基于DCNN和BiLSTM的单通道视听融合语音分离方法研究

兰朝凤,王顺博,郭小霞,韩玉兰,康守强
(哈尔滨理工大学测控技术与通信工程学院,黑龙江哈尔滨 150080)

摘 要: 近年来,随着语音处理及计算机技术的飞速发展,人机语音交互的重要性日益突出. 其中,语音分离是将目标语音从混合语音中分离出来的一项重要任务. 然而,在著名的"鸡尾酒会"等复杂开放环境下语音的分离远没有达到令人满意的效果. 针对现实生活中多说话人交流场景,本文以空洞卷积(Dilated Convolutions Neural Network,DCNN)和双向长短时记忆(Bi-directional Long Short-Term Memory,BiLSTM)为网络基础,提出一种视听融合的语音分离(DCNN-BiLSTM)模型. 该模型在训练过程中通过音频编号查找与之对应的视觉信息,视觉信息可以将音频聚焦在说话场景中该说话人上,以达到增强语音分离效果. 在AVSpeech数据集上进行实验测试,利用PESQ(Perceptual Evaluation of Speech Quality)、STOI(Short-Time Objective Intelligibility)和SDR(Signal-to-Distortion Ratio)指标评价分离效果. 研究表明,本文方法比经典的AVSpeech分离方法在语音分离能力上提高了3.37 dB.

关键词: 视听融合;空洞卷积;双向长短时记忆网络;单通道;语音分离
基金项目: 黑龙江省自然科学基金联合引导项目(No.LH2020F033);国家自然科学基金青年基金(No.11804068)
中图分类号: TP391.9 **文献标识码:** A **文章编号:** 0372-2112(2023)04-0914-08
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20210726

A Single Channel Audio-Visual Fusion Speech Separation Method Based on DCNN and BiLSTM

LAN Chao-feng, WANG Shun-bo, GUO Xiao-xia, HAN Yu-lan, KANG Shou-qiang
(*School of Measurement and Control Technology and Communication Engineering, Harbin University of Science and Technology, Harbin, Heilongjiang* 150080, *China*)

Abstract: In recent years, with the rapid development of speech processing and computer technology, it is becoming more and more prominent that the importance of human-computer speech interaction. Among them, speech separation is an important task to separate target speech from mixed speech. However, in the famous "Cocktail Party" and other complex open environment, the separation of speech is far from achieving satisfactory results. For the multi-speaker scenarios in real life, this paper is based on dilated convolutions neural network and bi-directional long short-term memory network, and presents an audio-visual fusion speech separation model DCNN-BiLSTM. In the training process, the model searches for the corresponding visual information through the audio number, and the visual information can focus the audio on the speaker in the speaking scene to enhance separation effect. Experimental tests are carried out on the AVSpeechs datasets, and the separation effect is evaluated by using PESQ, STOI and SDR indexes. The results show that the proposed method improves the speech separation ability by 3.73 dB compared with the traditional speech separation method.

Key words: audio-visual fusion; dilated convolution; bi-directional long short-term memory; single channel; speech separation
Foundation Item(s): Natural Science Foundation of Heilongjiang Province of China (No.LH2020F033); National Natural Science Foundation of China (No.11804068)

1 引言

语音分离的目标是从混合语音中分离出单个语音,也被称为鸡尾酒会问题[1]. 现实生活中,在复杂的开放环境中,对个体声音的倾听常常是在具有其他干扰声音存在的情况下发生的,虽然人类的听觉系统能够很容易在嘈杂环境中将注意力集中到一个感兴趣的

收稿日期:2021-06-07;修回日期:2022-01-30;责任编辑:孙瑶

个体声音上,但语音分离对计算机来说是难以实现的复杂任务[2]. 而单通道语音分离是从一维的混合语音中分离出多个语音的过程,对于多通道语音分离来说,单通道语音分离条件限制少,而且设备比较简单,所以单通道语音分离将是未来研究的热点,同时也是难点[3].

语音分离是语音信号处理中一个重要的组成部分,近年来国内外学者针对鸡尾酒会问题设计了很多计算机模型[4]. 在深度学习方法诞生之前,研究学者大多采用统计学方法解决语音分离问题,例如2006年提出的计算机场景分析(Computational Auditory Scene Analysis, CASA)[5]和非负矩阵分解(Non-negative Matrix Factorization, NMF)[6,7]是两种主流的研究方法,但是对于多说话人的语音分离来说CASA和NMF取得的效果有限. 近年来,随着深度学习领域的迅速发展,以深度神经网络(Deep Neural Network, DNN)为代表的深度模型[8]在源分离方面有显著的进展,与传统的语音分离方法相比,其性能有了很大的提高. 神经网络的典型模型是根据多个扬声器混合时频表示来估计源信号的时频掩码,这些模型将语音分离定义为有监督回归问题,对有监督的语音分离具有重要意义. 深度神经网络已经成功地应用于语音增强和语音分离领域中[9~12], Williamson等人[13]提出了一种基于DNN的语音分离方法,它在复域对幅度和相位谱同时增强,通过网络模型来估计在复域的理想比例掩模(Ideal Ratio Mask, IRM)的实虚分量,然后将这两组分量参数组成一个特征向量,代替幅值谱. 这类方法与说话人的标签序列排列有关,在训练时数据与标签不能完全匹配,最终导致输出端不能输出对应的目标. 为解决标签序列排列问题, Isik等人[14]提出通过深度聚类的方法对理想二值掩蔽(Ideal Binary Mask, IBM)进行分离来解决与说话人无关的语音分离问题. 该方法可以对语音特征进行高位可分离空间映射,然后通过聚类算法得出时频掩蔽目标,然后与混合语音计算后得出不同的分离语音信号,从而解决标签序列排列问题,在与说话人无关的语音分离问题上取得了一定的突破.

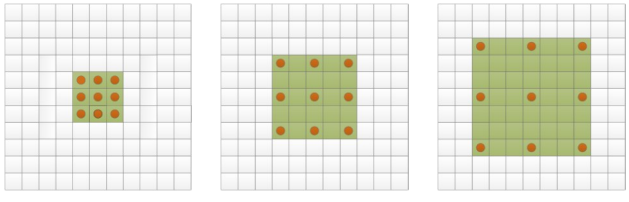
随着视觉技术的进步,视听方法也被用于语音分离,利用神经网络对听觉和视觉信号进行多模态融合,该方法引起了越来越多的研究人员的兴趣,主要包括试听语音识别[15~17]、从无声视频中预测语音或文本[18,19]、语音增强[20,21]以及语音分离[22,23]. Llagostera Casanovas等人[24]使用稀疏表示进行视听(Audio-Visual, AV)源分离,由于该方法依赖单活动区域来学习源特征,并且需要假设所有音频源都在屏幕上可以看到,而此时稀疏表示就受到了限制. 随后使用神经网络来解决该任务, Hou等人[25]提出了多任务基于卷积神经网络(Convolutional Neural Network, CNN)的模型,将音频流和视频流合并到统一的网络模型中,输出语音谱图,与纯语音分离相比具有很好的效果,有效地证实视觉信息的加入对于语音分离具有很好的效果. 另外, Ouyang等人[26]提出CNN具有计算效率高、可堆叠网络层数深的优势,能够保证高质量的训练结果.

Torfi等人[27]利用嘴唇信息进行视听匹配,由于输入嘴部特征作为辅助分离具有一定的局限性. Ephrat等人[28]用检测到的人脸特征代替嘴部特征,使语音分离效果进一步提高. 目前存在的AV语音分离方法的主要局限性是依赖说话者,这就意味着必须为每个说话者分别训练专用的模型,所以导致语音分离模型的适用性不高. Ephrat等人[28]为了解决其适用性不强的问题,故引入大规模的视听数据集——AVSpeechs. 由于现有的视听语音分离方法,大多依赖说话者本身,因此本文使用人脸特征来代替嘴部特征作为视频流的输入,解决嘴部嵌入的局限性,利用训练有素的人脸检测网络模型(Multi-Task Convolutional Neural Network, MTCNN),对输入的视频进行人脸检测,对检测到的人脸作为辅助输入对语音进行分离,并在大规模数据集AVSpeechs上训练,提出基于空洞卷积和双向长短时记忆网络的视听语音分离模型——DCNN-BiLSTM模型,并通过实验得出分离后的语音质量、可懂度和失真比,从而从客观角度评价该方法的语音分离性能.

2　分离模型

2.1　空洞卷积模型

2016年,空洞卷积在ICLR(International Conference on Learning Representation)会议上为解决图像分割问题被提出. 随着深度学习的发展,逐渐被用于语音领域. 与普通卷积不同的是,空洞卷积引入了一个超参数——扩张率(Dilation Rate),该参数定义了卷积核处理数据时各值之间的间距. 以3×3的卷积核为例,展示普通卷积和空洞卷积之间的区别,如图1所示.



(*a*) 普通卷积　(*b*) 空洞数为2的空洞卷积　(*c*) 空洞数为3的空洞卷积

图1　普通卷积和空洞卷积图

图1的3个子图是不同的卷积过程,其中大框表示输入图像,感受野为1,橙色的圆点代表3×3的卷积核,浅绿色区域表示卷积后的感受野. 图1(*a*)是普通的卷

积过程，空洞数为1，此时卷积后的感受野为3；图1(*b*)是空洞数为2的空洞卷积，卷积后的感受野为5；图1(*c*)是空洞数为3的空洞卷积，卷积后的感受野为7.

由图1可以看出，同样3×3的卷积，却可以与5×5和7×7的卷积等效，空洞卷积在不增加参数量的前提下，可以增大感受野，感受野的计算式如下：

$$k'=k+(k-1)\times(d-1) \tag{1}$$

$$S_i=\prod_{i=1}^{i}\text{Stride}_i \tag{2}$$

$$\text{RF}_{i+1}=\text{RF}_i+(k'-1)\times S_i \tag{3}$$

其中，k为空洞卷积核的大小；d为空洞数；k'为其等效的卷积核大小；Stride_i为第i层的步长；S_i为之前所有层的步长的乘积；RF_{i+1}为当前层的感受野；RF_i为上一层的感受野.

2.2 双向长短时记忆网络

长短时记忆网络于1997年被提出，是循环神经网络(Recurrent Neural Network，RNN)的变体，与RNN相比，可以学习到数据中的长期依赖关系，解决训练过程中梯度消失和梯度爆炸的问题. BiLSTM的提出是为了更好地对输入数据进行表达，可以通过使用对正向的时间序列和反向的时间序列分别进行训练，输出的数据可以获得上下文的信息.

BiLSTM网络的工作方式是在双向循环网络结构的基础上，将两个方向上的神经元用LSTM节点代替，其主要包括输入层、BiLSTM层、输出层. 一般来说，由于BiLSTM能够同时利用过去时刻和未来时刻的信息，会比单向的LSTM最终的预测更加准确. BiLSTM的结构如图2所示.

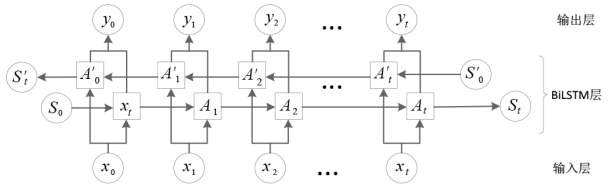


图2 BiLSTM基本结构

图2中，x_t和y_t表示t时刻的输入和输出参数，$A_0\to A_1\to\cdots\to A_t$为正向RNN，进行正向计算，$t$时刻的输入为$t$时刻的序列数据$x_t$和$t-1$时刻的输出$A_{t-1}$. $A'_t\to A'_{t-1}\to\cdots\to A'_0$为逆向RNN，进行逆向运算，$t$时刻的输入为$t$时刻的序列数据$x_t$和$t$+1时刻的输出$A'_{t+1}$. t时刻最终的输出值取决于A_{t-1}和A'_{t+1}. BiLSTM网络是由无数个记忆模块构成BiLSTM层，而每个BiLSTM层包含了控制语音信息传输的输入门、输出门和遗忘门. 其记忆模块如图3所示.

图3中，各个门函数可以表达为

输入门：

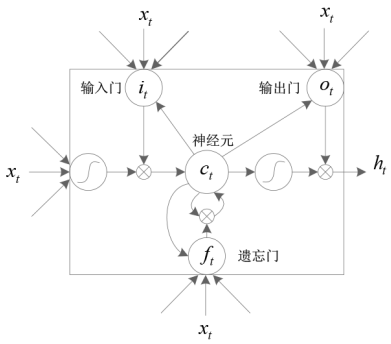


图3 长短时记忆模块

$$i_t=\sigma(\boldsymbol{W}_i\cdot[h_{t-1},x_t]+\boldsymbol{b}_i) \tag{4}$$

输出门：

$$o_t=\sigma(\boldsymbol{W}_o\cdot[h_{t-1},x_t]+\boldsymbol{b}_o) \tag{5}$$

遗忘门：

$$f_t=\sigma(\boldsymbol{W}_f\cdot[h_{t-1},x_t]+\boldsymbol{b}_f) \tag{6}$$

神经元：

$$C_t=f_t*C_{t-1}+i_t*\tanh(\boldsymbol{W}_C\cdot[h_{t-1},x_t]+\boldsymbol{b}_C) \tag{7}$$

其中，$\boldsymbol{W}_i$，$\boldsymbol{W}_o$，$\boldsymbol{W}_f$和$\boldsymbol{W}_C$分别表示输入x_t与记忆模块之间相互连接的矩阵；$\boldsymbol{b}_i$，$\boldsymbol{b}_o$，$\boldsymbol{b}_f$和$\boldsymbol{b}_C$表示为偏置向量；$\sigma(\cdot)$表示Sigmoid函数；$\tanh(\cdot)$表示双曲线的正切函数；h_t为t时刻隐含层的输出.

2.3 DCNN-BiLSTM

基于空洞卷积的增大感受野的特性和双向长短时记忆网络对数据正向和反向处理，以获得数据中依赖关系的能力，本文建立DCNN-BiLSTM视听语音分离模型. 此模型是由视频中检测到的人脸信息和与之相对应的音频信息组成的多模态结构，而后经过神经网络分离获取不同说话人特征. 输入视频信息是由视频流将视频中每一帧检测到的人脸缩略图与视频的音轨作为音频流的输入来构成的. 视频流在处理过程中使用基于多任务卷积神经网络MTCNN的人脸识别模型提取每个说话人缩略图的人脸嵌入，然后通过DCNN学习视觉特征. 音频流在处理过程中先对输入信号进行STFT变换获得频谱图，而后使用与视频流处理相似的DCNN学习音频特征，并通过连接学习到的视觉和音频特征创建联合的视听特征，最后利用BiLSTM和3个全连接层(Fully Connected layer，FC)对视听融合特征处理，最终输出视频中被检测到的与人脸相对应的复杂语谱图掩蔽，再使用混合语音乘以掩蔽以孤立每个说话人的语音信号，同时抑制其他干扰信号，以达到分离语音的效果. 语音分离处理过程的总体结构如图4所示.

图4中，采用的视觉和听觉特征作为输入，对于视频输入，给定一个包含多个扬声器的视频剪辑，使用现有的面部检测器查找每帧中的人脸，本文建立的此模型采用的是以25帧/秒的速度播放3 s的剪辑，每个说

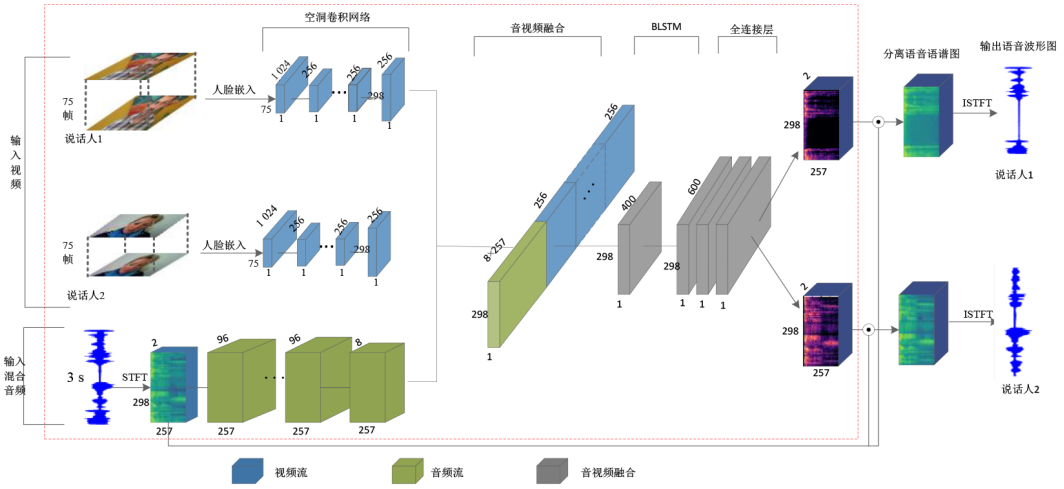


图4　DCNN-BiLSTM视听语音分离总框图

话人总共有75个面部缩略图. 该模型使用MTCNN人脸识别模型对每个检测到的人脸缩略图提取其人脸嵌入. 对于音频特征,对3 s的音频进行STFT,每个时频单元包含一个复数的实部和虚部,将此作为音频的输入. 另外,对音频采取μ律压缩操作,可以防止响亮的音频压倒弱的音频.

DCNN-BiLSTM模型的输出采用乘法频谱图掩蔽,其描述了干净的语音与背景干扰的时频关系,本文采用乘法频谱图掩蔽来抑制其他干扰信号,孤立说话人语音,有助于提升语音的分离效果.

2.4　分离性能评价

常用于评估分离效果的指标有3种:客观语音质量评估(Perceptual Evaluation of Speech Quality,PESQ)指标衡量语音的感知能力;用短时客观可懂度(Short-Time Objective Intelligibility,STOI)指标衡量分离语音的可懂度;源失真比(Signal-to-Distortion Ratio,SDR)指标衡量语音的分离能力. 其中,PESQ指标值在–0.5~4.5之间,得分越高表明被测试的语音具有越好的听觉语音质量;STOI指标在0~1之间,得分越高代表语音的可懂度越好. 分离信号的SDR定义为

$$\mathrm{SDR}=10\log_{10}\left(\frac{\left\|S_{\mathrm{target}}\right\|^2}{\left\|e_{\mathrm{interf}}+e_{\mathrm{noise}}+e_{\mathrm{artif}}\right\|^2}\right) \tag{8}$$

其中,S_{target}为分离出来的语音信号;e_{interf}为干扰信号;e_{noise}为加性噪声;e_{artif}为算法构件所产生的干扰信号.

本文利用上述3种评价方法,对提出的DCNN-BiLSTM语音分离模型性能进行评估.

3　实验与结果分析

3.1　实验数据及参数设置

3.1.1　实验数据

AVSpeechs数据集中语音长度在3~10 s之间,在每个片段中,视频中唯一可见的面孔和原声带中唯一可以听到的声音属于一个说话人. 该数据集包含了约4 700 h的视频片段,大约有15万个不同的说话者,跨越了各种各样不同性别的人、语音和面部姿态.

干净的语音剪辑来自AVSpeechs数据集,从数据集中不同长度的片段中截取3 s不重叠的语音片段,对于视频剪辑也是来自AVSpeechs数据集,同样截取与音频时间相对应的时长为3 s的视频段,本次实验随机选取500个干净语音两两混合后生成混合的语音数据库,再从此混合语音中选取20 000个可懂度相当的混合语音作为本次实验的数据集,其中90%作为训练集,剩余的10%作为测试集. 本文利用的混合语音按如下方式生成,即

$$\mathrm{Mix}=\mathrm{AVS}_j+\mathrm{AVS}_k \tag{9}$$

其中,AVS_j和AVS_k是来自AVSpeechs数据集的不同源视频的干净语音;Mix为生成的混合音频.

3.1.2　参数设置

DCNN-BiLSTM模型的音频流和视频流均由空洞卷积网络处理,在对视频流进行空洞卷积时不是在1 024维的人脸嵌入通道上进行,而是在时间轴上进行的. 在处理音频流时,所有音频被重新采样16 kHz,立体声音均被转化为左声道. STFT采用帧长25 ms,窗函数为汉宁窗,帧移为10 ms,FFT的大小为512,混合音频时长为3 s,采样频率为16 kHz,经计算输入音频的特征为257和298,本文对音频的数据处理参数设置如表1所示.

在处理视频流时,通过移除或者复制重新采样所有视频中人脸,以每秒25帧的采样率对3 s的视频进行采样,即可产生75个人脸嵌入作为输入视频流,另外,当特定的样本中遇到缺失的帧时,使用零向量代替人脸的嵌入,本文对视频的数据处理参数如表2所示.

3.2　分离结果与分析

采用本文提出的DCNN-BiLSTM语音分离模型对

表1 DCNN-BiLSTM模型的音频流空洞卷积层参数

Layer	Num Filter	Filter size	Dilation	Context
1	96	1×7	1×1	1×7
2	96	7×1	1×1	7×7
3	96	5×5	1×1	9×9
4	96	5×5	2×1	13×11
5	96	5×5	4×1	21×13
6	96	5×5	8×1	37×15
7	96	5×5	16×1	69×17
8	96	5×5	32×1	133×19
9	96	5×5	64×1	261×21
10	96	5×5	1×1	263×23
11	96	5×5	2×2	267×27
12	96	5×5	4×4	275×35
13	96	5×5	8×8	291×51
14	96	5×5	16×16	323×83
15	96	5×5	32×32	387×147
16	96	5×5	64×64	515×275
17	96	1×1	1×1	515×275

表2 DCNNBiLSTM模型视频流的空洞卷积层参数

Layer	Num Filter	Filter Size	Dilation	Context
1	256	7×1	1×1	7×1
2	256	5×1	1×1	9×1
3	256	5×1	2×1	13×1
4	256	5×1	4×1	21×1
5	256	5×1	8×1	37×1
6	256	5×1	16×1	69×1

两组干净语音信号合成的混合信号进行分离，并输出语谱图. 为了给出直观的观测结果，本文以其中输出的一组语音分离语谱图为例，2组干净语音混合后的语音语谱图、2组干净语音的语谱图及利用DCNN-BiLSTM语音分离模型对混合语音分离后得到的两组分离语音语谱图如图5所示.

图5(a)为2组干净语音混合后的语谱图，分别对比图5(b)中的语谱图(左)和图5(c)中的语谱图(左)，以及图5(b)中的语谱图(右)和图5(c)中的语谱图(右). 由此可见，该模型能够将源信号有效的分离出来.

语谱图仅仅通过定性观察，粗略评估是否能进行语音分离的手段，但并不能定量评估分离模型的分离效果. 因此以下利用SDR评价指标，在测试集中随机取3组声音数据：男生和男生、男生和女生、女生和女生，基于DCNN-BiLSTM语音分离模型对不同特性混合语音的语音分离效果进行评价，并与不同的模型进行分离效果对比，结果如表3所示. 其中，M1，M2，M3表示3位男士的声音；F1，F2，F3表示3位女士的声音. 在表3中，对于混合语音M1+M2来说，前者表示M1，后者表示M2，其他的混合语音亦是如此.

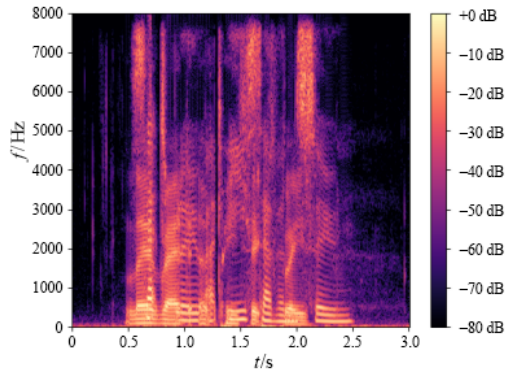
由表3可知，DCNN-BiLSTM语音分离模型对混合语音M1+M2进行分离，分离出M1语音的SDR值为15.1 dB，M2语音的SDR值为14.7 dB，相比于AV模型分离出M1语音的SDR值提升了6.2 dB，M2语音的SDR值提升了5.4 dB. 由表3中其他数据可以得出，DCNN-BiLSTM模型的语音分离效果有明显的提升. 另外，对比同性别说话人和异性说话人的语音分离效果可以看出，AV模型对男生和男生混合语音的分离平均SDR值为9.1 dB，而DCNN-BiLSTM对其语音的分离SDR值达到15.4 dB. 由表3中其他数据可知，无论是同性混合语音的分离还是异性混合语音的分离，DCNN-BiLSTM模型的分离效果均优于AV模型.

为对比分析AV模型和本文所用的DCNN-BiLSTM模型在2个语音混合情况下语音分离效果，并分析本文所提出模型层数不同时的语音分离性能，取测试集数据利用PESQ值、STOI值和SDR值进行评价，分离后语音的评价结果如表4所示.

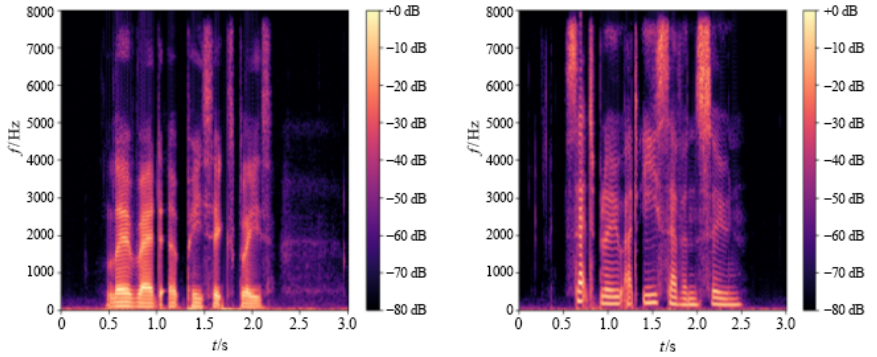
由表4可知，DCNN-BiLSTM-15模型分离出语音1的SDR值为14.15 dB，语音2的SDR值为15.88 dB，相比于AV模型分离出语音1的SDR值提升了1.55 dB，语音2的SDR值提升了4.08 dB，对比表4中的PESQ值和STOI值，说明DCNN-BiLSTM-15模型对语音的分离效果优于AV模型.

采用与空洞卷积感受野相同的普通卷积神经的CNN-BiLSTM模型分离出语音1的SDR值为12.73 dB，语音2的SDR值为12.02 dB，比DCNN-BiLSTM-15模型分离出的语音1的SDR值降低1.42 dB，语音2的SDR值降低了3.86 dB，对比表4中的PESQ值和STOI值，说明使用空洞卷积神经不仅能减少网络的复杂度，而且可以分离出质量更好的语音. DCNN-BiLSTM(去除视频分支)模型分离出语音1的SDR值为12.35 dB，语音2的SDR值为11.45 dB，相比于DCNN-BiLSTM-15模型分离出语音1的SDR值降低了1.80 dB，语音2的SDR值降低了4.43 dB，对比表4中的PESQ值和STOI值，说明视频分支的加入对语音分离有很好辅助分离的作用.

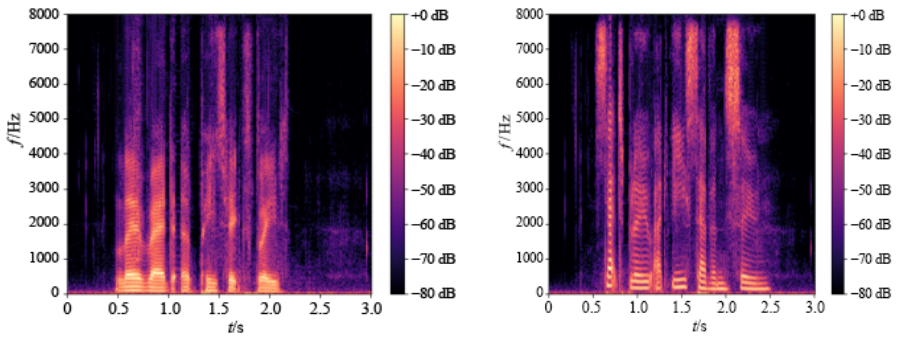
另外，DCNN-BiLSTM-17模型分离出语音1的SDR值为15.75 dB，语音2的SDR值为16.11 dB，而相比于DCNN-BiLSTM-15模型分离出的语音1和语音2的SDR值分别提升了1.6 dB和0.23 dB，这表明增加网络层数会得到更好的分离效果. AV模型分离出的语音1的SDR值为12.6 dB，语音2的SDR值为11.8 dB，二者的平均SDR值为12.2 dB，DCNN-BiLSTM-17模型分离出的语音1和语音2的平均SDR值为15.93 dB，比AV模型分离出语音的平均SDR值高3.73 dB.



(*a*)　混合语音语谱图



(*b*)　干净语音语谱图



(*c*)　分离得到的语音语谱图

图5　频谱图

表3　不同性别语音分离的SDR值

混合语音	AV模型[28]			DCNN-BiLSTM模型		
	前者	后者	平均值	前者	后者	平均值
M1+M2	8.9	9.3	9.1	15.1	14.7	15.4
M1+M3	8.7	9.4		16.0	15.8	
F1+F2	9.5	10.5	10.0	11.3	16.1	14.5
F1+F3	10.3	9.8		16.2	14.4	
M1+F1	11.2	10.3	10.5	17.1	14.0	16.6
M1+F2	9.8	10.5		16.7	18.4	

表4 语音分离总体评价

评价方法	语音1			语音2		
	PESQ	STOI	SDR	PESQ	STOI	SDR
AV模型[28]	2.50	0.71	12.60	2.26	0.66	11.80
CNN-BiLSTM	2.65	0.75	12.73	2.54	0.69	12.02
DCNN-BiLSTM(去除视频分支)	2.40	0.68	12.35	2.11	0.56	11.45
DCNN-BiLSTM-15	3.24	0.80	14.15	3.04	0.86	15.88
DCNN-BiLSTM-17	3.35	0.87	15.75	3.30	0.90	16.11

4 总结

本文针对单通道语音分离,提出一种基于DC神经网络和BiLSTM神经网络的视听融合语音分离模型——DCNN-BiLSTM模型. 采用DCNN神经网络对音频和嵌入人脸特征进行数据处理,由于DCNN神经网络能够在参数个数保持不变的情况下增大卷积核的感受野,因此通过DCNN神经网络能够获得更多的音频和视觉特征,得到更为全面的视听融合特征. 采用BiLSTM神经网络对视听融合特征进行训练,能够学习数据中长期依赖的关系,通过对正反向的视听融合数据进行训练,能够获得更多依赖关系. 实验结果表明,本文提出的DCNN-BiLSTM模型分离出的语音在PESQ,STOI和SDR这3个指标上都要好于AV模型,且改变网络层数能够进一步提高语音分离效果.

参考文献

[1] BELL A J, SEJNOWSKI T J. An information-maximization approach to blind separation and blind deconvolution [J]. Neural Computation, 1995, 7(6): 1129-1159.

[2] 葛宛营, 张天骐, 范聪聪, 等. 噪声情况下采用稀疏非负矩阵分解与深度吸引子网络的人声分离算法[J]. 声学学报, 2021, 46(1): 55-66.
GE W Y, ZHANG T Q, FAN C C, et al. Monaural noisy speech separation combining sparse non-negative matrix factorization and deep attractor network[J]. Acta Acustica, 2021, 46(1): 55-66. (in Chinese)

[3] 朱阁. 基于深度学习的单通道语音分离技术研究[D]. 南京: 南京邮电大学, 2020.
ZHU G. Research on Single-Channel Speech Separation Technology Based on Deep Learning[D]. Nanjing: Nanjing University of Posts and Telecommunications, 2020. (in Chinese)

[4] CHEN J J, MAO Q R, QIN Y C, et al. Latent source-specific generative factor learning for monaural speech separation using weighted-factor autoencoder[J]. Frontiers of Information Technology & Electronic Engineering, 2020, 21 (11): 1639-1650.

[5] WANG D L, BROWN G S. Computational Auditory Scene Analysis: Principles, Algorithms, and Applications [M]. Hoboken: Wiley Interscience, 2006.

[6] SCHMIDT M N, OLSSON R K. Single-channel speech separation using sparse non-negative matrix factorization [C]//Proceedings of Interspeech 2006 - Ninth International Conference on Spoken Language Processing. Pittsburgh: ISCA, 2006: 1652-1661.

[7] ZHOU W L, ZHU Z, LIANG P Y. Speech denoising using Bayesian NMF with online base update[J]. Multimedia Tools and Applications, 2019, 78(11): 15647-15664.

[8] SUN L, DU J, DAI L R, et al. Multiple-target deep learning for LSTM-RNN based speech enhancement[C]//2017 Hands-free Speech Communications and Microphone Arrays (HSCMA). San Francisco: IEEE, 2017: 136-140.

[9] KOLBÆK M, TAN Z H, JENSEN J. Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2017, 25(1): 153-167.

[10] SALEEM N, KHATTAK M I, QAZI A B. Supervised speech enhancement based on deep neural network[J]. Journal of Intelligent & Fuzzy Systems, 2019, 37(4): 5187-5201.

[11] SALEEM N, KHATTAK M I, ALI M, et al. Deep neural network for supervised single-channel speech enhancement[J]. Archives of Acoustics, 2019, 44: 3-12.

[12] ZÃO L, COELHO R, FLANDRIN P. Speech enhancement with EMD and Hurst-based mode selection[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014, 22(5): 899-911.

[13] WILLIAMSON D S, WANG Y X, WANG D L. Complex ratio masking for monaural speech separation[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2016, 24(3): 483-492.

[14] ISIK Y, ROUX J L, CHEN Z, et al. Single-channel multi-speaker separation using deep clustering[EB/OL]. (2016) [2021]. https://arxiv.org/abs/1607.02173.

[15] FENG W J, GUAN N Y, LI Y, et al. Audio visual speech

recognition with multimodal recurrent neural networks [C]//2017 International Joint Conference on Neural Networks (IJCNN). Anchorage: IEEE, 2017: 681-688.

[16] MROUEH Y, MARCHERET E, GOEL V. Deep multimodal learning for audio-visual speech recognition[C]// 2015 IEEE International Conference on Acoustics, Speech and Signal Processing. South Brisbane: IEEE, 2015: 2130-2134.

[17] NGIAM J, KHOSLA A, KIM M, et al. Multimodal deep learning[C]//Proceedings of the 28th International Conference on Machine Learning. Washington: DBLP, 2009: 1-9.

[18] 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. Honolulu: IEEE, 2017: 3444-3453.

[19] EPHRAT A, HALPERIN T, PELEG S. Improved speech reconstruction from silent video[C]//2017 IEEE International Conference on Computer Vision Workshops. Venice: IEEE, 2017: 455-462.

[20] 颜霖煌. 基于图像边缘保持滤波技术的语音增强算法研究[D]. 广州: 广州大学, 2020.
YAN L H. Research on Speech Enhancement Algorithm Based on Image Edge Preserving Filter[D]. Guangzhou: Guangzhou University, 2020. (in Chinese)

[21] GABBAY A, EPHRAT A, HALPERIN T, et al. Seeing through noise: Visually driven speaker separation and enhancement[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing. Calgary: IEEE, 2018: 3051-3055.

[22] TAN K, XU Y, ZHANG S X, et al. Audio-visual speech separation and dereverberation with a two-stage multimodal network[J]. IEEE Journal of Selected Topics in Signal Processing, 2020, 14(3): 542-553.

[23] GU R Z, ZHANG S X, XU Y, et al. Multi-modal multi-channel target speech separation[J]. IEEE Journal of Selected Topics in Signal Processing, 2020, 14(3): 530-541.

[24] LLAGOSTERA CASANOVAS A, MONACI G, VANDERGHEYNST P, et al. Blind audiovisual source separation based on sparse redundant representations[J]. IEEE Transactions on Multimedia, 2010, 12(5): 358-371.

[25] 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.

[26] OUYANG Z H, YU H J, ZHU W P, et al. A fully convolutional neural network for complex spectrogram processing in speech enhancement[C]//2019 IEEE International Conference on Acoustics, Speech and Signal Processing. Brighton: IEEE, 2019: 5756-5760.

[27] TORFI A, IRANMANESH S M, NASRABADI N, et al. 3D convolutional neural networks for cross audio-visual matching recognition[J]. IEEE Access, 2017, 5: 22081-22091.

[28] EPHRAT A, MOSSERI I, LANG O, et al. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation[J]. ACM Transactions on Graphics, 2018, 37(4): 112.

作者简介

兰朝凤　女,1981年出生,黑龙江绥化人. 博士后,副教授,博士生导师. 主要研究方向为人工智能AI算法、音频信号分析与处理、噪声控制、机器视觉、生物医学信号处理与建模.
E-mail: lanchaofeng@hrbust.edu.cn

王顺博　女,1995年出生,河南周口人. 硕士研究生. 主要研究方向为音频信号分析与处理.
E-mail: 1530984071@qq.com