基于深度视听模型的鸡尾酒会问题研究现状与展望

卢慧君,蔡敦波,黄智国,杭 涛,冯清松,钱 岭
(中移(苏州)软件技术有限公司,江苏 苏州 215000)

摘 要:“鸡尾酒会问题”目前依然是语音处理领域很有挑战的一个问题,该问题的核心是多说话人语音分离。目前对于以上问题的研究取得了较大的进展,但缺少一个系统,简洁的分析和总结。文章围绕“鸡尾酒会问题”的解决方案,总结了语音处理领域多说话人语音分离方法的发展:(1)分析了经典的语音分离方法,包括谱减法、维纳滤波、计算听觉场景分析等;(2)分析了引入深度学习思想后出现的语音分离方法,包括初期的深度音频的方法和其后出现的深度视觉听觉的方法,重点评述了基于深度学习的视觉听觉方法的主要算法思想和效果方面的新进展;(3)总结了目前深度视听方法中常用视听数据集的特点。文末对深度视听模型解决鸡尾酒会问题的现状以及当前存在的挑战进行了评述,并展望未来的研究方向。

关键词:鸡尾酒会问题;多说话人语音分离;深度学习;深度视听方法;视听数据集
中图分类号:TP183 **文献标识码**:A **文章编号**:1673-629X(2021)0008-08

Research State and Frontiers of Cocktail Party Problem Based on Deep Audio-visual Models

LU Hui-jun,CAI Dun-bo,HUANG Zhi-guo,HANG Tao,FENG Qing-song,QIAN Ling
(China Mobile (Suzhou) Software Technology Co.,Ltd,Suzhou 215000,China)

Abstract:The "cocktail party problem" is still a very challenging problem in the field of speech processing. The core of the problem is the separation of multi-speaker speech. At present,the research on the above issues has made great progress,but it lacks a systematic, concise analysis and summary. The solutions of the "cocktail party problem" are focused on and the development of multi-speaker speech separation methods in the field of speech processing is summarized. Firstly,the classic speech separation methods are analyzed briefly,including spectral subtraction,Wiener filtering,and computational auditory scene analysis. Secondly,the deep learning based speech separation methods are analyzed in-depth,including the auditory methods and deep audio-visual methods,and particularly reviews the new development of deep audio-visual models. Thirdly,the commonly used audio-visual datasets are reviewed. At the end,deep audio-visual models to solve the cocktail party problem and current challenges are reviewed,and the future directions of research are discussed.

Key words:cocktail party problem;multi-speaker speech separation;deep learning;deep visual-audio model;visual-audio datasets

0 引 言

人脑可将听觉注意力集中在特定刺激上,同时滤除其他刺激。此种注意力选择机制,使人即使身处嘈杂的酒会,依然可以集中精力于自己关心和参与的对话之中。著名心理学家 Colin Cherry 于 1953 发表论文[1]将这种现象命名为“鸡尾酒会效应”。此后几十年间,研究者一直致力于探究和建模人类的此种能力,提出了很多在不同场景下实现“鸡尾酒会效应”智能的人工智能方法。其中,针对多个说话人同时讲话的自然场景,能够识别、分离和跟踪目标说话人的语音,是其中的一系列关键问题。相应的技术推动了工业界自动语音识别(automatic speech recognition,ASR)系统的快速发展。在 ASR 系统中,由于麦克风采集到的声音中可能包括噪声、其他人说话的声音、混响等干扰,直接进行语音识别将会影响到识别准确率。现代语音识别系统在语音识别的前端增加了语音分离模块,把目标说话人声音和其它干扰分开,以提高语音识别系统鲁棒性。

语音分离问题在很多业务场景中均处于核心的地位,文中将把说明重点放到语音分离技术方面。更进一步地,文中会更多关注近几年发展起来的视听特征相结合的多说话人语音分离解决方案,对这类的模型

收稿日期:2020-07-23
基金项目:中国移动通信集团有限公司应用基础研究项目(R202111101114MP)
作者简介:卢慧君(1991-),女,硕士,研究方向为人工智能与图像处理。

进行介绍和阐述。对视听模型的特殊关注一方面起源于人类听觉机制的研究——听觉过程伴随着视觉机制。基于视觉和听觉信息结合的多感知特征表示的研究,代表了目前很活跃的前沿领域;另一方面,基于深度学习的视听模型方法,在解决鸡尾酒问题上取得了很大的进展。由于互联网可以提供大量的单通道视频,视觉听觉数据集得以不断地发展和更新,为深度学习模型的训练提供了很好的保证。鸡尾酒问题的解决方案,可以直接应用到例如会议字幕,多方人机交互以及听力辅助等场景,具备很好的工业前景。

1　鸡尾酒会问题

在一个鸡尾酒会上,混杂着多个说话人的语音,音乐和乐器等声音。在如此嘈杂环境中,人类可以有选择地注意特定的声音,而忽略其他的声音,如图 1 所示。

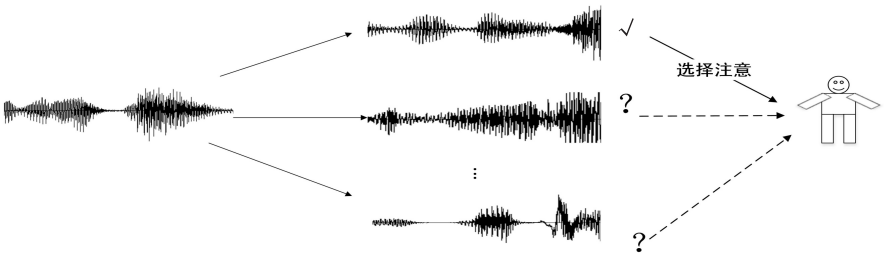

图 1　人类在鸡尾酒会中的听觉选择机制示意图

计算机模仿人类这种能力,需要面对两大挑战:其一是多说话人语音分离问题或者更广泛地是,多种混杂声音的分离;其二是选择、跟踪特定的说话人或声源,并实现在分离出的不同声音之间切换的能力。解决“鸡尾酒会问题”,主要是把多种混合的声音分离开来,并可以选择注意其中的一种或两种声音信号。研究人员试图用算法解决这样的问题,让计算机在嘈杂环境下,具备像人类一样的听觉注意能力。但目前来看,达到这样的目标依然困难重重。

2　语音分离的传统方法

语音分离旨在从多个说话人的音频混合中分离出单个说话人的语音,一般有两类描述:其一,根据分离语音中的噪声干扰信号的不同,可以分为三类:一类是语音增强,干扰信号为环境噪声;一类是解混响问题,干扰信号为说话人的空间反射信号;第三类即多说话人语音分离问题,干扰信号为其他说话人的语音,也是解决鸡尾酒会问题时面临的核心问题。其二,根据声源采集端的数目,又可以分为单通道语音分离和多通道语音分离。单通道语音是指使用单个麦克风采集的声音。分离方法主要是利用语音的频谱属性——音调的连续性、谐波结构、常见的声母等,建立基于统计的模型、基于规则的模型或者基于分解的模型。传统的单通道语音分离方法包括谱减法(spectral subtraction,SS)、维纳滤波(Wiener filtering,WF)、计算听觉场景分析(computational auditory scene analysis,CASA)等。多通道语音则是麦克风阵列采集的声音,分离方法可以利用声源的空间属性,相比单通道语音分离难度更小一些。常规的多通道语音分离技术,例如独立成分分析(independent component analysis,ICA)、波束成形(beamforming,BF)等。

对这些经典的语音分离技术的简单介绍如下:

谱减法:假设噪声是平稳或缓慢变化的信号,谱减法是一种通过从含噪声信号频谱中减去平均噪声频谱得到估计值,来恢复在加性噪声中观察到的信号的功率或幅度频谱的方法。谱减法的优点是计算复杂度低,缺点是由于噪声的随机变化,频谱相减会导致短时幅度或功率谱的计算值出现负数。

维纳滤波:维纳滤波器是诺伯特·维纳(Norbert Wiener)在 1940 年提出的。该方法使用统计模型(贝叶斯方法)估算信号,解决了平稳信号的信号估计问题。信号可能包含已被加性噪声破坏的未知感兴趣信号,维纳滤波器可从已知被损坏的信号中滤除噪声,提供感兴趣的基础信号的估计。

计算听觉场景分析法:计算听觉场景分析(CASA)就是让计算机模拟人类处理听觉信号的过程,具备从复杂混合声源中感知声音、处理声音、解释声音的能力。CASA 可分为两大类别:一类是图式驱动型 CASA;另一类是数据驱动型 CASA。

独立成分分析法:独立成分分析(ICA)是一种基于统计计算,用于揭示随机变量、测量值、信号集的隐藏因子的技术。在满足一定条件的情况下,ICA 能够从多路观测信号中,较好地分离出这些独立的成分(源或因子)。ICA 是一种比主成分分析和因子分析更强大的技术。

波束成形法:又称为空间滤波器。该方法通过设计麦克风阵列的间隔,尺寸,结构,从而达到增强特定方向的信号,削减其他方向的干扰信号的目的。但是,

当目标源和干扰源位置靠近时或在回声场景中,波束成形对声源方向的判定变得模糊,效果变差。

总的来说,以上这些传统的语音分离方法,只能应用于一些相当简单的情况,例如固定扬声器、有限词汇表、不同性别说话人的语音分离场景。随着深度学习理论和方法的发展,基于深度学习的方法开始应用到语音处理领域。此类方法的出现,给语音处理领域带来了崭新的活力和较大的发展。

3 语音分离深度模型的初期发展

基于深度学习的语音分离方法,是指从训练数据中学习语音、说话人和噪音的特征,从而实现分离语音的目标。根据有没有引入视觉信息,可以把目前基于深度学习的语音分离方法分为两类:一类是基于纯语音特征的方法,另一类是结合视觉听觉特征的方法,简称视听模型方法。这两大类方法均在不断发展,都是很活跃的研究方向。本章节简单介绍基于深度学习的语音分离方法的早期发展。

早期基于深度学习的模型结构一般较为简单,根据训练目标的不同,可以分为三类:基于时频掩蔽的方法、基于频谱映射的方法和基于信号近似的方法。基于时频掩蔽的算法得到目标语音,是传统的计算听觉场景分析(CASA)常见的思路。最早被利用的掩蔽信号为理想二值掩蔽(ideal ratio mask,IBM),其定义如下:

$$\mathrm{IBM}(t,f)=\begin{cases}1 & \text{若 SNR}(t,f) > \mathrm{LC}\\ 0 & \text{其他}\end{cases} \tag{1}$$

其中,

$$\mathrm{SNR}(t,f)=10\log_{10}\frac{\|S(t,f)\|^2}{\|N(t,f)\|^2} \tag{2}$$

LC 是一个阈值,一般设置为零。理想二值掩蔽里面的非零数值标注出了目标语音中主导的时频单元(time-frequency unit)。

华人科学家汪德亮团队首次使用深度神经网络学习带噪声信号的特征和时频掩蔽之间的映射关系,把鸡尾酒会问题变成了一个机器学习问题。他们把一个听觉信号在时域和频域两个维度进行表示,得到一个二维矩阵,矩阵中的每个元素称为一个时频单元。基于理想二值掩蔽(IBM),该时频单元要么属于目标声源,要么就是背景噪声。这样,问题就转化成一个分类问题,使得模型学习的难度变小。

随后的一些研究对时频掩蔽的计算目标进行了改进,比如理想浮值掩蔽(ideal ratio mask,IRM)、幅度谱掩蔽(spectral magnitude mask,SMM)、相位敏感掩蔽(phase sensitive mask,PSM)、复数理想浮值掩蔽(complex ideal ratio mask,cIRM)等。其中,cIRM 是复数域的掩蔽,通过估计频谱的实部和虚部,间接估计了幅度和相位,可以完美地从含噪语音中重构出纯净的语音。目前时域和频域上的语音分离方法大多都是基于掩蔽的思想。

基于频谱映射的方法,是通过有监督学习,让网络模型学习含噪语音的频谱与干净语音的频谱之间的映射关系,把问题转化成一个回归问题。频谱映射可以选用幅度谱、功率谱、对数功率谱、梅尔谱和 Gammatone 功率谱——模拟人耳耳蜗滤波后的特征。

基于信号近似的方法主要有幅度信号近似(magnitude signal approximation,MSA)、相位敏感信号近似(phase-sensitive signal approximation,PSA),复数谱近似(complex spectrum approximation,CSA)。信号近似法预测一个隐式的掩蔽,计算谱间误差,最大化信噪比。

除了训练目标的各种变化,基于深度学习的语音分离方法采用的网络结构也存在较大差异。常用的网络结构有深度神经网络(deep neural network,DNN)、双向的长短记忆(bidirectional long short-term memory,BLSTM)循环神经网络、残差网络(residual network,ResNet)、U-Net 等。为了更好地理解各种模型的差别,下面重点介绍一下高频使用的网络结构的定义和特点:

DNN:深度神经网络(DNN)[2]是深度学习的基础,可以理解为有很多隐藏层的神经网络。DNN 有时也叫做多层感知机(multi-layer perceptron,MLP)。DNN 内部的神经网络层可以分为三类,即输入层、隐藏层和输出层。一般来说第一层是输入层,最后一层是输出层,而中间的层数都是隐藏层。

BLSTM:长短时记忆网络(long short-term memory,LSTM)是 1997 年出现的循环神经网络(recurrent neural network,RNN)的一种变体。2005 年,Graves 提出了将 LSTM 与 BRNN 结合到一起,即是 BLSTM。BLSTM 对正向的时间序列和反向的时间序列分别进行训练,可以获得上下文相关的长时信息。与 BRNN 相比,BLSTM 可以更好地处理梯度消失和爆炸的问题[3]。

ResNet:深度残差网络(ResNet)[4]是过去几年中计算机视觉和深度学习领域开创性的工作。ResNet 不仅使训练数百甚至数千层网络成为可能,而且具有很优越的性能。F(X)+X 组成的结构称为残差块(residual block),如图 2 所示。多个相似的残差块串联构成 ResNet。

U-Net:U-Net,顾名思义,是个 U 型的网络[5],U 型左侧可视为一个编码器,右侧可视为一个解码器。编码器有四个子模块,每个子模块包含两个卷积层,之

后连接一个最大池化层,即下采样层。采用的编码器(下采样)-解码器(上采样)结构和跳跃连接是一种非常经典的设计方法。目前已有许多新的卷积神经网络设计方式,但很多仍延续了 U-Net 的核心思想(见文献[12])。

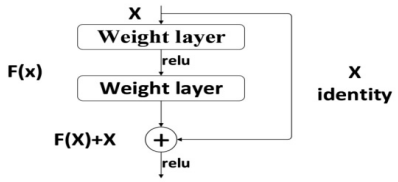

图 2　残差块

研究者把这些网络结构应用到语音处理的任务中来,为语音处理带来了新的活力。但如何学习语音流的高效的特征表示依然存在诸多挑战。由于语音的视觉信号(例如,嘴唇运动),可以用来学习更好的特征表示,结合视觉信息的语音分离解决方案应运而生。视觉听觉联合特征表示在解决鸡尾酒会问题方面取得了很大的进步。

4　深度视听方法重要进展

引入视觉信息对语音进行处理之后,基于深度学习的视听语音处理方法取得了重大进展。该方法广泛应用于以下几类任务[6-8]:视频或图像与音频之间的相互生成问题[9],比如语音预测;定位发声源问题[10],包括视听事件定位,发声对象定位和乐器的定位,已经有研究人员针对乐器发声实现了像素级别的定位[11];音频和视频的匹配问题[12];多说话人的语音分离问题等[13-23]。本章节主要讨论多说话人分离这一应用方向的技术发展。结合了视觉特征之后,语音分离模型在有噪声的环境中,分离效果均得到了很大提升。

基于深度学习的视听模型方法一般包含以下几个模块:特征提取,时频分解,分离模型,以及波形合成。模型的结构,一般是分为视觉处理分支和音频处理分支,分别完成视觉信息、语音信息的特征编码和表示。视觉特征提取基于三维卷积神经网络(3D CNN)[12,14],2D 或 3D ResNet[15,17,21],简单的一维空洞卷积[13,23],或者直接最基本的卷积神经网络和深度神经网络[18-20];一些模型基于编码和解码网络结构[16,21]。音频特征提取一般是转换到频域进行。常见的处理方法是短时傅里叶变换(short-time Fourier transform,STFT)和短时傅里叶逆变换(inverse short-time Fourier transform,ISTFT)[12,13,15,18-20]。近来,一些研究人员在时域对语音信号进行处理[14,21-23],也取得了较好的实验结果。分离模型一般由深度神经网络结合全连接网络层组成,由于语音信号有上下文信息,BLSTM 的网络结构处理语音就更加合适,目前也出现了基于 BLSTM 的研究工作[13,19-20];还有部分工作结合 U-Net 网络结构,建立回归模型[14],实现 2 个人的混合语音分离,分离语音聆听效果清晰,失真度小。

在以上提及的研究工作中,重点介绍一下 Torfi[12],Ephrat[13],Afouras[15],Wu[21]的方法。

Torfi[12]等人基于三维卷积神经网络(3D CNN)结构,寻找音频和视频流之间的对应关系。这里称该模型为 3D-CNN-FC。该模型首先将视听两种模态映射到表示空间中,然后使用学习到的多模态特征评估视听流的对应性。模型使用两个深度神经网络分支,处理视觉信息和语音信息,两个子网络通过全连接神经网络层实现融合,如图 3 所示。视觉分支网络的输入是从视频中提取的 60 * 100 像素大小的嘴唇图像;语音分支网络的输入为语音信号的梅尔频率能量系数(Mel-frequency energy coefficients,MFEC)及其一阶导数、二阶导数。

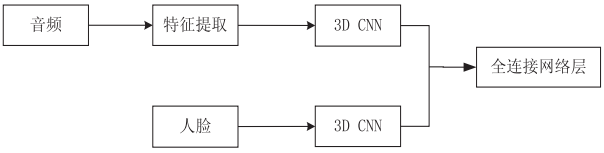

图 3　3D-CNN-FC 算法模型框架

模型基于视听数据集 LRW,AVD 进行训练,均等错误率(equal error rate,EER)指标提升了 20%,平均精度(average precision,AP)指标提升了 7%[12]。该方法可直接应用于多说话人场景中的说话人识别,也可视为解决多说话人语音分离问题的一种视听模型方法。

2018 年,谷歌研究团队的 Ephrat[13]等人构建了一个视听联合模型,这里称该模型为 AV-BLSTM-FC。该模型使用空洞卷积进行视觉特征提取,如图 4 所示。算法基于复数浮值掩蔽(complex ratio mask,cRM),利用网络直接预测一个复数掩蔽(complex mask),该复数掩蔽与原始的含噪语音频谱图进行复数乘积运算,再经过短时傅里叶逆变换,得到时域的语音信号。损失函数定义为干净语音频谱和网络分离出的频谱之间的均方误差。作者构建了一个新的视听数据集——AVSpeech 作为训练集,该数据集包含来自网络的数千小时的视频片段,包括嘈杂的真实场景,比如激烈采访,吵闹的酒吧,玩闹的孩子们等。基于 Mandarin 数据集测试,比 Hou[18]等人提出的模型语音感知质量指标,即 PESQ(perceptual evaluation of speech quality)提高了 0.08,信号失真比 SDR(signal-to-distortion ratio)提高了 3.3 dB[13]。基于 TCD-TIMIT 数据集测试,PESQ 比 Aviv[32]等人提出的模型提高了 0.39,SDR 提高了 3.7 dB[13]。其中,PESQ 指标用于评价分

离得到的语音的清晰度,取值范围为(-0.5,4.5);SDR 指标与分离得到的语音中残留的噪声量相关,用来评价信号的强度或噪声的强度,数值越大,代表语音质量越好。该模型是与说话人无关的,最多可实现 3 个说话人的语音分离。

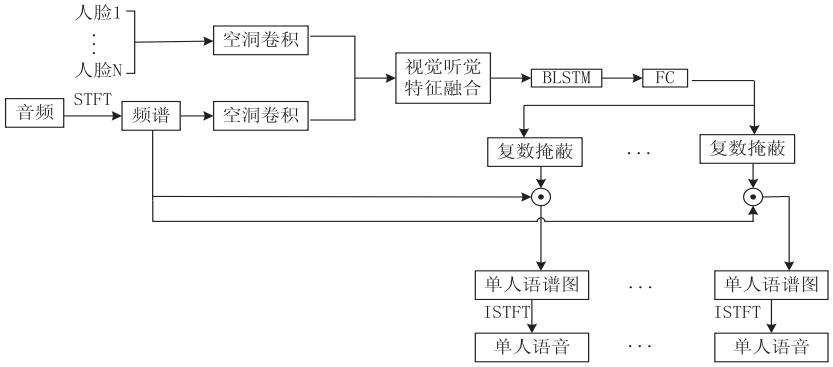

图 4 AV-BLSTM-FC 算法模型框架

前期的一些算法,直接使用含噪语音的相位作为处理后语音的相位,用以合成干净的时域语音信号。这种方法是认为相位信息包含很少的语音信号能量,影响较小,但最近的研究表明,相位信息对于语音信号的聆听质量有较大影响。因此,Afouras[15] 提出了一个视听语音增强网络,结合视觉信息,同时预测音频的幅度和相位,将说话人的声音分离出来。该方法使用 3D Resnet 提取视觉特征,这里称该模型为 AV-amplitude-phase,算法框架如图 5 所示。

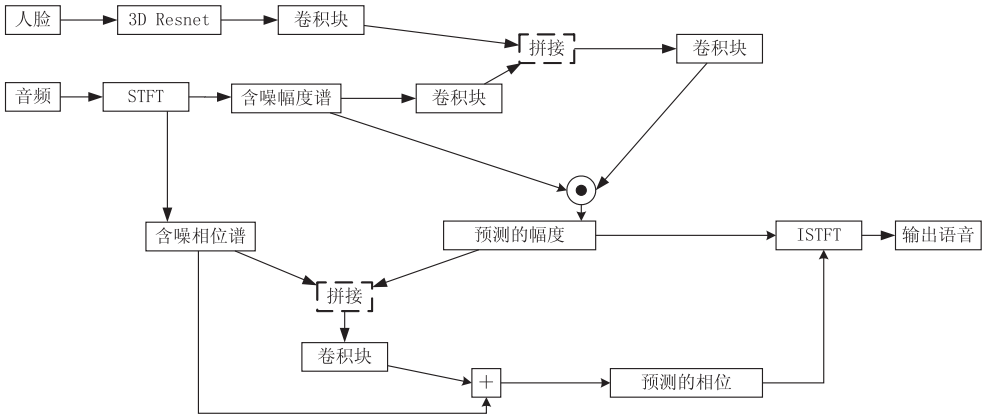

图 5 AV-amplitude-phase 算法框架

模型的损失函数由两部分组成:一部分来自幅度预测,即对预测的语音幅值和真实的语音幅值间的最小绝对值误差(L1 损失)求最小值;另一部分来自相位预测,即对预测相位和真实相位的余弦相似度求最大值。对于未出现在训练集中的说话人,模型依然可以进行语音分离处理。在实验阶段,该方法最多实现了 5 个说话人的语音分离,而且比 Gabbay[17] 提出的方法的语音分离聆听效果清晰保真。基于 LRS2 数据集测试,当说话人数目为 5 人时,算法的 PESQ 指标比直接使用含噪语音相位的方法,提高了 0.1,WER 指标降低了 0.3[15]。数量上看,提高不是很多,但人耳听起来,不同步的谐波带来的“机器人”效应明显改善。但该模型不是端到端的训练,训练分三个阶段:首先训练幅度子网,其次冻结该子网,训练相位子网,最后进行端到端的微调。

Wu[21] 等人提出了一种新的时域视听模型架构,这里称为 AV-TasNet,用于从单声道混合语音中提取目标说话者的语音。该模型结构延续了 TasNet(时域语音分离网络),支持多模态的学习,同时将视听语音分离从频域扩展到了时域。模型使用 3D 卷积神经网络和 Resnet-18 提取视觉特征,算法框架如图 6 所示。

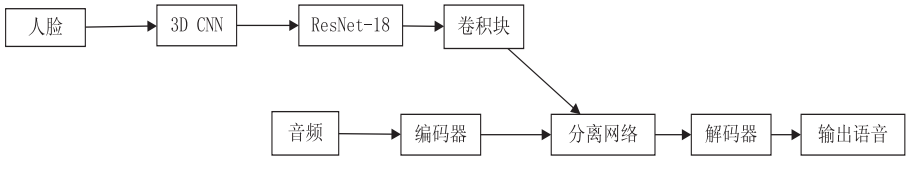

图 6 AV-TasNet 算法框架

模型的目标函数是尺度不变信噪比(scale-invariant source-to-noise ratio,Si-SNR),并且也基于该指标对模型效果进行评价。

$$\text{Si-SNR} = 20\log_{10}\frac{\|\alpha \cdot s_t\|}{\|s_e - \alpha \cdot s_t\|} \tag{3}$$

其中,s_e 是估算的信号,s_t 是目标源信号,α 是最优尺度因子,定义如下:

$$\alpha = \frac{s_e^{\mathrm{T}} s_t}{s_t^{\mathrm{T}} s_t} \tag{4}$$

基于最近发布的 BBC-LRS2 数据集进行的模拟混音实验表明,与仅使用音频的 TasNet[33] 方法相比,可以分别在 2 个和 3 个说话人的情况下,分别提高了 3 dB +和 4 dB +。Si-SNR 指标类似于 SDR,是评价分离语音质量的指标。

5　视听数据集(visual-audio dataset)

训练阶段使用的数据集是基于深度学习的模型性能提升的关键因素。数据集的大小,数据的多样性,包括语料的多样性、噪声的多样性、说话人的多样性等,均可以影响模型的泛化能力和性能。目前应用到视听语音处理的数据集,主要分为两种:在实验室受控环境下录制的数据集和从大型视频网站取材构建的数据集。来源于网络视频的数据集,贴近自然场景,数量大,将会是未来的研究热点。

表 1 是近年来在基于深度学习的语音分离视听模型高频使用的数据集。这些数据集包含视频和音频信息,简称为视听数据集。

表 1　视听数据集

数据集	特点			
	语音文本/语音内容	录音人数	语音时长	其他特点
LRS[8]	句子	多人	4 960 小时; 118 116 个句子;词汇量 17 428;	自然场景下的视听数据集
AVSpeech[13]	多种语言	150 000	约 4 700 小时; 每段 3-10 秒;	说话人年龄、性别、视角 d 多样化,每段只有一个人出现
Mandarin Sentences Corpus[18]	汉语	1	320 句;每句 3-4 秒,每句 10 个汉字;	视频采集在光照充足,安静环境,说话人角度为正面
GRID Corpus[34]	51 个词	34(男:18,女:16)	1 000 句;每句 3 秒;每句 6 个词;	最常用的视听数据集之一,缺点是词汇量小
TCD-TIMIT[35]	英语	60	12 000 段视频	其中包含 3 名唇语者。
VoxCeleb[36]	\	1 251	10 万段	\
VoxCeleb2[37]	\	6 112	100 万段	\
AudioSet[38]	人类与动物声音、乐器与音乐流派声音、日常的环境声音;	\	2 084 320 段; 每段大约 10 秒;	音频所在的 YouTube 视频的 ID,开始时间,结束时间 以及标签(可能是多标签)
BBC-LRS2[8,39]	句子	多人	数千句子	来自 BBC 新闻的五千个小时的视频,对齐字幕,做了嘴唇位置等预处理
LRW[39]	500 个词	数百人	1 000 片段	\
CUAVE[40]	5 遍数字 0-9	36	每个数字 180 遍;	\

以上表格所列视听数据集中,以 GRID Corpus 和 TCD-TIMIT 在视听语音处理模型中最为常用。这两个数据集均为实验室环境下录制制作,与基于互联网视频构建的视听数据集,如 AVSpeech 相比,它们视频环境单一,数据量也相对较小。深度视听方法的一个发展趋势即为把语音处理问题的背景从受控场景转移到自然场景下。类似 AVSpeech 这样取材于自然场景视频的视听数据集,对于算法的性能提升将会非常重要。

6　结束语

结合视觉和听觉多感知特征的深度学习模型,在解决多说话人语音分离问题任务上取得了很大的进步。目前存在的挑战主要有以下几点:一是绝大多数

模型只能适用于已知的说话人。意味着模型推理阶段只能处理训练集中出现过的说话人的语音分离;二是说话人数目的限制与不确定的问题。目前最多实现五、六个人的语音分离,一些模型需要说话人数目已知;三是模型的处理速度。绝大多数模型达不到实时处理的效果;四是语音分离质量。尤其是在说话人数目增多的情况下;五是适用自然场景的能力。一些模型的训练和测试基于实验室采集的干净语音信号或受控场景下的语音信号,一旦迁移到自然场景下的语音分离任务,模型效果会变得很差;六是缺少广泛应用的自然环境下视听数据集。

目前鸡尾酒会问题的主要应用方向可以分为两类:一类是不断提升算法水平,使得在现实场景中语音识别问题上,算法可以模拟达到人类的水平,比如 iPhone 和 iPad 上的智能个人助理,天猫精灵等;第二类是算法帮助有听觉障碍的人,比如助听器的设计。互联网视频网站的蓬勃发展,提供了海量的自然场景下单通道的视频,这就为基于视觉听觉的无监督学习提供了潜在的,大量的数据,这也使得基于深度学习的视听模型方法具备很大的发展潜力。

参考文献:

[1] CHERRY E C. Some experiments on the recognition of speech, with one and with two ears[J]. The Journal of the Acoustical Society of America,1953,25(5):975-979.

[2] BENGIO Y. Learning deep architectures for AI[M]. [s. l.]:Now Publishers Inc,2009.

[3] SCHUSTER M,PALIWAL K K. Bidirectional recurrent neural networks[J]. IEEE Transactions on Signal Processing,1997,45(11):2673-2681.

[4] HE K,ZHANG X,REN S,et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas:IEEE,2016:770-778.

[5] RONNEBERGER O,FISCHER P,BROX T. Unet:convolutional networks for biomedical image segmentation[C]//International conference on medical image computing and computer-assisted intervention. [s. l.]: Springer, 2015: 234-241.

[6] ZHU H,LUO M,WANG R,et al. Deep audio-visual learning:a survey[J]. International Journal of Automation and Computing,2001,18:351-376.

[7] QIAN Y,WENG C,CHANG X,et al. Past review,current progress,and challenges ahead on the cocktail party problem[J]. Frontiers of Information Technology & Electronic Engineering,2018,19(1):40-63.

[8] CHUNG J S, ZISSERMAN A P. Lip reading in profile[C]//The British machine vision association and society for pattern recognition. London:[s. n.],2017.

[9] EPHRAT A,HALPERIN T,PELEG S. Improved speech reconstruction from silent video[C]//Proceedings of the IEEE international conference on computer vision workshops. Venice:IEEE,2017:455-462.

[10] ARANDJELOVIC R, ZISSERMAN A. Objects that sound[C]//Proceedings of the European conference on computer vision (ECCV). Germany:[s. n.],2018:435-451.

[11] ZHAO H,GAN C,ROUDITCHENKO A,et al. The sound of pixels[C]//Proceedings of the European conference on computer vision (ECCV). Germany:[s. n.],2018:570-586.

[12] 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.

[13] 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):1-11.

[14] OWENS A,EFROS A A. Audio-visual scene analysis with self-supervised multisensory features[C]//Proceedings of the European conference on computer vision (ECCV). Germany:[s. n.],2018:631-648.

[15] AFOURAS T,CHUNG J S,ZISSERMAN A. The conversation:deep audio-visual speech enhancement[J]. arXiv:1804.04121,2018.

[16] GABBAY A,SHAMIR A,PELEG S. Visual speech enhancement[J]. arXiv:1711.08789,2017.

[17] 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 (ICASSP). Canada: IEEE,2018:3051-3055.

[18] HOU J C,WANG S S,LAI Y H,et al. Audio-visual speech enhancement using multimodal deep convolutional neural networks[J]. arXiv:1703.10893,2017.

[19] LU R,DUAN Z,ZHANG C. Audio - visual deep clustering for speech separation[J]. IEEE/ACM Transactions on Audio,Speech,and Language Processing,2019,27(11):1697-1712.

[20] LU R,DUAN Z,ZHANG C. Listen and look:audio - visual matching assisted speech source separation[J]. IEEE Signal Processing Letters,2018,25(9):1315-1319.

[21] WU J,XU Y,ZHANG S X,et al. Time domain audio visual speech separation[C]//2019 IEEE automatic speech recognition and understanding workshop (ASRU). Singapore: IEEE,2019:667-673.

[22] AFOURAS T,CHUNG J S,SENIOR A,et al. Deep audio-visual speech recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2018.

[23] LUO Y,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.

[24] 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). Hawaii:IEEE, 2017:3444-3453.

[25] OWENS A,ISOLA P,MCDERMOTT J,et al. Visually indicated sounds[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas:IEEE, 2016:2405-2413.

[26] AYTAR Y, VONDRICK C, TORRALBA A. Soundnet: learning sound representations from unlabeled video[C]// Advances in neural information processing systems. Barcelona:[s. n.],2016:892-900.

[27] WANG W, COSKER D, HICKS Y, et al. Video assisted speech source separation[C]//IEEE international conference on acoustics, speech, and signal processing. America:IEEE, 2005:v/425-v/428.

[28] KHAN F,MILNER B. Speaker separation using visually-derived binary masks[C]//Auditory-visual speech processing (AVSP). France:[s. n.],2013.

[29] SMARAGDIS P,CASEY M. Audio/visual independent components[C]//International symposium on independent component analysis and blind signal separation. Japan: ICA, 2003:709-714.

[30] SEDIGHIN F, BABAIE-ZADEH M, RIVET B, et al. Two multimodal approaches for single microphone source separation[C]//2016 24th European signal processing conference (EUSIPCO). Hungary:IEEE,2016:110-114.

[31] RIVET B,GIRIN L,JUTTEN C. Mixing audiovisual speech processing and blind source separation for the extraction of speech signals from convolutive mixtures[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2006, 15(1):96-108.

[32] GABBAY A,SHAMIR A,PELEG S. Visual speech enhancement using noise-invariant training[J]. arXiv:1711.08789, 2017.

[33] LUO Y,MESGARANI N. Tasnet:time-domain audio separation network for real-time, single-channel speech separation[C]//2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). Canada: IEEE, 2018:696-700.

[34] COOKE M,BARKER J,CUNNINGHAM S,et al. An audio-visual corpus for speech perception and automatic speech recognition[J]. The Journal of the Acoustical Society of America,2006,120(5):2421-2424.

[35] HARTE N,GILLEN E. TCD-TIMIT:an audio-visual corpus of continuous speech[J]. IEEE Transactions on Multimedia, 2015,17(5):603-615.

[36] NAGRANI A, CHUNG J S, ZISSERMAN A. Voxceleb: a large-scale speaker identification dataset[J]. arXiv:1706.08612,2017.

[37] CHUNG J S, NAGRANI A, ZISSERMAN A. Voxceleb2: deep speaker recognition[J]. arXiv:1806.05622,2018.

[38] GEMMEKE J F,ELLIS D P W,FREEDMAN D,et al. Audio set:an ontology and human-labeled dataset for audio events [C]//2017 IEEE international conference on acoustics, speech and signal processing (ICASSP). New Orleans: IEEE,2017:776-780.

[39] CHUNG J S,ZISSERMAN A. Lip reading in the wild[C]// Asian conference on computer vision. [s. l.]: Springer, 2016:87-103.

[40] PATTERSON E K,GURBUZ S,TUFEKCI Z,et al. Moving-talker,speaker-independent feature study, and baseline results using the CUAVE multimodal speech corpus[J]. EURASIP Journal on Advances in Signal Processing,2002,2002(11):208541.

(上接第 7 页)

pression recognition under partial occlusion[C]//2014 tenth international conference on intelligent information hiding and multimedia signal processing. Kitakyushu:[s. n.],2014:211-214.

[48] ZHANG L,VERMA B,TJONDRONEGORO D,et al. Facial expression analysis under partial occlusion: a survey[J]. ACM Computing Surveys,2018,51(2):25.1-25.49.

[49] LI Y, LIU S, YANG J, et al. Generative face completion [C]//2017 IEEE conference on computer vision and pattern recognition (CVPR). Honolulu, HI, USA:IEEE,2017:5892-5900.

[50] 王素琴,高宇豆,张加其. 基于生成对抗网络的遮挡表情识别[J]. 计算机应用研究,2019,36(10):3112-3115.

[51] 王海涌,梁红珠. 基于改进的 GAN 的局部遮挡人脸表情识别[J]. 计算机工程与应用,2020,56(5):141-146.

[52] 姚乃明,郭清沛,乔逢春,等. 基于生成式对抗网络的鲁棒人脸表情识别[J]. 自动化学报,2018,44(5):865-877.

[53] DUAN Q,ZHANG L. BoostGAN for occlusive profile face frontalization and recognition[J]. arXiv:1902.09782,2019.