原因可能在于随着噪声对语音的影响越来越小，网络在受到噪声高干扰的情况下会更多地依赖视觉特征辅助语音以判断背景噪声。

图9为各网络模型每帧输入的参数量和计算量对比。L2L因其过深的卷积层数和较大的卷积核尺寸与数量，参数量和计算量较大；VSE因其巨大的卷积核数和尺寸，以及视听特征拼接后输入多层线性层中导致模型的复杂度很高，参数量和计算量最大；LAVSE因其大尺寸的卷积核与大参数量的LSTM导致其较大的参数量；纯音频没有视觉特征的输入，计算量和参数量均为最低；本文网络的模型中优化了卷积核的尺寸与数量，且输入的LSTM的参数量仅为LAVSE的$\frac{1}{4}$，拥有较小的参数量，但因MCU中较多的卷积运算增加了计算量；本文网络-L是本文视听编码层数仅为2层的网络，复杂度相比本文网络稍低一些。

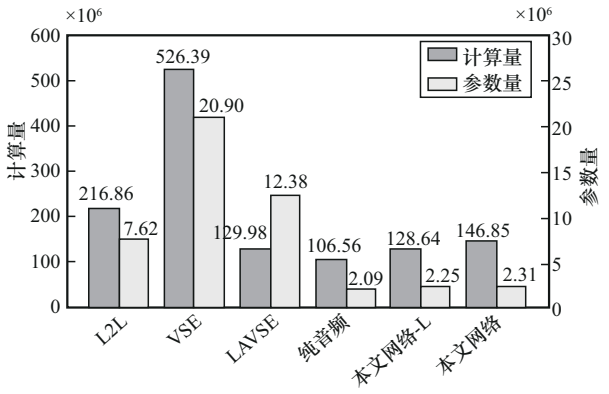


图9 各网络模型每帧输入的参数量和计算量对比

从表9、表10和图9可以看出，本文网络-L相比其他视听模型的计算量和参数量均为最低，与计算量相当的LAVSE相比，PESQ和STOI性能分别提升了45.72%和41.26%。通过对比分析上述模型的计算

量、参数量以及性能提升，本文网络虽然在参数量上稍高于LAVSE，但除纯音频以外，所提出的低配置本文网络-L的复杂度最低，且性能提升远优于其他模型。最后，可以说明本文提出模型的语音可懂度和质量最优，模型最高效。

图10为不同网络模型的语音增强的波形对比，以在−5 dB下随机选取的含犬吠声的男性语音为例，图10(a)、图10(b)、图10(c)分别为含噪语音、纯净语音、纯音频语音增强的波形，图10(d)、图10(e)、图10(f)为3种不同对比网络的语音增强波形，图10(g)为本文网络的波形。从首尾的沉默段可以看出，本文的视听语音增强相比纯音频语音增强能够较好地去除沉默段的噪声，且相比其他网络可以较好地还原出纯净语音，语音的失真度更小，对噪声的抑制效果最好。

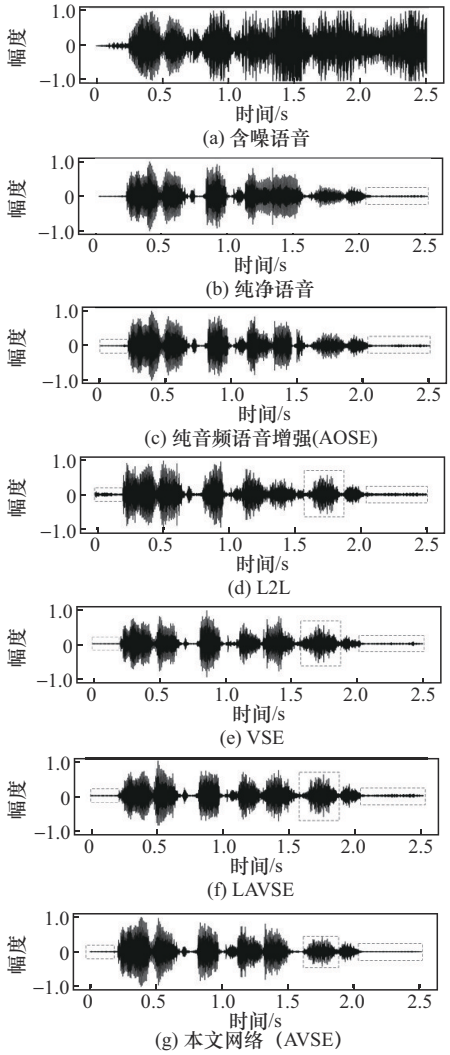


图10　不同网络模型的语音增强的波形对比

4　结束语

本文提出一种在低信噪比下多级音频和视频特征深度融合的视听语音增强网络。在听觉编码网络提出一种多支路协作单元，交互学习多支路的差异性特征以提取更加丰富的听觉特征；在每层视觉和听觉编码网络间提出一种视听注意力融合模块，更好地学习多感官信息间的关联性；在视听融合网络中提出一种融合加权模块，补充学习多层编码网络所遗失的低精度特征。本文方法在TMSV和LGRID视听数据集上进行的实验验证结果表明，在低信噪比下的视听语音增强比纯音频语音增强的平均PESQ和STOI分别提升38.95%和33.92%，证明了视觉信息的添加能够更好地学习到多模态间相关性，更好地消除语音沉默段噪声，且通过对比其他网络，在参数量仅为对比网络的11.05%~30.31%情况下，平均PESQ、STOI分别提升52.30%~74.06%、46.74%~67.15%，表明本文在性能和参数量上取得较好平衡，具有一定的优越性。后续的研究会对视觉信息和网络模块进行优化处理，以进一步降低模型复杂度。

参考文献：

[1] 张睿, 张鹏云, 孙超利. 基于多域融合及神经架构搜索的语音增强方法[J]. 通信学报, 2024, 45(2): 225-239.
ZHANG R, ZHANG P Y, SUN C L. Speech enhancement method based on multi-domain fusion and neural architecture search[J]. Journal on Communications, 2024, 45(2): 225-239.

[2] PARK S R, LEE J W. A fully convolutional neural network for speech enhancement[C]//Proceedings of the Interspeech 2017. Piscataway: IEEE Press, 2017: 1993-1997.

[3] TAN K, WANG D L. A convolutional recurrent neural network for real-time speech enhancement[C]//Proceedings of the Interspeech 2018. Piscataway: IEEE Press, 2018: 3229-3233.

[4] WANG K, HE B B, ZHU W P. CAUNet: context-aware U-Net for speech enhancement in time domain[C]//Proceedings of the 2021 IEEE International Symposium on Circuits and Systems (ISCAS). Piscataway: IEEE Press, 2021: 1-5.

[5] ZION GOLUMBIC E, COGAN G B, SCHROEDER C E, et al. Visual input enhances selective speech envelope tracking in auditory cortex at a “cocktail party”[J]. Journal of Neuroscience, 2013, 33(4): 1417-1426.

[6] ARNAL L H, WYART V, GIRAUD A L. Transitions in neural oscillations reflect prediction errors generated in audiovisual speech[J]. Nature Neuroscience, 2011, 14(6): 797-801.

[7] GABBAY A, SHAMIR A, PELEG S. Visual speech enhancement[C]// Proceedings of the Interspeech 2018. Piscataway: IEEE Press, 2018: 1170-1174.

[8] HOU J C, WANG S S, LAI Y H, et al. Audio-visual speech enhancement using multimodal deep convolutional neural networks[J]. IEEE Transactions on Emerging Topics in Computational Intelligence, 2018, 2(2): 117-128.

[9] CHUANG S Y, TSAO Y, LO C C, et al. Lite audio-visual speech enhancement[C]//Proceedings of the Interspeech 2020. Piscataway: IEEE Press, 2020: 1131-1135.

[10] CHUANG S Y, WANG H M, TSAO Y. Improved lite audio-visual speech enhancement[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022, 30: 1345-1359.

[11] IUZZOLINO M L, KOISHIDA K. AV(SE)^2: audio-visual squeeze-excite speech enhancement[C]//Proceedings of the ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2020: 7539-7543.

[12] XU X M, WANG Y, XU D X, et al. VSEGAN: visual speech enhancement generative adversarial network[C]//Proceedings of the ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2022: 7308-7311.

[13] RAMESH K, XING C, WANG W P, et al. Vset: a multimodal transformer for visual speech enhancement[C]//Proceedings of the ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2021: 6658-6662.

[14] LI Y K, ZHANG X M. Lip landmark-based audio-visual speech enhancement with multimodal feature fusion network[J]. Neurocomputing, 2023, 549: 126432.

[15] HEGDE S B, PRAJWAL K R, MUKHOPADHYAY R, et al. Visual speech enhancement without A real visual stream[C]//Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision (WACV). Piscataway: IEEE Press, 2021: 1926-1935.

[16] 张天骐, 柏浩钧, 叶绍鹏, 等. 基于注意力门控膨胀卷积网络的单通道语音增强[J]. 电子与信息学报, 2022, 44(9): 3277-3288.
ZHANG T Q, BAI H J, YE S P, et al. Monaural speech enhancement based on attention-gate dilated convolution network[J]. Journal of Electronics & Information Technology, 2022, 44(9): 3277-3288.

[17] LI B, WU Z W, WANG Y H. Cross-modal mask fusion and modality-balanced audio-visual speech recognition[C]//Proceedings of the 2022 IEEE 4th International Conference on Power, Intelligent Computing and Systems (ICPICS). Piscataway: IEEE Press, 2022: 371-375.

[18] HAN K, WANG Y H, CHEN H T, et al. A survey on vision transformer[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(1): 87-110.

[19] WANG K, HE B B, ZHU W P. Cptnn: cross-parallel transformer neural network for time-domain speech enhancement[C]//Proceedings of the 2022 International Workshop on Acoustic Signal Enhancement (IWAENC). Piscataway: IEEE Press, 2022: 1-5.

[20] YU B, ZHANG Z, ZHAO D, et al. Audio-visual speech enhancement with deep multi-modality fusion[C]//Proceedings of the 2022 5th International Conference on Information Communication and Signal Processing (ICICSP). Piscataway: IEEE Press, 2022: 143-147.

[21] ALGHAMDI N, MADDOCK S, MARXER R, et al. A corpus of audio-visual Lombard speech with frontal and profile views[J]. The Journal of the Acoustical Society of America, 2018, 143(6): 523-529.

[22] SALAMON J, JACOBY C, BELLO J P. A dataset and taxonomy for urban sound research[C]//Proceedings of the 22nd ACM International Conference on Multimedia. New York: ACM Press, 2014: 1041-1044.

[23] EPHRAT A, MOSSERI I, LANG O, et al. Looking to listen at the cocktail party[J]. ACM Transactions on Graphics, 2018, 37(4): 1-11.

[作者简介]

张天骐（1971– ），男，四川眉山人，重庆邮电大学教授、博士生导师，主要研究方向为通信信号的调制解调、盲处理、图像语音信号处理、神经网络实现以及FPGA、VLSL实现。

沈夕文（2000– ），男，安徽滁州人，重庆邮电大学硕士生，主要研究方向为语音增强、语音信号处理。

唐娟（2000– ），女，四川德阳人，重庆邮电大学硕士生，主要研究方向为卫星扩频信号捕获。

谭霜（2001– ），女，重庆人，重庆邮电大学硕士生，主要研究方向为数字水印、信息隐藏技术。