文章编号：1003-0530(2024)07-1208-10

基于扩张卷积和Transformer的视听融合语音分离方法

刘宏清* 谢奇洲 赵 宇 周 翊
（重庆邮电大学通信与信息工程学院，重庆 400065）

摘 要：为了提高语音分离的效果，除了利用混合的语音信号，还可以借助视觉信号作为辅助信息。这种融合了视觉与音频信号的多模态建模方式，已被证实可以有效地提高语音分离的性能，为语音分离任务提供了新的可能性。为了更好地捕捉视觉与音频特征中的长期依赖关系，并强化网络对输入上下文信息的理解，本文提出了一种基于一维扩张卷积与Transformer的时域视听融合语音分离模型。将基于频域的传统视听融合语音分离方法应用到时域中，避免了时频变换带来的信息损失和相位重构问题。所提网络架构包含四个模块：一个视觉特征提取网络，用于从视频帧中提取唇部嵌入特征；一个音频编码器，用于将混合语音转换为特征表示；一个多模态分离网络，主要由音频子网络、视频子网络，以及Transformer网络组成，用于利用视觉和音频特征进行语音分离；以及一个音频解码器，用于将分离后的特征还原为干净的语音。本文使用LRS2数据集生成的包含两个说话者混合语音的数据集。实验结果表明，所提出的网络在尺度不变信噪比改进（Scale-Invariant Signal-to-Noise Ratio Improvement，SI-SNRi）与信号失真比改进（Signal-to-Distortion Ratio Improvement，SDRi）这两种指标上分别达到14.0 dB与14.3 dB，较纯音频分离模型和普适的视听融合分离模型有明显的性能提升。

关键词：语音分离；视听融合；多头自注意力机制；扩张卷积
中图分类号：TN912.3 **文献标识码：**A **DOI：**10.16798/j.issn.1003-0530.2024.07.003

引用格式：刘宏清，谢奇洲，赵宇，等. 基于扩张卷积和Transformer的视听融合语音分离方法[J]. 信号处理，2024，40(7)：1208-1217. DOI：10.16798/j.issn.1003-0530.2024.07.003.
Reference format：LIU Hongqing，XIE Qizhou，ZHAO Yu，et al. Audio-visual fusion speech separation method based on dilated convolution and Transformer[J]. Journal of Signal Processing，2024，40(7)：1208-1217. DOI: 10.16798/j.issn.1003-0530.2024.07.003.

Audio-visual Fusion Speech Separation Method Based on Dilated Convolution and Transformer

LIU Hongqing* XIE Qizhou ZHAO Yu ZHOU Yi
（School of Communication and Information Engineering，Chongqing University of Posts and Telecommunications，Chongqing 400065，China）

Abstract: To improve the performance of speech separation，other than using the mixed speech signal，the visual signal may also serve as auxiliary information. This multimodal modeling method that integrates visual and audio signals has been proven to effectively improve the performance of speech separation and provides new possibilities for speech separation tasks. To better capture the long-term dependencies in visual and audio features and enhance the network's under-

收稿日期：2023-06-12；修回日期：2023-09-07
*通信作者：刘宏清 hongqingliu@cqupt.edu.cn *Corresponding Author：LIU Hongqing，hongqingliu@cqupt.edu.cn
基金项目：重庆市自然科学基金面上项目(CSTB2022NSCQ-MSX0990)；重庆市教委科学技术研究项目(KJQN202000612)
Foundation Items：Sponsored by Natural Science Foundation of Chongqing，China（CSTB2022NSCQ-MSX0990）；Supported by science and technology research program of Chongqing Education Commission of China（KJQN202000612）

standing of contextual information in the input, this study proposes a time-domain audio-visual fusion speech separation model based on a one-dimensional dilated convolution and Transformer. The traditional audio-visual fusion speech separation method based on the frequency domain is applied to the time domain, avoiding the information loss and phase reconstruction problems caused by time-frequency transformation. The proposed network architecture consists of four modules: a visual feature extraction network, which extracts lip embedding features from video frames; an audio encoder, which converts the mixed speech into feature representation; a multimodal separation network, which consists of an audio subnetwork, video subnetwork, and Transformer network and uses visual and audio features for speech separation; and an audio decoder, which restores the separated features to clean speech. This study uses the LRS2 dataset to generate a dataset containing the mixed speech of two speakers. Experimental results reveal that the proposed network attains 14.0 dB and 14.3 dB improvements in scale-invariant signal-to-noise ratio and signal-to-distortion ratio metrics, respectively, significantly outperforming both the pure audio separation and universal audio-visual fusion models.

Key words: speech separation; audio-visual fusion; multi-head self-attention mechanism; dilated convolution

1　引言

语音分离任务旨在将目标说话者的声音从背景噪声中分离出来，该任务通常也被称为“鸡尾酒会问题”[1-2]。人类的听觉系统具有从多个声源中提取单个声源的能力，因此就算是在鸡尾酒会这样的嘈杂环境中，人类也能毫不费力地在其他说话人和背景噪声中分辨出目标说话人的声音，但这对计算机来说是非常困难的。近年来，基于深度学习的语音分离方法发展迅速，提出了多种纯音频语音分离算法。Hershey等人提出的深度聚类算法（DPCL）[3]在两个和三个说话人的语音分离任务上表现出优异的性能，分离效果超越了传统的基于信号处理或者非负矩阵分解（NMF）的方法。随后Yu等人提出了排列不变性训练准则（PIT）和句子级别的排列不变性训练准则（uPIT）以解决说话人无关的语音分离中的排列组合问题[4-5]。Luo等人提出了一种时域单通道的端到端语音分离网络（Conv-TasNet）[6]，较以往的网络取得了较大的性能提升，奠定了时域单通道语音分离的基础。在Conv-TasNet提出之后，Luo等人设计的双路径循环神经网络（DPRNN）在对较长语音序列的分离中展现出了更好的性能[7]。同年Chen等人提出的双路径Transformer网络（DPTNet）将改进后的Transformer引入到双路径循环神经网络中[8]，使得模型对极长的语音序列建模有效，体现了基于多头自注意力机制的Transformer在语音分离任务中的有效性。但由于纯音频语音分离模型只使用了混合的语音信息，并且需要预先确定目标说话者的数量。因此在目标说话者数量未知的语音分离任务中，纯音频语音分离模型往往表现不佳，削弱了它们应用场景的广泛性。

与纯音频语音分离模型相比，融合了视觉与听觉信息的视听语音分离模型额外提取了目标说话者的视觉特征，在许多实际的场景里，视听语音分离模型已经被证明优于纯音频语音分离模型[9-10]。由于视听融合语音分离模型输入了指定说话者的面部视频帧，因此仅需从混合语音中分离出目标语音，不再需要预先确定说话者数量，同时也避免了标签排列问题。Ephrat提出的与说话人无关的分离模型[11]采用了一种频域视听融合策略，将混合音频信号转换为频域表示，并输入每一位说话者的面部特征，由此预测出每位说话者的音频掩码，最后将掩码与频域表示逐元素相乘得到干净的音频信号，此模型在分离性能与泛化能力上相较以往的纯音频模型有了显著的提升。Lu等人构建了一个能学习声音与唇部运动对应关系的模型[12]，以此辅助仅基于音频的分离模型。Luo等人利用视听输入和i-vector说话者嵌入[13]，将单声道语音信号分解为多个属于不同说话者的语音片段。Gao等人构建了一种新的模型架构[14]，利用说话者的面部运动信息识别说话者，从而帮助模型更好地区分不同的说话者，该模型获得了当时最先进的结果。Wu提出了一种时域视听语音分离模型[15]，其将Conv-TasNet推广到了多模态学习，他提出了一种新的视频编码器，可以从视频流中提取说话者的唇部嵌入，并将其与音频编码器的输出结合起来。Li等人使用一种深度神经网络来提取音频信号和视觉信息的特征[16]，模型使用注意力机制结合音频和视觉信息，最后输出分离后的单个说话者的语音信号。最近，Wu结合了注意力机制在时域对低质量视频进行语音分离[17]，使用注意力机制来帮助模型选择与音频特征相关的视觉特征，性能优于其基线模型。尽管当前的视听语音分离模型已经有了一定的性能提升，但在尺度不变信噪比改进（SI-SNRi）与信号失真比改进（SDRi）这两种客观评价指标上，相较于纯音频语音分离网络，它们的表现还有待提高。并

且,尽管现有的模型同时接收了视觉与音频输入,但却难以充分利用两种信息之间的相关性。此外,视听语音分离网络在处理音频与视频信号时,为了更好地捕捉各模态信息以提高融合效果,对上下文信息的理解和长期关系的捕捉是至关重要的,由于传统的视听语音分离模型通常使用了卷积神经网络(CNN)和循环神经网络(RNN)来处理音频和视频特征,网络架构存在一定的局限性,因此不能很好地捕捉序列数据中的长期依赖关系和理解上下文信息。

针对传统视听分离网络不能很好的理解输入语音的上下文信息与输入图像不同位置之间的关联这两个问题,我们采用了在自创网络中加入扩张卷积的解决方法。针对传统网络对捕捉输入序列中不同位置之间的长距离依赖关系的困难,我们提出了在自创网络的语音与视频支路分别接入Transformer网络的解决方案。同时,鉴于当前视听语音分离模型普遍存在客观评价指标不高的问题,我们提出了一种视听子网络架构提取音频和视频输入的深度特征,并在其中加入了优化后的自定义一维深度可分离卷积,以达到提升客观评价指标的目的。此外,我们在视频子网络中选择了更合适处理图像的归一化方式,并选择了更优的融合策略。最终我们提出了基于一维扩张卷积与Transformer的时域端到端网络结构,结合预训练的唇部嵌入提取器从视频帧中提取唇部嵌入,以获取视觉信息来辅助语音分离任务。一维扩张卷积让模型能够在保持尺寸较小的情况下增大感受野,可以对语音信号的远距离依赖关系进行建模,并通过不同的扩张率来实现多尺度的特征提取,以增强模型的表达能力,这有助于解决堆叠多层卷积神经网络所增加的计算量与网络复杂度的问题。本文模型采用Transformer模块分别对音频与视频支路进行特征提取,利用多头自注意力机制建模音频的长期依赖关系与视频的时序和空间关系,从而更准确地获取多模态的语义信息与长距离依赖关系。实验结果表明,与纯音频语音分离网络Conv-TasNet和现有的普适视听融合分离网络相比,提出网络具有明显的性能优势。

2 基于一维扩张卷积与Transformer的视听分离网络

2.1 视听融合网络整体框架

使用频域方法分离语音时,由于语音信号经过短时傅里叶变换(STFT)后会被转换到频域中,因此分离结果的时间和频率分辨率会受到STFT参数的影响。如果STFT的窗口太大,时间分辨率会变差,而如果窗口太小,频率分辨率会变差。本文提出的基于一维扩张卷积和Transformer的时域视听融合语音分离模型,解决了传统频域语音分离方法存在问题。网络的整体架构如图1所示,提出网络主要由以下四部分组成:视觉特征提取网络由预训练的唇部嵌入提取器组成,用于从视频帧中获取唇部嵌入。音频编码器将输入的混合语音信号进行编码,并转换为特征表示。多模态分离网络主要由音频子网络、视频子网络,以及Transformer网络组成,目的是生成视频中对应说话人的掩码。音频解码器的作用是解码分离后的特征,将该结果重构到时域中,输出对应说话人的干净语音。

输出干净语音的计算过程如式(1)所示。

$$s_i = \mathrm{Decoder}\left(a_c \odot m_i\right) \tag{1}$$

其中,s_i是a_c与m_i逐元素乘法后经由音频解码器输出的对应说话人分离结果,a_c为音频编码器输出的音频特征,m_i为多模态分离网络输出的掩码,a_c和m_i的计算过程如式(2)与式(3)所示。

$$a_c = \mathrm{Encoder}\left(x\right) \tag{2}$$

$$m_i = \mathrm{Separator}\left(v_e, a_e\right) \tag{3}$$

式(2)中的x为输入网络的混合语音信号,式(3)中的v_e和a_e分别为v_c和a_c经过视频与音频预处理得到,其中v_c如式(4)所示。

$$v_c = \mathrm{Extrator}\left(v\right) \tag{4}$$

其中v为输入网络的指定说话者面部视频信号,v_c为v通过视觉特征提取网络得到的唇部嵌入。

2.2 视觉特征提取网络

本文提出的网络旨在分离出指定的一位说话者的纯净语音,视觉特征提取网络的目的是从输入的指定说话者面部视频帧中提取唇部嵌入v_c,通过说话者唇部的运动信息帮助分离网络区分目标说话者的声音以及其他无关说话者的声音,唇部嵌入特征的加入可以有效地辅助语音分离任务,从而更好地帮助网络理解语音信号中的上下文信息。

为解决传统的深度神经网络在层数增加时可能会面临梯度消失或梯度爆炸的问题,在特征提取网络中引入残差网络(ResNet)[18]以帮助训练更深层次的网络。视觉特征提取网络中的唇部嵌入提取器由一个三维卷积层和一个标准的18层ResNet组成,类似于文献[19]中的工作。在训练本文提出网络前,唇部嵌入提取器已经在LRW数据集[20]上完成了预训练,具备了从面部视频帧中提取唇部嵌

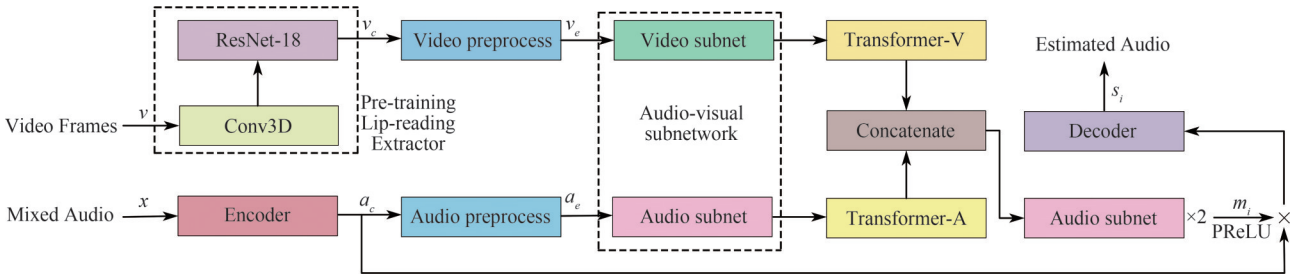


图 1　视听融合语音分离网络整体架构

Fig. 1　Overall architecture of audio-visual fusion speech separation network

入的能力。视觉特征提取网络中的残差网络由17个卷积层和1个全连接层组成。在包含指定说话者面部的视频帧v输入网络后，首先使用三维卷积层对每一个视频帧进行卷积操作，从而对视频帧进行特征提取并建立时空特征表示，然后每一帧通过一个标准的18层ResNet网络，进而生成唇部嵌入特征v_c。

2.3　音频编码器与解码器

音频编码器和解码器旨在实现语音信号的表示转换与信号重建，两者分别对语音信号进行一维卷积和一维转置卷积操作。音频编码器接收并编码原始的混合语音信号x，将语音信号转换为更高级别的特征表示a_c，以便下游的多模态分离网络更好地理解这些信号。音频解码器接收多模态分离网络的输出掩码m_i与特征表示a_c逐元素乘法后的结果，通过一维转置卷积对此结果进行信号重建，将高维度的掩码降维并重构为对应说话者的语音波形s_i。

假设音频编码器输入维度为$1\times T$的混合语音信号，T表示输入语音的长度。为了让提出模型更容易处理较长的语音输入，我们将输入信号切分为S秒一段，采样率设置为R Hz，并在训练时丢弃那些长度小于2 R个采样点的语音段。音频编码器采用一维卷积对输入语音进行编码，假设卷积输出通道数为N，卷积核的长度为L。为减少输出的时间步长，减少计算复杂度，设置卷积步长为$L/2$，最终音频编码器得到维度为$N\times(2T/L)$的混合语音嵌入a_c。音频解码器中的一维转置卷积与编码器采取相反的操作，其中卷积参数的设置与编码器相同，最终输出对应说话者的纯净语音。

2.4　多模态分离网络

在音频编码器的输出a_c与唇部嵌入v_c输入多模态分离网络前，为简化计算复杂度，并保留关键的特征信息，a_c与v_c分别输入语音与视频预处理模块从而生成a_e与v_e。语音预处理模块包括一个全局归一化层（gLN）和一个一维逐点卷积，用于减小样本间的差异并提供有用的语音特征，最后输出a_e。视频预处理模块由一个一维卷积组成，旨在对v_c进行降维操作，较低的维度可提供更紧凑的特征表示，使分离网络识别并捕捉唇部运动的关键信息。

多模态分离网络旨在预测指定说话者的掩码m_i，本文提出了一种结合一维扩张卷积与Transformer的多模态分离网络。在音频子网络和视频子网络对a_e与v_e分别进行深度特征提取后，将音频与视频的特征分别输入Transformer网络，以提取它们不同时间步之间的依赖关系，并进行语义建模与上下文编码。为跨模态融合视听信息，在拼接融合模块，首先使用最近邻插值法将v_e在时间维度上对齐a_e，然后在特征维度上，将插值后的v_e与a_e进行拼接操作，最后通过一维卷积层降维得到与融合前维度相同的音频特征。此时的音频特征已融合了视觉信息。视觉特征由经过预训练的唇部嵌入提取器提取，具有良好的初始表示能力，而音频特征则需要进行更多的迭代和交互，以更好捕捉音频特征的时序和相关性。为提高模型对音频信号的表征能力和泛化能力，融合后的音频特征输入两层堆叠的音频子网络进一步提取特征，最终输出估计的掩码m_i。下面详细介绍多模态分离网络中的音频子网络、视频子网络，以及Transformer网络。

2.4.1　视听子网络

视听子网络由视频子网络与音频子网络组成，此网络的具体结构如图2所示，图中的红色与蓝色箭头分别表示了音频子网络与视频子网络中输入特征的传递方向。

音频子网络与视频子网络分别接收a_e与v_e作为输入，两者分别输入Audio Conv1D模块与Video Conv1D模块，在从一维点卷积模块输出后，两者分别输入音频与视频的自定义一维深度可分离卷积，在输出视听子网络时保持输入与输出维度不变。视听子网络的核心是一维扩张卷积，扩张卷积通过在卷积核的元素间插入空洞来增大卷积核大小，在没有额外增加参数量的情况下，提高卷积核的感受野从而促进模

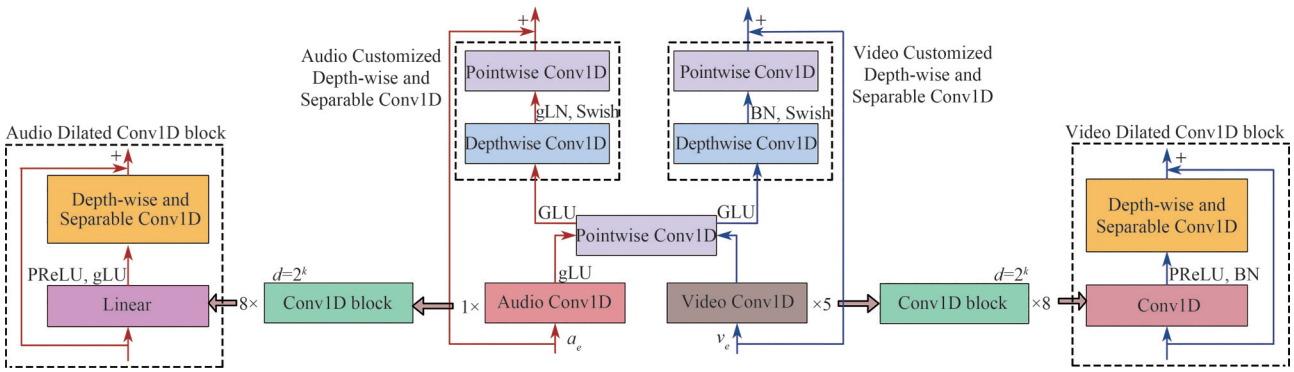


图2 视听子网络结构

Fig. 2 Structure of audio-visual subnet

型学习更多的上下文信息。空洞的大小由膨胀率决定,膨胀率为1的扩张卷积与普通卷积相同。

音频子网络主要由以下几部分组成:一个Audio Conv1D模块、全局层归一化(gLN)、一维逐点卷积、门控线性单元(GLU)以及音频自定义一维深度可分离卷积。Audio Conv1D模块采用了8层包含一维扩张卷积的一维卷积块,扩张卷积指数增长的膨胀率d为2^k,其中$k\in\{0,1,\cdots,7\}$。每个一维卷积块由以下几部分组成:线性层、激活函数PReLU、gLN以及一维深度可分离卷积。一维卷积块中的深度可分离卷积[21]展现出卷积的膨胀性,在深度可分离卷积中,深度卷积的卷积核间距为2^k并且依次增大,以提取到更广泛的音频特征信息。在Audio Conv1D模块对a_e进行深度特征提取后,对音频特征进行gLN操作以减少特征之间的偏差,并提高模型对不同音频样本的泛化能力。随后的一维逐点卷积对特征进行线性变换,以提取时间序列特征。GLU通过门控机制对输入进行筛选,从而选择和强调有用的音频信息。设计自定义的一维深度可分离卷积旨在更好地处理本文中提取语音特征的任务。音频子网络中的自定义一维深度可分离卷积在原深度可分离卷积的基础上添加了gLN与自适应激活函数Swish。其中,基于整个样本集进行归一化的gLN操作在处理音频等序列数据时,进一步地处理不同时间步之间的依赖关系。相较于传统的ReLU激活函数,包含可训练参数的Swish激活函数可以自适应地调整函数的形态,使得在深层模型上表现更好。实验部分证实了视听子网络中自定义一维深度可分离卷积对提出网络带来的提升。

视频子网络的架构与音频子网络相似,但在设计时做了针对性的改变使得视频子网络更适合处理视觉特征。实验部分验证了视频子网络中的批量归一化操作相较于全局层归一化操作的适用性,因此本文中视频子网络中的归一化方法采用批量归一化。为提取图像中更高级别的特征,以达到更好的泛化效果,我们在网络中堆叠了5个Video Conv1D模块。我们将一维卷积块中的线性层替换为了卷积操作,以减少参数量与计算量,并更好地利用局部信息。

2.4.2 Transformer网络

Transformer通常由编码器与解码器两部分组成[22],编码器和解码器都由多层网络堆叠而成。其中编码器的输出是对应输入序列的特征表示,这些特征表示可以被继续用于下游任务,因此Transformer的编码器相当于一种特征提取器。本文中的Transformer网络特指编码器部分,网络架构如图3所示,主要包含:多头自注意力机制、层归一化(LN)、残差连接和全连接前馈网络。

在本文中,音频支路和视频支路的Transformer网络仅在它们的输入维度上有所不同。Transformer网络由两个相同的层堆栈组成,网络的输入

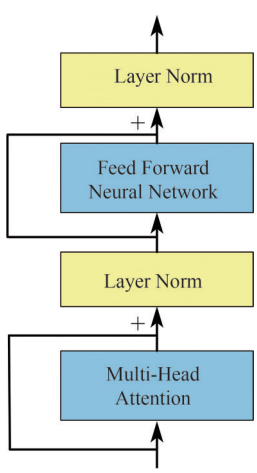


图3 Transformer网络结构

Fig. 3 Structure of Transformer network

与输出的维度保持不变。我们将多头注意力的头数设置为h，每个头会对输入的序列进行一次自注意力操作，得到h种不同的注意力分布，再将这几种注意力分布拼接起来，以得到最终的注意力分布。多头自注意力机制可以在不同的特征子空间中对输入序列进行关注，进而提高模型鲁棒性。由于提出的网络层数较深，为减少模型的过拟合，我们为多头自注意力层与前馈网络层设置了随机失活（dropout）概率。通过Transformer网络中的多头自注意力机制，获得了音频特征的注意力分布，从而提取输入音频特征中的语义信息。由于音频序列具有时序特征，多头自注意力机制同时还能学习序列中的长期依赖关系。对于输入的唇部视觉特征，Transformer网络同时对其时序关系与空间特征进行建模，以在下游任务中更准确地预测出输入视频帧所对应说话者的声音信号。

3　实验结果与分析

3.1　训练目标与评价指标

本网络的训练目标是最大化尺度不变信噪比（SI-SNR），为计算估计信号与原始信号之间的差异，我们使用负的尺度不变信噪比作为损失函数。尺度不变信噪比定义为：

$$s_{\text{target}}=\frac{\langle s_i,s\rangle s}{\|s\|^2} \tag{5}$$

$$e_{\text{noise}}=s_i-s_{\text{target}} \tag{6}$$

$$\text{SI-SNR}=10\log_{10}\frac{\|s_{\text{target}}\|^2}{\|e_{\text{noise}}\|^2} \tag{7}$$

其中，s与s_i分别为原始语音信号和估计语音信号。为确保指标的尺度不变性，计算前将s与s_i进行零均值归一化处理。

我们将尺度不变信噪比改进（SI-SNRi）与信号失真比改进（SDRi）作为模型分离性能的客观评价指标。这两个指标分别根据尺度不变信噪比与信号失真比（Signal-to-Distortion Ratio，SDR）计算得出，两者分别定义为式（8）与式（9）：

$$\text{SI-SNRi}=\text{SI-SNR}(s_i,s)-\text{SI-SNR}(x,s) \tag{8}$$

$$\text{SDRi}=\text{SDR}(s_i,s)-\text{SDR}(x,s) \tag{9}$$

其中，s与s_i为说话者的原始语音与网络输出的估计语音，x为输入网络的混合语音信号。

3.2　数据集

本文在实验中使用开源的LRS2数据集[23]建立了视听融合语音分离数据集，LRS2数据集中的每个语句均来自于BBC电视节目，且每个语句的长度不超过100个字符。通过混合LRS2数据集中单人说话者的语音，创建了包含两个说话者混合语音的数据集。其中，首先从LRS2数据集中随机选择不同的说话者，然后使用−5 dB到5 dB之间的随机信噪比将语音混合以得到混合语音。数据集中每段语音的长度为2 s，采样率设置为16000 Hz，并在训练时丢弃长度小于32000个采样点的语音段。包含说话者面部的视频以每秒25帧的速度采样2 s，以保证和语音同步。最终生成了包含20000条混合语音的训练集、5000条混合语音的验证集以及3000条混合语音的测试集，这三个集合中的说话者互不重叠。

3.3　实验参数设置

在视觉特征提取网络中，我们使用卷积核大小为5×7×7，步幅大小为1×2×2的三维卷积层。在音频编码器与解码器中，我们将卷积输出通道数N设置为512，卷积核长度L设置为20。将视听子网络中扩张卷积的卷积核大小设置为3。我们使用初始学习率为1×10^{-3}的AdamW优化器，优化器的权重衰减设置为0.1，当验证集的损失在连续的4个周期内没有减小时，我们将学习率减半。实验设置150个训练周期，当连续12个周期内验证损失均未减小时，实验将提前终止。模型超参数的总结如表1所示。本文提出的模型在2×NVIDIA RTX A6000 32G GPUs上进行训练。

表1　实验中模型的超参数设置

Tab. 1　Hyper-parameters setting of the model in the experiment

符号	参数描述	参数值
N	编码器输出通道数	512
L	编码器中的卷积核大小	20
v_c	唇部嵌入特征维度	512
a_e	语音预处理输出维度	512
v_e	视频预处理输出维度	64
d	扩张卷积的膨胀率	1,2,…,128
R	语音采样率	16000
h	多头注意力的头数	4
Dropout	随机失活率	0.1
S	语音分段长度/s	2.0
Batch-size	批量大小	5
Num-workers	并行工作数量	8

3.4　结果评估

为了验证本文提出的基于扩张卷积和Transformer的语音分离模型中相关模块的有效性，我们设计了多组消融实验，实验的结果均证实了相关模块对网络性能的提升。首先，通过实验以验证本文核心的扩张卷积操作与Transformer模块的效果。

为验证在时域特征提取中的扩张卷积操作对模

型性能的影响，我们移除了视听子网络中一维卷积块所包含的扩张卷积操作，同时保持网络其他结构与参数不变，并与完整的模型进行性能对比。模型的对比如表 2 所示，本文提出的模型在 SI-SNRi 与 SDRi 上，相较于移除扩张卷积操作的模型均有 2.1 dB 的性能提升。这凸显了扩张卷积在网络设计中的关键作用，因为其提高了卷积核的感受野从而促使模型学习到更多的上下文信息，增强了模型对整体输入数据的理解能力，从而达到能好的性能表现。

为了检验 Transformer 网络对模型性能的影响，我们设置了一组对比实验。保持网络其他结构与参数不变，将完整的模型与仅移除音频与视频支路上 Transformer 网络的模型进行性能比较，模型的比较如表 2 所示。由结果可知，本文提出模型在 SI-SNRi 与 SDRi 上，相较于移除 Transformer 网络的模型均有 0.4 dB 的性能提升。这证明，Transformer 网络可以显著提升提出模型的分离性能，因为多头自注意力机制可以捕捉输入音频特征与视频特征的全局信息，为下游网络输入自注意力机制处理后的特征，从而使下游网络可以更好地预测音频信号。

表 2 提出模型与移除模型模块的消融效果比较

Tab. 2 Comparison of the ablation effects between the proposed model and the model with removed module

模型	SI-SNRi/dB	SDRi/dB
移除扩张卷积	11.9	12.2
移除自定义模块	12.8	13.1
原模块替换为普通模块	13.1	13.4
移除 Transformer	13.6	13.9
提出模型	**14.0**	**14.3**

针对视听子网络中自定义的一维深度可分离卷积与视频子网络中采用的批量归一化操作，分别进行了消融实验以验证设计部分的有效性。

为验证视听子网络中自定义的一维深度可分离卷积对模型性能的影响，我们将提出模型、移除音频子网络与视频子网络中自定义的一维深度可分离卷积模块的模型和将其仅替换为不含归一化操作与激活函数的普通一维深度可分离卷积的模型进行对比。实验时保持视听子网络中其他结构与所有参数不变，表 2 展示了性能对比结果。本文提出模型在 SI-SNRi 与 SDRi 上，相较于移除自定义模块的模型均有 1.2 dB 的性能提升，相较于替换为普通模块的模型均有 0.9 dB 的性能提升。有效证明了模块设计的有效性，同时也证明了自定义的深度可分离卷积相较于普通的深度可分离卷积对本网络的适用性。

为验证视频子网络中的批量归一化操作相较于全局层归一化操作的适用性，我们将提出模型与使用全局层归一化操作的视频子网络模型进行性能比较，并且在消融实验中加入不包含任何归一化操作的视频子网络模型，以验证归一化操作是否改进了该子网络。所有实验保持网络结构与参数不变，结果如表 3 所示。本文提出模型在 SI-SNRi 与 SDRi 上，相较于不含归一化的模型均有 1.0 dB 的性能提升，相较于全局层归一化的模型均有 0.7 dB 的性能提升。证明了归一化操作的加入对视频子网络性能的提升，同时证明了本模型采用的批量归一化操作相较于全局层归一化操作对视频子网络的适用性。

表 3 视频子网络中不同归一化方式的比较

Tab. 3 Comparison of different normalization methods in video subnet

模型	SI-SNRi/dB	SDRi/dB
不含归一化	13.0	13.3
全局层归一化	13.3	13.6
批量归一化	**14.0**	**14.3**

我们设置了一组对比实验以选择提出网络的视听特征融合策略。在实验中对比了拼接与求和这两种特征融合方法，并保持网络其他结构与参数不变，表 4 展示了两种方法的结果。采用拼接方法的提出网络相较于采用求和的模型在 SI-SNRi 与 SDRi 上均有 0.2 dB 的性能提升。求和方法可能无法充分利用模态之间的相关性，而拼接方法可以增加特征维度与提供全面的特征表示，因此在提出模型的多模态特征融合中具有优势。

表 4 拼接与求和的比较

Tab. 4 Comparison between Concatenation and Summation

模型	SI-SNRi/dB	SDRi/dB
求和方法	13.8	14.1
拼接方法	**14.0**	**14.3**

为定量评估本文提出模型与其他模型的性能差异，我们使用 SI-SNRi 与 SDRi 这两种指标，对比了之前提出的纯音频语音分离模型与音视频融合语音分离模型。使用与本文相同的数据集训练并评估了文献[6]提出的纯音频语音分离模型 Conv-TasNet，其与本文提出模型的对比如表 5 所示。同时，在从两位说话者中进行语音分离这一任务下，与一些近年来基于 LRS2 数据集的其他视听融合分

表5　提出模型与纯音频语音分离模型比较

Tab. 5　Comparison between the proposed model and audio-only speech separation model

模型	SI-SNRi/dB	SDRi/dB
Conv-TasNet[6]	10.3	10.7
本文提出模型	**14.0**	**14.3**

离模型进行了性能对比，此项对比中所使用结果均来自原始论文，原始论文中未给出的结果均使用“-”替代。与其他模型的相关对比如表6所示。

由表5可知，相较于Conv-TasNet网络，本文提出网络性能明显更优，在SI-SNRi与SDRi上分别有3.7 dB与3.6 dB的性能提升。时域语音分离模型Conv-TasNet有着优秀的泛化能力与分离性能，这证明在时域中，相较于传统的纯音频语音分离模型，本文提出的有视频输入作为辅助的视听融合模型具有明显优势。

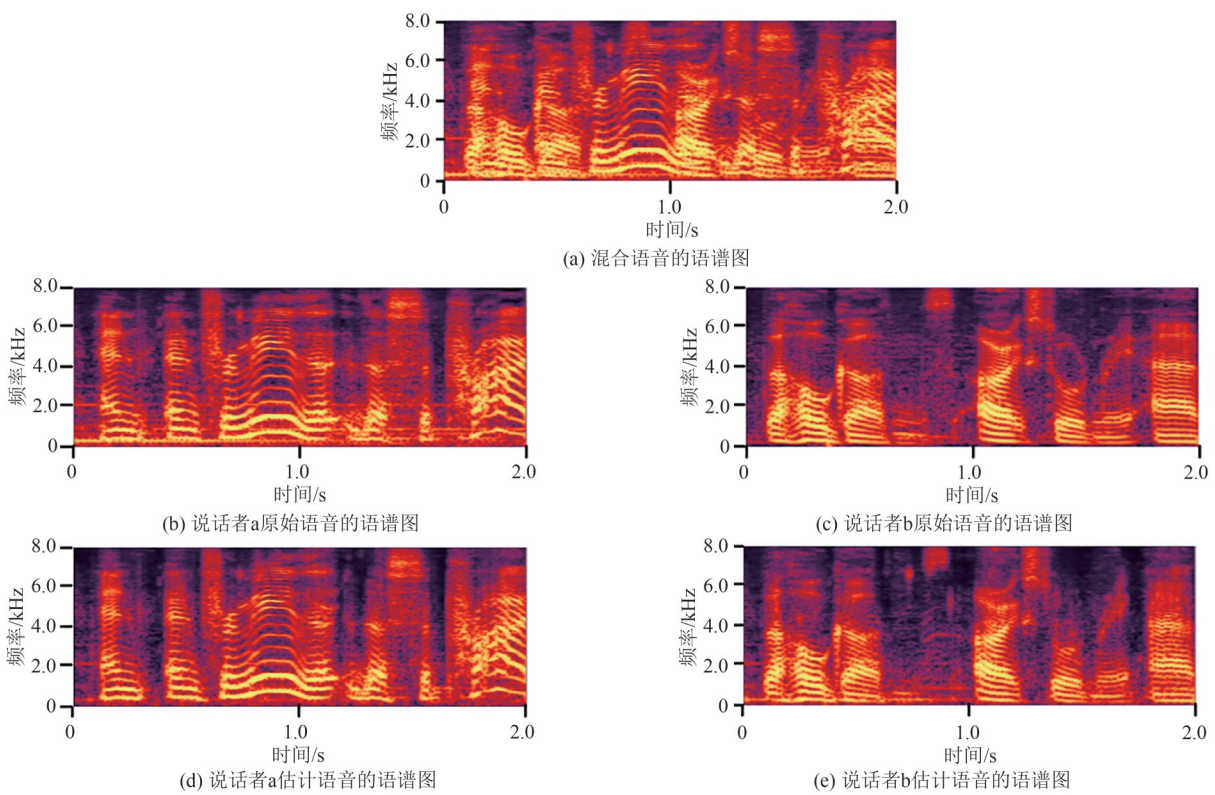
表6　提出模型与视听融合语音分离模型比较

Tab. 6　Comparison between the proposed model and audio-visual speech separation model

模型	SI-SNRi/dB	SDRi/dB
LWTNet[24]	-	10.8
文献[25]中的模型	-	11.6
CaffNet-C[9]	-	12.5
文献[15]中的模型	13.3	-
本文提出模型	**14.0**	**14.3**

由表6可知，相较于其他基于LRS2数据集的多模态模型，本文提出模型仍存在不同程度的性能优势。对比频域视听融合处理模型LWTNet[24]与CaffNet-C[9]，本文提出模型在SDRi上分别取得了3.5 dB与1.8 dB的性能提升，展现出本文在时域进行分离任务的优势。相较于文献[25]中采取双向长短时记忆网络的时域分离模型，提出模型在SDRi上取得了2.7 dB的性能增益，与文献[15]中受Conv-TasNet启发的时域视听分离模型对比，提出模型在SI-SNRi上仍然取得0.7 dB的性能提升。这证明，在时域条件下，本文中加入Transformer结构并多次堆叠一维扩张卷积的设计思路，可以更深度地提取特征并关注到更多有用的信息，从而更好地服务于下游的融合模块，使得融合模块从音频和视频中融合更多的有用特征，以提高模型的分离性能。

为使分离后的结果可视化，在模型训练完成后，我们在测试集中随机选择了一条混合语音，并



(a) 混合语音的语谱图

(b) 说话者a原始语音的语谱图

(c) 说话者b原始语音的语谱图

(d) 说话者a估计语音的语谱图

(e) 说话者b估计语音的语谱图

图4　提出模型分离结果的语谱图

Fig. 4　Spectrogram of the separation results of the proposed model

进行了两次分离实验以分别分离出两位说话者的估计语音。图 4 展示了混合语音、原始语音与估计语音的语谱图。从图 4 可以看出，说话者 a 与说话者 b 分离后的估计语谱图有效的恢复了原始语谱图，与客观指标一致。

4 结论

本文针对现有网络对上下文信息理解困难和对长期关系捕捉困难的问题，创新的提出了一种基于一维扩张卷积与 Transformer 的时域视听融合语音分离模型。提出模型同时接收混合语音与待分离说话者的面部视频帧作为输入，从视频帧中提取出唇部特征以辅助语音分离任务，并直接输出预测的指定说话者的纯净语音。通过利用 Transformer 网络，提出模型弥补了现有视听语音分离模型对序列的长期依赖关系关注不够的问题，同时多次堆叠一维扩张卷积，强化了模型对输入语音的上下文信息与输入图像不同位置之间关联的理解，创新提出的视听子网络架构提升了对音频和视频特征的深度提取效果。消融实验的结果显示，Transformer 网络、扩张卷积操作、创新设计的自定义模块与视频子网络中的归一化方式均对本文提出模型有明显的性能提升。提出模型在 SI-SNRi 与 SDRi 这两种指标上，明显优于纯音频语音分离模型 Conv-TasNet，对比现有基于 LRS2 数据集的视听语音分离模型也有着不同程度的性能提升。后续工作将对比本文中的 Transformer 网络与跨模态注意力机制对模型性能的影响，并继续对特征融合模块深入研究，进一步优化本文提出的方法，设计出分离性能更好的视听融合语音分离模型。

参考文献

[1] ZHANG Yixuan, CHEN Zhuo, WU Jian, et al. Continuous speech separation with recurrent selective attention network[C]//ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore, Singapore. IEEE, 2022: 6017-6021.

[2] SUBAKAN C, RAVANELLI M, CORNELL S, et al. Real-M: Towards speech separation on real mixtures [C]//ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore, Singapore. IEEE, 2022: 6862-6866.

[3] HERSHEY J R, CHEN Zhuo, LE ROUX J, et al. Deep clustering: Discriminative embeddings for segmentation and separation[C]//2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Shanghai, China. IEEE, 2016: 31-35.

[4] YU Dong, KOLBÆK M, TAN Zhenghua, et al. Permutation invariant training of deep models for speaker-independent multi-talker speech separation [C]//2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New Orleans, LA, USA. IEEE, 2017: 241-245.

[5] KOLBÆK M, YU Dong, TAN Zhenghua, et al. Multi-talker speech separation with utterance-level permutation invariant training of deep recurrent neural networks [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2017, 25(10): 1901-1913.

[6] LUO Yi, MESGARANI N. Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 27(8): 1256-1266.

[7] LUO Yi, CHEN Zhuo, YOSHIOKA T. Dual-path RNN: Efficient long sequence modeling for time-domain single-channel speech separation[C]//ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain. IEEE, 2020: 46-50.

[8] CHEN Jingjing, MAO Qirong, LIU Dong. Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation[C]//Interspeech 2020. ISCA: ISCA, 2020: 2642-2646.

[9] LEE J, CHUNG S W, KIM S, et al. Looking into your speech: Learning cross-modal affinity for audio-visual speech separation[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, TN, USA. IEEE, 2021: 1336-1345.

[10] MONTESINOS J F, KADANDALE V S, HARO G. A cappella: Audio-visual singing voice separation[EB/OL]. 2021: arXiv: 2104.09946. https://arxiv.org/abs/2104.09946.

[11] EPHRAT A, MOSSERI I, LANG O, et al. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation [EB/OL]. 2018: arXiv: 1804.03619. https://arxiv.org/abs/1804.03619.

[12] LU Rui, DUAN Zhiyao, ZHANG Changshui. Listen and look: Audio-visual matching assisted speech source separation[J]. IEEE Signal Processing Letters, 2018, 25(9): 1315-1319.

[13] LUO Yiyu, WANG Jing, WANG Xinyao, et al. Audio-visual speech separation using I-vectors[C]//2019 IEEE 2nd International Conference on Information Communication and Signal Processing (ICICSP). Weihai, China. IEEE, 2020: 276-280.

[14] GAO Ruohan, GRAUMAN K. VisualVoice: audio-visual speech separation with cross-modal consistency [C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, TN, USA. IEEE, 2021: 15490-15500.
[15] WU Jian, XU Yong, ZHANG Shixiong, et al. Time domain audio visual speech separation[C]//2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). Singapore. IEEE, 2020: 667-673.
[16] LI Chenda, QIAN Yanmin. Deep audio-visual speech separation with attention mechanism[C]//ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain. IEEE, 2020: 7314-7318.
[17] WU Yifei, LI Chenda, BAI Jinfeng, et al. Time-domain audio-visual speech separation on low quality videos [C]//ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore, Singapore. IEEE, 2022: 256-260.
[18] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA. IEEE, 2016: 770-778.
[19] AFOURAS T, CHUNG J S, ZISSERMAN A.The conversation: Deep audio-visual speech enhancement[C]// Interspeech 2018.ISCA: ISCA, 2018: 1400-1404.
[20] CHUNG J S, SENIOR A, VINYALS O, et al. Lip reading sentences in the wild[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA. IEEE, 2017: 3444-3453.
[21] CHOLLET F. Xception: Deep learning with depthwise separable convolutions[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA. IEEE, 2017: 1800-1807.
[22] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[EB/OL]. 2017: arXiv: 1706.03762. https://arxiv.org/abs/1706.03762.
[23] AFOURAS T, CHUNG J S, SENIOR A, et al. Deep audio-visual speech recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44 (12): 8717-8727.
[24] AFOURAS T, OWENS A, CHUNG J S, et al. Self-supervised learning of audio-visual objects from video [EB/OL]. 2020: arXiv: 2008.04237. https://arxiv.org/abs/2008.04237.
[25] TRUONG T D, DUONG C N, DE VU T, et al. The right to talk: An audio-visual transformer approach[C]// 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, QC, Canada. IEEE, 2022: 1085-1094.

作者简介

刘宏清　男，1980年生，黑龙江人。重庆邮电大学教授，博士，主要研究方向为统计信息处理、语音信号处理。
E-mail：hongqingliu@cqupt.edu.cn

谢奇洲　男，1999年生，四川人。重庆邮电大学通信与信息工程学院硕士研究生，主要研究方向为深度学习、语音信号处理。
E-mail：857806593@qq.com

赵　宇　男，1987年生，黑龙江人。重庆邮电大学讲师，博士，主要研究方向为阵列信号处理、语音信号处理。
E-mail：zhaoyu@cqupt.edu.cn

周　翊　男，1974年生，四川人。重庆邮电大学教授，博士，主要研究方向为语音信号处理、参数估计。
E-mail：zhouy@cqupt.edu.cn

（责任编辑：刘建新）