软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2019,30(Suppl.(2)):25−34

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于知识蒸馏和生成对抗网络的远场语音识别*

邬 龙[1,2], 黎 塔[1], 王 丽[1], 颜永红[1,2,3]

[1](语言声学与内容理解重点实验室(中国科学院 声学研究所),北京 100190)

[2](中国科学院大学,北京 100049)

[3](新疆民族语音语言信息处理实验室(中国科学院 新疆理化技术研究所),乌鲁木齐 830011)

通讯作者: 黎塔, E-mail: lita@hccl.ioa.ac.cn

摘 要: 为了进一步利用近场语音数据来提高远场语音识别的性能,提出一种基于知识蒸馏和生成对抗网络相结合的远场语音识别算法.该方法引入多任务学习框架,在进行声学建模的同时对远场语音特征进行增强.为了提高声学建模能力,使用近场语音的声学模型(老师模型)来指导远场语音的声学模型(学生模型)进行训练.通过最小化相对熵使得学生模型的后验概率分布逼近老师模型.为了提升特征增强的效果,加入鉴别网络来进行对抗训练,从而使得最终增强后的特征分布更逼近近场特征.AMI 数据集上的实验结果表明,该算法的平均词错误率(WER)与基线相比在单通道的情况下,在没有说话人交叠和有说话人交叠时分别相对下降 5.6%和 4.7%.在多通道的情况下,在没有说话人交叠和有说话人交叠时分别相对下降 6.2%和 4.1%.TIMIT 数据集上的实验结果表明,该算法获得了相对7.2%的平均词错误率下降.为了更好地展示生成对抗网络对语音增强的作用,对增强后的特征进行了可视化分析,进一步验证了该方法的有效性.

关键词: 远场语音识别;知识蒸馏;生成对抗式网络;多任务学习;语音增强

中文引用格式: 邬龙,黎塔,王丽,颜永红.基于知识蒸馏和生成对抗网络的远场语音识别.软件学报,2019,30(Suppl.(2)):25–34. http://www.jos.org.cn/1000-9825/19015.htm

英文引用格式: Wu L, Li T, Wang L, Yan YH. Distant speech recognition based on knowledge distillation and generative adversarial network. Ruan Jian Xue Bao/Journal of Software, 2019,30(Suppl.(2)):25−34 (in Chinese). http://www.jos.org.cn/1000-9825/19015.htm

Distant Speech Recognition Based on Knowledge Distillation and Generative Adversarial Network

WU Long[1,2], LI Ta[1], WANG Li[1], YAN Yong-Hong[1,2,3]

[1](Key Laboratory of Speech Acoustics and Content Understanding (Institute of Acoustics, Chinese Academy of Sciences), Beijing 100190, China)

[2](University of Chinese Academy of Sciences, Beiing 100049, China)

[3](Xinjiang Laboratory of Minority Speech and Language Information Processing (Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences), Urumqi 830011, China)

Abstract: In order to further utilize near-field speech data to improve the performance of far-field speech recognition, this paper proposes an approach to integrate knowledge distillation with the generative adversarial network. In this work, a multi-task learning structure is firstly proposed to jointly train the acoustic model with feature mapping. To enhance the acoustic modeling, the acoustic model trained with far-field data (student model) is guided by an acoustic model trained with near-field data (teacher model). Such

* 基金项目: 国家自然科学基金(11590774, 11590770); 新疆维吾尔自治区重大科技专项(2016A03007-1); 中国科学院声学研究所青年英才计划(QNYC201602)

Foundation item: National Natural Science Foundation of China (11590774, 11590770); Key Science and Technology Project of the Xinjiang Uygur Autonomous Region (2016A03007-1); IACAS Young Elite Researcher Project (QNYC201602)

收稿时间: 2019-07-15; 采用时间: 2019-11-04

training process makes the student model mimics the behavior of the teacher model by minimizing the Kullback-Leibler Divergence. To improve the speech enhancement, an additional discriminator network is introduced to distinguish the enhanced features from the real clean ones. The distribution of the enhanced features is further pushed towards that of the clean features through this adversarial multi-task training. Evaluated on AMI single distant microphone data, the method achieves 5.6% relative non-overlapped word error rate (WER) and 4.7% relative overlapped WER decrease over the baseline model. Evaluated on AMI multi-channel distant microphone data, the method achieves 6.2% relative non-overlapped WER and 4.1% relative overlapped WER decrease over the baseline model. Evaluated on the TIMIT data, the method can reach 7.2% WER reduction. To better demonstrate the effects of generative adversarial network on speech enhancement, the enhanced features is visualized and the effectiveness of this method is verified.

Key words: distant speech recognition; knowledge distillation; generative adversarial network; multi-task learning; speech enhancement

近年来,深度学习理论在机器学习领域兴起,其对语音识别技术产生了深远影响.2012 年,Hinton 提出的基于深度神经网络隐马尔可夫模型(deep nerual network hidden Markov model,简称 DNN-HMM)的混合声学建模方案推动了语音识别系统取得了突破性的进展[1].目前,虽然神经网络声学模型很大程度上降低了近场语音识别系统的识别错误率,但识别系统对远场语音的识别准确率仍远低于对近场语音的识别准确率.远场语音中存在的背景噪声、混响以及人声干扰是影响语音识别技术广泛实用化的一个关键因素.故此,如何提升对远场语音的识别准确率在理论和实践两个方面便具有了极为重要的意义.

迄今为止,前人已经提出不少旨在提高远场语音识别性能的方法.这些方法大致可以概括成 3 类:一是采用两阶段训练算法,即先对语音信号进行增强,然后将增强后的信号进行后端声学建模;二是直接利用远场语音数据训练声学模型,并在训练过程中加入混响信息等以及使用更复杂的神经网络结构来进行声学建模;三是将语音信号增强和声学建模放在一个框架里面进行联合优化.

第 1 类方法主要是在去混响和降噪两方面对语音信号进行增强.其中,单通道场景下,Boll 提出了谱减法,通过在频域减去观测信号中的噪声成分来增强语音信号[2].Ephraim 使用分别在频域和对数幅度谱域使用维纳滤波,给出了统计意义下的最小均方误差解[3,4].Mohammadiha 使用了带约束的 NMF 算法并用于有监督和无监督的语音增强,实现了噪声 6 和语音的分离[5,6].随着深度学习技术的快速发展,基于深度神经网络(deep nerual network,简称 DNN)的语音增强方法可以分为两种.一种为频谱映射,即利用 DNN 将带噪声和混响信号的频谱映射成纯净信号频谱[7].另一种是掩蔽估计(mask estimation),基于分类的思想,以各种掩蔽作为 DNN 的训练目标[8,9].在多通道场景下,通常使用波束形成算法来对多通道语音信号进行增强.传统的波束形成算法主要包括固定波束形成,如延迟相加算法[10]、超指向波束形成算法[11]等,以及自适应波束形成,如线性约束最小方差[12]、广义旁瓣抵消算法[13]等.自适应波束形成,也称作自适应滤波,它的另外一种方法是基于信号最小均方误差估计的多通道维纳滤波[14].其中包括语音失真加权的多通道维纳滤波[15]、最小方差无失真响应[16]、广义特征值波束形成[17]等.

第 2 类方法主要在输入特征和声学模型拓扑结构方面提升识别准确率.在输入特征层面,采用多条件训练[18]以及将噪声或混响相关的信息参数化与声学特征拼接作为声学模型的输入[19].在声学模型拓扑方面,采用残差网络、长短时记忆网络(LSTM)等建模能力更强的神经网络来建模[20].

第 3 类方法主要是将语音增强和声学建模放在一个框架里面进行联合优化,从而解决语音增强的优化方向和语音识别字错误率下降的方向不一致的问题.单通道情况下,Gao 等人提出将特征映射和声学建模单独训练,然后再将二者连接使用交叉熵准则联合更新微调参数[21].而特征映射部分通常通过最小化增强后的特征与近场特征之间的均方误差(MMSE)来去除信号中的混响和噪声干扰,从而达到语音增强的目的.然而 MMSE 准则对噪声存在同方差、不相关的假设,这种假设对于真实场景下的语音往往不成立.近年来对抗学习在图像领域取得了巨大的成功,在语音领域被用于语音增强[22,23]、语音转换、声学模型自适应以及说话人自适应等.通过对抗训练可以将不同领域的数据映射成具有相同分布的数据.多通道情况下,Zhong 等人则是通过 LSTM 网络预测波束形成系数[24].Heymann 等人使用神经网络估计统计意义上最优波束形成器的掩蔽值并与声学建模联合训练[25].Sainath 等人提出直接使用原始时域波形信号训练多通道声学模型,用 CNN 网络来实现空间滤波和时间滤波,从而模拟传统的波束形成对信号进行增强[26].

上述 3 类方法固然可以有效地提高远场语音的识别准确率,但只是将近场语音作为语音增强的训练目标或者将其作为声学模型训练数据,并未最大限度地挖掘近场语音的知识.Hinton 等人提出“知识蒸馏”的概念,即利用多个复杂的神经网络构成的组合模型指导一个简单的神经网络进行训练[27].前者被称为老师模型,后者为学生模型.学生模型在训练的过程中模拟老师模型的后验概率分布.在语音识别领域,Li 提出利用一个复杂的 DNN 声学模型指导一个小型的 DNN 声学模型进行训练,从而达到模型压缩的目的[28].Yi 将教师-学生训练思想用于远场语音识别[29].上述方法的核心思想便是采用相对熵来最小化教师模型和学生模型之间后验概率分布的差异.

鉴于此,本文首先利用多任务学习技术将语音增强和声学模型进行联合训练,使得语音增强的优化方向和语音识别字错误率下降的方向保持一致.然后,利用教师-学生训练策略,即用近场数据训练的声学模型来指导远场声学模型的训练,提升远场声学模型的建模性能.最后将生成对抗网络加入上述多任务学习框架中,从而进一步减少增强后的特征和近场特征之间的差异,提高语音增强部分的性能.通过将知识蒸馏和生成对抗网络结合在一起,分别在语音增强和声学建模两方面进一步利用近场数据来提高远场语音识别的性能.本文在 AMI[30]数据集上的实验结果表明,与基线相比,在单通道的情况下,在没有说话人交叠和有说话人交叠时 WER 分别相对下降 5.6%和 4.7%.在多通道的情况下,在没有说话人交叠和有说话人交叠时 WER 分别相对下降 6.2%和 4.1%.在 TIMIT[31]数据集上的实验结果表明,该算法获得了相对 7.2%的 WER 下降.

1 基于语音增强和声学模型联合优化的多任务学习框架

由于背景噪声、混响以及人声干扰等因素,在进行远场声学建模之前通常需要对远场语音信号进行增强.研究发现,语音增强虽然能提高信号的质量,但是并不一定能降低语音识别的字错误率.为此,本文采用多任务学习技术,将特征映射和声学建模放在一起进行联合优化.该算法如图 1 所示,神经网络输出有两个分支,一个是预测三因子状态的后验概率输出,另一个是经过神经网络映射的近场语音特征.

假定输入远场语音特征 $X=\{x_1,...,x_T\}$, 以及与之相对应的近场语音特征 $Y=\{y_1,...,y_T\}$. 基于特征映射的语音增强网络试图学习一个非线性的映射函数 F 将输入的远场特征 X 映射成 $Y^*=\{y_1^*,...,y_T^*\}$. 通过最小化 Y 与 Y^*之间的均方误差 $L_F(\theta_f)$来进行增强:

$$L_F(\theta_f)=\frac{1}{T}\sum_{i}^{T}(F(x_i)-y_i)^2 \tag{1}$$

而对于声学模型 M_s 而言,输入是经过上述增强后的特征 Y^*,并通过最小化声学概率和标注之间的交叉熵 $L_F(\theta_f)$来优化:

$$\begin{aligned}L_{M_s}(\theta_{M_s},\theta_f)&=-\frac{1}{T}\sum_{i=1}^{T}\log P(s_i\mid x_i;\theta_{M_s},\theta_f)\\&=-\frac{1}{T}\sum_{i=1}^{T}\log(M(F(x_i)))\end{aligned} \tag{2}$$

其中,θ_f 为增强网络的参数,θ_{M_s} 为声学模型的参数. $S=\{s_1,...,s_T\}$ 是远场语音每一帧对应的声学标签.对于语音增强和声学模型联合优化的多任务学习而言,联合优化的目标函数 L_{AFM} 可以表示为

$$L_{AFM}=\beta L_F(\theta_f)+(1-\beta)L_{M_s}(\theta_{M_s},\theta_f) \tag{3}$$

其中,β是插值系数,用于权衡语音增强的代价和声学模型代价之间的重要性.

2 知识蒸馏

本文中的知识蒸馏是指用近场数据训练老师模型,而学生模型在训练过程中使用老师模型的后验概率(软标签)而不是远场语音对应的标注信息 0-1 向量(硬标签),如图 2 所示.在学生模型训练的过程中,尽量使其逼近老师模型的后验概率分布,模仿老师的行为.

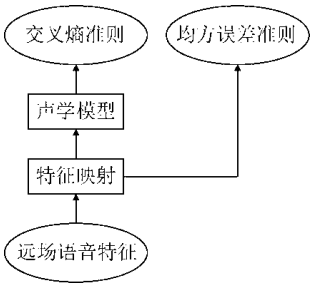

图 1 基于语音增强和声学建模的多任务学习框架

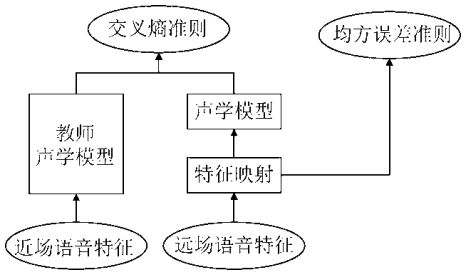

图 2 基于知识蒸馏的多任务学习

假设 x_i^{sdm} 代表远场特征, x_i^{ihm} 代表与之相对应的近场特征.i 代表远场语音帧序号,Q 代表声学状态集合,q 代表每一帧预测的具体状态. $P_{Teacher}(q|x_i^{ihm})$ 和 $P_{Student}(q|x_i^{sdm})$ 分别代表教师模型和学生模型的声学后验概率.二者之间后验概率分布的差异 $KL(P_{Teacher} \| P_{Student})$ 用相对熵(KL divergence)来最小化:

$$KL(P_{Teacher} \| P_{Student}) = \sum_{i=1}^{T} \sum_{q \in Q} P_{Teacher}(q \mid x_i^{ihm}) \log\left(\frac{P_{Teacher}(q \mid x_i^{ihm})}{P_{Student}(q \mid x_i^{sdm})}\right) \tag{4}$$

最小化上述公式等价于优化 $L_{M_s}^{TS}(\theta_f, \theta_{M_s})$:

$$L_{M_s}^{TS}(\theta_f, \theta_{M_s}) = -\sum_{i=1}^{T} \sum_{q \in Q} P_{Teacher}(q \mid x_i^{ihm}) \log(P_{Student}(q \mid x_i^{sdm})) \tag{5}$$

由公式(5)可以看出,优化该公式的最小值也即求交叉熵的最小值.其与式(2)标准交叉熵唯一不同的是训练所需要的标签为老师模型的后验概率分布而不是硬标签.

3 生成对抗网络

虽然基于 MMSE 准则的神经网络特征映射能对远场语音信号起到增强的作用,但是该准则对信号里面的噪声有同方差、不相关的假设.而现实中的远场信号中的非平稳噪声和混响很难满足该假设.因此,本文借助对抗学习的思想,如图 3 所示,在上述多任务学习的框架中加入一个鉴别网络 D,它的输入是增强后的特征以及近场语音特征,输出是属于干净特征(近场特征)的概率,即:

$$P(y_i \in C) = D(y_i) \tag{6}$$

$$P(y_i^* \in E) = 1 - D(y_i^*) \tag{7}$$

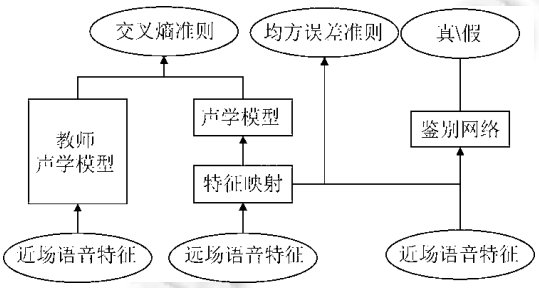

图 3 基于知识蒸馏和生成对抗网络的多任务学习

式(6)、式(7)中,c 和 E 分别代表干净特征和带噪特征的集合.鉴别器的代价函数:

$$L_D(\theta_f,\theta_d)=\frac{1}{T}\sum_{i=1}^{T}[\log P(y_i\in C)+\log P(y_i^*\in E)]$$
$$=\frac{1}{T}\sum_{i=1}^{T}\log D(y_i)+\log[1-D(F(x_i))] \tag{8}$$

为了使得增强后的特征分布更接近近场特征的分布,本文使用对抗学习的策略来训练鉴别器 D 和生成器 F.对于鉴别器的参数 θ_d,通过最小化 $L_D(\theta_f,\theta_d)$来进行更新.对于生成器的参数 θ_f,通过最大化 $L_D(\theta_f,\theta_d)$来进行更新.通过最大最小之间的博弈训练,最终使得增强后的特征足够接近近场特征.

在多任务学习的框架中,需要考虑将经过教师-学生学习策略训练的声学模型加入总的代价函数中,即:

$$L_{total}=L_{AFM}(\theta_f,\theta_{M_s})-\lambda L_D(\theta_f,\theta_d) \tag{9}$$

通过对抗多任务学习技术对式(9)进行联合优化如下:

$$\hat{\theta}_f=\arg\min_{\theta_f}L_{total}(\theta_f,\hat{\theta}_d) \tag{10}$$

$$\hat{\theta}_d=\arg\min_{\theta_d}L_{total}(\hat{\theta}_f,\theta_d) \tag{11}$$

其中,$\hat{\theta}_f$ 和 $\hat{\theta}_d$ 分别代表了在训练过程中生成器和鉴别器的最优参数.本文使用基于随机梯度(SGD)下降的反向传播算法来进行更新:

$$\theta_f\leftarrow\theta_f-\mu\left[\frac{\partial L_{AFM}(\theta_f)}{\partial\theta_f}-\lambda\frac{\partial L_D(\theta_f,\theta_d)}{\partial\theta_f}\right] \tag{12}$$

$$\theta_d\leftarrow\theta_d-\mu\frac{\partial L_D(\theta_f,\theta_d)}{\partial\theta_d} \tag{13}$$

其中,μ为学习率.在推理的时候,只有生成网络 F 和声学模型 M_s 被用于最终的识别.

4 实验结果与分析

4.1 实验数据及配置

本文首先在真实录制的公开数据集 AMI 上进行了相关的实验,验证所提出算法的性能.此数据集包含大约 100h 的会议录音数据.声学信号的录制采用头戴式麦克风(录制近场语音),以及 8 通道的麦克风阵列(录制远场语音).依照 AMI 语料官方网址对数据的划分方式,我们将录音数据分为 3 部分:78 小时的训练集、9 小时的开发集以及 9 小时的远场语音测试集.在本实验中,本文首先在单通道数据上验证了算法的性能,单通道远场语音数据取的是第 1 个麦克风录制的数据.然后在多通道数据上进一步验证算法的性能.多通道实验中,首先将 8 个通道的训练和测试数据经过波束形成进行增强,然后用增强后的数据进行上述算法验证.

为了进一步验证该算法的有效性,本文亦在模拟数据集上进行了相关的实验.首先将 TIMIT 的训练集和测试集均混上真实房间的冲击响应,从而获得模拟的远场语音并将原来干净的 TIMIT 数据作为近场语音.然后在构建的模拟数据集上进行相关的实验.其中,TIMIT 是朗读数据集,由 630 个说话人每人说 10 句话组成总共 6 300 句话,为了获得混响数据,将混响时间 T60 设置为 0.7s,具体混响数据的产生方法均参照文献[31].

本文基于 Tensorflow 来搭建语音识别系统,并基于 kaldi 来生成数据强制对齐的结果.实验中,使用 40 维的 fbank 特征来进行语音增强和声学建模.对于 AMI 数据集而言,声学模型输出 3 992 个状态,对于 TIMIT 数据集而言,声学模型输出 1 934 个状态.采用截断的反向传播(back propagation through time,简称 BPTT)算法来更新模型参数,截断长度为 20 帧.基线声学建模网络为 7 层的前向神经网络.其中输入层上下文各扩展 6 帧组成总共 13 帧的 520 维向量.5 层隐含层的维度为 2 048 并衔接上线性整流函数(RELU)作为激活函数.对于增强网络而言,本文采用 6 层前向神经网络,输入层也进行了上下文 6 帧的扩展,4 层隐含层的维度是 1 024,激活函数也为 RELU.输出是 40 维相应的近场 fbank 特征.

4.2 实验结果与分析

首先,在不引入多任务学习框架的情况下,探究使用近场、远场以及近场和远场混合数据(多条件训练)作为训练数据对远场语音识别结果的影响,见表 1~表 3.表 1 和表 2 分别是 3 个声学模型在 AMI 单通道和多通道数据集下的结果.表 3 是在 TIMIT 模拟数据下的结果.

Table 1 The WER (%) comparison of near-field, far-field and multi-condition acoustic models on single channel AMI dataset

表 1 近场声学模型、远场声学模型、多条件声学模型在单通道 AMI 数据集上的 WER(%)比较

声学模型	WER(over)	WER(non-over)
IHM	76.5	72.9
SDM	54.0	44.5
MCT	**53.0**	**43.2**

Table 2 The WER (%) comparison of near-field, far-field and multi-condition acoustic models on multichannel AMI dataset

表 2 近场声学模型、远场声学模型、多条件声学模型在多通道 AMI 数据集上的 WER(%)比较

声学模型	WER(over)	WER(non-over)
IHM	76.5	72.9
SDM	49.0	39.3
MCT	**48.4**	**38.6**

Table 3 The WER (%) comparison of near-field, far-field and multi-condition acoustic models on TIMIT dataset

表 3 近场声学模型、远场声学模型、多条件声学模型在 TIMIT 数据集上的 WER(%)比较

声学模型	WER
IHM	65.5
SDM	42.1
MCT	**41.6**

其中,SDM 和 IHM 分别表示的是用远场和近场数据训练的声学模型,MCT 表示的是将远场和近场数据混合在一起训练的声学模型.WER(over)和 WER(non-over)指的是在 AMI 测试的时候有无说话人交叠的字错误率.可以看出,直接用近场数据训练的模型测试远场数据性能非常差,说明远场数据中的噪声和混响会严重降低识别系统的性能.比较 MCT 和 SDM 模型可知,在训练的时候加入近场数据会进一步提高识别的准确率.在 AMI 单通道情况下,相较于 SDM 模型,MCT 模型 WER(over)和 WER(non-over)相对下降 1.9%和 2.9%.在 AMI 多通道情况下,MCT 模型 WER(over)和 WER(non-over)相对下降 1.2%和 1.8%.在 TIMIT 数据集上,MCT 模型相较于 SDM 模型 WER 相对下降 1.2%.

接下来,我们采用多任务学习框架,探究了语音增强和声学建模一起联合训练对远场语音识别 WER 的影响.在训练中加入教师-学生训练策略,进一步提升了系统在远场条件下的识别率.具体实验结果见表 4~表 6.

其中,MCT-MSE 表示引入多任务学习框架,将语音增强和声学模型一起进行联合训练.可以看出,相较于 MCT 模型,MCT-MSE 模型在 AMI 单通道情况下,WER(over)和 WER(non-over)相对下降 1.7%和 1.6%.在 AMI 多通道情况下,WER(over)和 WER(non-over)相对下降 0.4%和 0.8%.在 TIMIT 数据集上,WER 相对下降 1.7%.其中,在多通道情况下,加入语音增强联合训练后 WER 下降得比单通道情况下要少,是因为多通道的数据已经经过前端的波束形成起到了前端增强的作用.

为了尽可能地挖掘近场语音中的知识,本文采用教师-学生学习策略来辅助声学模型训练,见表 4~表 6 中最后一行 MCT-MSE-TS.可以看出,相较于 MCT 模型,MCT-MSE 模型在 AMI 单通道情况下,WER(over)和 WER(non-over)相对下降 3.8%和 5.1%.在 AMI 多通道情况下,WER(over)和 WER(non-over)相对下降 3.3%和 4.7%.在 TIMIT 数据集上,WER 相对下降 6.5%.

Table 4 The WER(%) comparison on single channel AMI dataset when applying the multitask learning and knowledge distillation

表 4 加入多任务学习和知识蒸馏技术后,模型在 AMI 单通道数据集上的 WER(%)比较

声学模型	WER(over)	WER(non-over)
MCT	53.0	43.2
MCT-MSE	52.1	42.5
MCT-MSE-TS	**51.0**	**41.0**

Table 5 The WER (%) comparison on multichannel AMI dataset when applying the multitask learning and knowledge distillation

表 5 加入多任务学习和知识蒸馏技术后,模型在 AMI 多通道数据集上的 WER(%)比较

声学模型	WER(over)	WER(non-over)
MCT	48.4	38.6
MCT-MSE	48.2	38.3
MCT-MSE-TS	**46.8**	**36.8**

Table 6 The WER (%) comparison on TIMIT dataset when applying the multitask learning and knowledge distillation

表 6 加入多任务学习和知识蒸馏技术后,模型在 TIMIT 数据集上的 WER(%)比较

声学模型	WER
MCT	41.6
MCT-MSE	40.9
MCT-MSE-TS	**38.9**

最后,为了使得增强后的特征分布更接近近场特征的分布,本文在 MCT-MSE-TS 的基础上,加入鉴别网络进行对抗学习.在远场语音上的识别结果见表 7~表 9.其中,表中的最后一行 MCT-MSE-TS-GAN 表示加入鉴别网络进行对抗训练.可以看出,相较于 MCT 基线模型,加入多任务学习、知识蒸馏和对抗学习后的模型在 AMI 单通道情况下,WER(over)和 WER(non-over)相对下降 4.7%和 5.6%.在 AMI 多通道情况下,WER(over)和 WER(non-over)相对下降 4.1%和 6.2%.在 TIMIT 数据集上,WER 相对下降 7.2%.此结果验证了对抗学习,提升了语音增强的性能,从而提升了远场识别的准确率.

Table 7 The WER(%) comparison on single channel AMI dataset when applying generative adversarial network

表 7 加入生成对抗网络后,模型在 AMI 单通道数据集上的 WER(%)比较

声学模型	WER(over)	WER(non-over)
MCT	53.0	43.2
MCT-MSE-TS	51.0	41.0
MCT-MSE-TS-GAN	**50.5**	**40.8**

Table 8 The WER(%) comparison on multichannel AMI dataset when applying generative adversarial network

表 8 加入生成对抗网络后,模型在 AMI 多通道数据集上的 WER(%)比较

声学模型	WER(over)	WER(non-over)
MCT	48.4	38.6
MCT-MSE-TS	46.8	36.8
MCT-MSE-TS-GAN	**46.4**	**36.2**

Table 9 The WER(%) comparison on TIMIT dataset when applying generative adversarial network

表 9 加入生成对抗网络后,模型在 TIMIT 数据集上的 WER(%)比较

声学模型	WER
MCT	41.6
MCT-MSE-TS	38.9
MCT-MSE-TS-GAN	**38.6**

4.3 语音增强后的特征可视化分析

为了更好地展示生成对抗网络对语音增强的影响.本文将远场语音、与之相对应的近场语音以及经过MCT-MSE 模型增强后的语音和加入生成对抗网络后 MCT-MSE-TS-GAN 模型增强后的语音进行了对比分析.如图 4、图 5 所示.本文展示了 AMI 单通道和多通道语音增强后的特征,在图中,SDM 表示远场语音的特征,IHM 表示近场语音的特征,GAN 表示经过 MCT-MSE-TS-GAN 模型增强后的特征,MSE 表示 MCT-MSE 模型增强后的特征.通过对比,可以看出,将声学模型和语音增强一起联合优化的多任务学习确实能起到增强作用,如方框所示,语音部分得到了增强,噪声部分得到了抑制.但是,和 GAN 相比较,MSE 会出现过度平滑的问题,经过对抗网络训练后的特征保留了更多的特征细节,从而使得特征的分布更接近真实的近场数据.因此,在侧面也可以反映出加入生成对抗网络后,能够在一定程度上提高语音增强部分的性能.

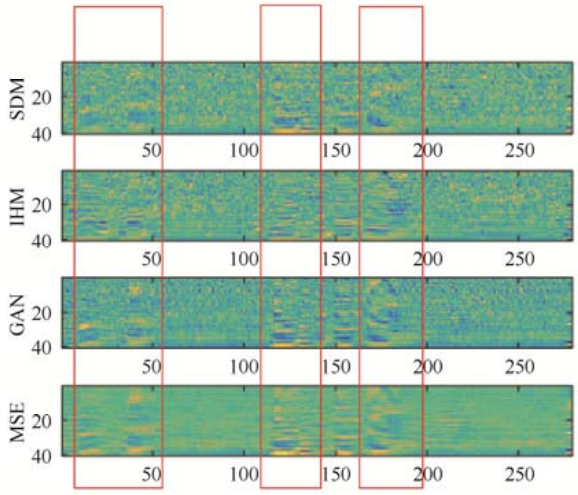

图 4 AMI 单通道语音增强后的特征

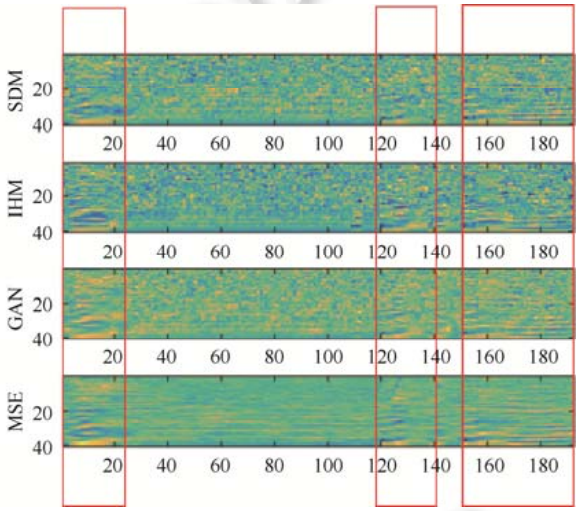

图 5 AMI 多通道语音增强后的特征

5 总 结

针对远场语音识别,本文首先提出一种多任务学习框架,将语音增强和声学建模放在一起进行联合优化.然后分别对上述框架中声学建模和语音增强部分进行改进,进而提升系统远场语音识别的准确率.在声学建模层面,采用教师-学习训练策略充分挖掘近场语音中的知识辅助声学模型的训练,提高声学模型的建模能力.在语音增强部分,采用对抗学习技术使得增强后的特征分布更接近近场特征的分布,从而提升系统的降噪能力.在 AMI 数据集上的实验结果表明,与基线相比,在单通道的情况下,在没有说话人交叠和有说话人交叠时 WER 分别相对下降 5.6%和 4.7%.在多通道的情况下,在没有说话人交叠和有说话人交叠时 WER 分别相对下降 6.2%和 4.1%.在 TIMIT 数据集上的实验结果表明,该算法获得了相对 7.2%的 WER 下降.为了更好地展示生成对抗网络对语音增强的作用,本文对增强后的特征进行了可视化分析,进一步验证了该方法的有效性.

References:

[1] Hinton G, Deng L, Yu D, *et al*. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Processing Magazine, 2012,29(6):82−97.

[2] Boll SF. Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. on Acoustics Speech & Signal Processing, 1979,27(2):113−120.

[3] Ephraim Y, Malah D. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. IEEE Trans. on Acoustics, Speech, and Signal Processing, 1984,32(6):1109−1121.

[4] Ephraim Y, Malah D. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Trans. on Acoustics, Speech, and Signal Processing, 2003,33(2):443−445.

[5] Lee DD, Seung HS. Algorithms for non-negative matrix factorization. In: Proc. of the Neural Information Processing Systems. 2000. 556−562.

[6] Mohammadiha N, Smaragdis P, Leijon A. Supervised and unsupervised speech enhancement using nonnegative matrix factorization. IEEE Trans. on Audio Speech & Language Processing, 2017,21(10):2140−2151.

[7] Xu Y, Du J, Dai L, *et al*. A regression approach to speech enhancement based on deep neural networks. IEEE Trans. on Audio, Speech, and Language Processing, 2015,23(1):7−19.

[8] Wang D, Chen J. Supervised speech separation based on deep learning: An overview. IEEE Trans. on Audio, Speech, and Language Processing, 2018,26(10):1702−1726.

[9] Wang Y, Narayanan A, Wang D, *et al*. On training targets for supervised speech separation. IEEE Trans. on Audio, Speech, and Language Processing, 2014,22(12):1849−1858.

[10] Anguera X, Wooters C, Hernando J. Acoustic beamforming for speaker diarization of meetings. IEEE Trans. on Audio, Speech and Language Processing, 2007,15(7):2011−2022.

[11] Yaakov B, *et al*. Asymmetric beampatterns with circular differential microphone arrays. In: Proc. of the 2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). 2017. 190−194.

[12] Frost OL. An algorithm for linearly constrained adaptive array processing. Proc. of the IEEE, 1972,60(8):926−935.

[13] Griffiths LJ, Jim CW. An alternative approach to linearly constrained adaptive beamforming. IEEE Trans. on Antennas and Propagation, 1982,30(1):27−34.

[14] Simon D, Moonen M. GSVD-based optimal filtering for single and multimicrophone speech enhancement. IEEE Trans. on Signal Processing, 2002,50:2230−2244.

[15] Simon D, Spriet A, Wouters J, *et al*. Speech distortion weighted multichannel wiener filtering techniques for noise reduction. In: Speech Enhancement. Springer-Verlag, 2005. 199−228.

[16] Zelinski R. A microphone array with adaptive post-filtering for noise reduction in reverberant rooms. In: Proc. of the Int'l Conf. on Acoustics. IEEE, 1988. 2578−2581.

[17] Warsitz E, Haeb-Umbach MR. Blind acoustic beamforming based on generalized eigenvalue decomposition. IEEE Trans. on Audio, Speech and Language Processing, 2007,15(5):1529−1539.

[18] Ko T, Peddinti V, Povey D, *et al*. A study on data augmentation of reverberant speech for robust speech recognition. In: Proc. of the IEEE Int'l Conf. on Acoustics. IEEE, 2017. 5220−5224.

[19] Seltzer ML, Yu D, Wang Y. An investigation of deep neural networks for noise robust speech recognition. In: Proc. of the Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2013. 7398−7402.

[20] Zhang Y, Chen G, Yu D, *et al*. Highway long short-term memory RNNS for distant speech recognition. In: Proc. of the Int'l Conf. on Acoustics, Speech, and Signal Processing. 2016. 5755−5759.

[21] Gao T, Du J, Dai LR, *et al*. Joint training of front-end and back-end deep neural networks for robust speech recognition. In: Proc. of the IEEE Int'l Conf. on Acoustics. IEEE, 2015. 4375−4379.

[22] Pascual S, Bonafonte A, Serra J. Segan: Speech enhancement generative adversarial network. In: Proc. of the Annual Conf. of the Int'l Speech Communication Association. 2017. 3642−3646.

[23] Meng Z, *et al*. Adversarial feature-mapping for speech enhancement. In: Proc. of the IEEE Int'l Conf. on Acoustics. IEEE, 2018. 3259−3263.

[24] Meng Z, Watanabe S, Hershey JR, *et al*. Deep long short-term memory adaptive beamforming networks for multichannel robust speech recognition. In: Proc. of the Int'l Conf. on Acoustics, Speech, and Signal Processing. 2017. 271−275.

[25] Heymann J, Drude L, Boeddeker C, *et al*. Beamnet: End-to-end training of a beamformer-supported multi-channel ASR system. In: Proc. of the IEEE Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017. 5325−5329.

[26] Sainath TN, Weiss RJ, Wilson KW, *et al*. Multichannel signal processing with deep neural networks for automatic speech recognition. IEEE/ACM Trans. on Audio, Speech, and Language Processing, 2017,25(5):965−979.

[27] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. Computer Science, 2015,14(7):38−39.

[28] Li J, Zhao R, Huang J, *et al*. Learning small-size DNN with output-distribution-based criteria. In: Proc. of the Conf. of the Int'l Speech Communication Association. 2014. 1910−1914.

[29] Yi J, Tao J, Wen Z, *et al*. Distilling knowledge using parallel data for far-field speech recognition. arXiv Preprint arXiv: 1802.06941, 2018.

[30] Carletta J, Ashby S, Bourban S, *et al*. The AMI meeting corpus: A pre-announcement. In: Proc. of the Int'l Conf. on Machine Learning. 2005. 28−39.

[31] Ravanelli M, Brakel P, Omologo M, *et al*. Batch-normalized joint training for DNN-based distant speech recognition. In: Proc. of the Spoken Language Technology Workshop. 2016. 28−34.

邬龙(1991－),男,河南信阳人,学士,主要研究领域为远场语音识别.

王丽(1985－),女,副研究员,主要研究领域为语音识别声学建模.

黎塔(1982－),男,博士,研究员,主要研究领域为语音识别,语音信号处理,人机交互,海云计算.

颜永红(1967－),男,博士,研究员,博士生导师,CCF 专业会员,主要研究领域为语音信号处理,听感知,人机交互,海云计算.