DOI: 10.19650/j.cnki.cjsi.J2210872

基于感知条件网络的可控语音增强模型*

袁文浩,屈庆洋,梁春燕,夏 斌
(山东理工大学计算机科学与技术学院 淄博 255000)

摘 要:为了给不同听者在不同场景下提供更好的语音增强主观听觉感受,提出了一种基于感知条件网络的可控语音增强模型。首先设计分位数损失函数来对语音的高估和低估进行权衡,并以此来指导网络的训练,通过调节网络输出中的语音损失和噪声残留水平,来控制模型的输出特性。然后为了让单个网络具有可变的输出特性,引入条件网络,利用分位数损失函数中与听者感知相关的分位值产生条件信息来对含噪语音特征进行调制,建立了可控的语音增强模型。实验结果表明,设计的分位数损失函数能够有效调节增强语音中的语音损失和噪声残留水平;基于感知条件网络建立的可控语音增强模型,能够提供可由听者主动控制的增强语音输出特性,使听者获得更好的语音增强体验。

关键词: 语音增强;深度学习;深度神经网络;条件网络;损失函数
中图分类号: TN912.3 TH701 **文献标识码**: A **国家标准学科分类代码**: 510.40

Controllable speech enhancement model based on perceptual conditional network

Yuan Wenhao, Qu Qingyang, Liang Chunyan, Xia Bin
(*School of Computer Science and Technology, Shandong University of Technology, Zibo 255000, China*)

Abstract: To provide better subjective auditory perception of speech enhancement for different listeners in different environments, a controllable speech enhancement model based on the perceptual conditional network is proposed. First, a quantile loss function is designed to balance the overestimation and underestimation of speech, which is used to guide the training of network. In this way, the output characteristics of model are controlled by adjusting the level of noise residual and speech distortion in the output of the network. Then, to make a single speech enhancement network has variable output characteristics, the conditional network is introduced. The conditional information is generated by the quantile value related to auditory perception in the quantile loss function to modulate the noisy speech features, and a controllable speech enhancement model is established. The experimental results show that, the designed quantile loss function can effectively adjust the level of residual noise and speech distortion in the enhanced speech, and the proposed controllable speech enhancement model based on the perceptual conditional network can provide variable characteristics of enhanced speech that can be actively controlled by the listener. The listener can get a better speech enhancement experience.

Keywords: speech enhancement; deep learning; deep neural network; conditional network; loss function

0 引 言

传统语音增强方法一般基于统计学原理,其对于平稳噪声具有较好的处理效果,但是在低信噪比和非平稳噪声条件下性能会急剧下降。针对传统语音增强方法的不足,近年来,研究人员将深度学习技术应用于语音增强,提出了基于深度神经网络的语音增强方法。基于深度神经网络的语音增强方法利用大量语音和噪声样本数据进行网络训练,建立起含噪语音和增强语音之间的映射关系,相比传统方法显著提高了语音增强性能[1]。

研究人员对基于深度神经网络的语音增强方法开展了广泛的研究,提出了多种不同形式的语音增强网络(speech enhancement network, SE-NET)。根据语音增强

收稿日期:2022-12-13 Received Date: 2022-12-13
*基金项目:山东省自然科学基金(ZR2022MF330,ZR2021MF017)、国家自然科学基金(61701286)项目资助

网络所采用的特征形式,基于深度神经网络的语音增强方法可以分为时域、时频域以及时域和时频域结合 3 类。

时域方法直接以含噪语音的波形特征作为输入,通过语音增强网络处理,输出相应增强语音的波形,是一种端到端的处理方法。时域方法不需要进行信号的时频分析,具有更简单的处理流程,且能够充分利用语音的相位信息。但是,由于时域波形特征缺乏明显的结构信息,需要更加复杂的网络结构对其进行建模[2-8]。

时频域方法以含噪语音短时傅里叶变换后的时频谱特征作为输入,根据是否在语音增强网络中考虑相位信息,时频域方法又可以分为两类。第 1 类方法在网络的设计中忽略语音的相位信息,采用含噪语音短时傅里叶变换后的幅度谱、功率谱或者对数功率谱特征作为输入,通过语音增强网络处理,计算得到相应增强语音的幅度谱,然后使用含噪语音的相位谱来重构增强语音[9-14]。第 2 类方法在网络设计中考虑了相位信息,采用含噪语音短时傅里叶变换后的复数谱特征作为输入,通过语音增强网络处理,计算得到相应增强语音的复数谱,然后重构得到增强语音[15-22]。

基于时域和时频域方法各自的优缺点,研究人员尝试将两种方法进行结合,提出了基于时域和时频域结合的方法。该类方法将一个时域语音增强网络与一个时频域语音增强网络进行级联,同时估计纯净语音的时域波形和时频域频谱[23-25]。

经过近几年的快速发展,基于深度神经网络的语音增强方法的研究工作已经取得了诸多成果。但是,在 INTERSPEECH 2021 DNS 挑战赛的比赛结果分析中,赛事组织者指出,对于由 speech MOS 测量的主观语音质量,所有参加比赛的 19 支队伍中只有 2 支队伍取得了比含噪语音更好的结果[26],表明现有的大多数语音增强方法在追求噪声抑制性能的同时,牺牲了语音保持的性能。

实际上,对于所有的语音增强方法,增强语音与纯净语音之间的误差都可以归纳为两个方面:语音损失和噪声残留。对于一个已经训练好的语音增强网络,由于其参数是固定的,对于一段含噪语音,其输出的增强语音中语音损失和噪声残留水平也是固定的。而不同听者对语音损失和噪声残留的接受程度是不同的,同一网络对不同含噪语音的语音保持和噪声抑制水平也是不同的;因此,单一固定的语音增强网络不能为不同听者在不同噪声场景下提供持续良好的语音增强主观听觉感受。基于以上分析,为了给听者提供更好的语音增强主观体验,本文提出一种基于感知条件网络的可控语音增强模型,该模型具备可控的多种不同输出特性,能够依据听者的主观听觉感受权衡增强语音中的语音损失和噪声残留水平。

1 可控语音增强模型

1.1 基于网络切换的可控语音增强模型

对于语音增强的效果,基于耳朵的主观听觉感受,人们虽然不能进行增强语音质量和可懂度的精确评价,但是能够对增强语音中语音损失的严重程度和噪声残留的多少进行粗略判断,这一判断能够反映语音增强系统在当前场景下的输出特性,可以作为反馈帮助调整优化语音增强网络,将这一主观判断定义为感知反馈。

建立可控的语音增强模型的关键是如何训练与感知反馈相关的具有不同输出特性的一系列语音增强网络。本文以损失函数的设计作为网络训练时输出特性的控制手段,在损失函数中对语音保持和噪声抑制水平进行参数化调整,分别为语音保持和噪声残留设计独立的损失函数 L_s 和 L_d,模型的整体损失函数为两个损失函数的加权和,即:

$$L = \lambda \times L_s + (1 - \lambda) \times L_d \tag{1}$$

其中,L_s 衡量的是语音损失水平,L_d 衡量的是噪声残留的水平,λ 是权衡两者之间关系的权重参数。显然,不同的 λ 对应不同的损失函数,因此对应不同的 λ 可以训练不同的语音增强网络,每个语音增强网络在语音保持和噪声抑制之间建立了不同的权衡关系,也即拥有不同的输出特性。在进行语音增强时,听者可以根据自己的主观听觉感受在不同的网络之间进行切换,选择具有更好听感的网络;也即,听者可以通过调节 λ 的值来控制模型的输出特性,实现语音增强模型的动态调整,因此 λ 可以作为联系听者主观听觉感知和模型输出特性的可控系数,如图 1 所示。

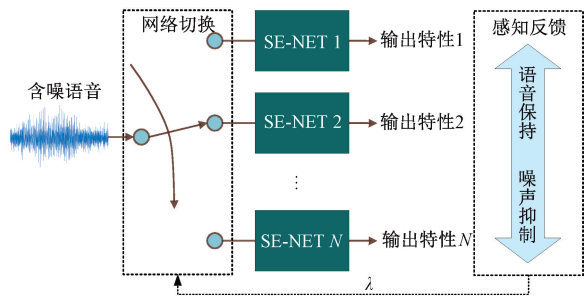


图 1 基于网络切换的可控语音增强模型

Fig. 1 Controllable speech enhancement model based on network switching

1.2 分位数损失函数

需要注意的是,在真实语音中,语音损失和噪声残留并非是独立存在的,两者往往共同出现,相互影响,难以

进行独立的衡量,式(1)中的损失函数是无法实现的。针对该问题,考虑对于语音的高估往往会引入噪声成分残留,而对于语音的低估往往会带来语音成分的损失,本文为时频域语音增强设计一种分位数损失函数,通过损失函数对语音的高估和低估进行权衡,间接来控制增强语音中的语音损失和噪声残留水平。

在时频域,对于语音频谱的幅度掩蔽,分位数损失函数定义为:

$$L_{quantile}(R,\hat{R}) = \max(\lambda \times (\hat{R}-R),(1-\lambda)\times(R-\hat{R})) \tag{2}$$

其中,$R = |S|/|Y|$、$\hat{R} = |\hat{S}|/|Y|$分别是理想幅度谱掩蔽及其估计值,S、Y 和 $\hat{S}$ 分别代表含噪语音谱、纯净语音谱和估计语音谱。由式(2)可知,分位数损失函数能够通过分位值 $\lambda(0<\lambda<1)$ 对语音幅度谱的高估和低估给予不同的惩罚,如果分位值 λ 取较小的值,那么在网络训练时,为了使损失函数下降的更快,网络会侧重于输出 $\lambda\times(\hat{R}-R)$,即使 $(\hat{R}-R)$ 取得大于 0 的值,此时会产生对语音幅度谱的高估,容易引入噪声;而如果分位值 λ 取较大的值,网络则侧重输出 $(1-\lambda)\times(R-\hat{R})$,即使 $(R-\hat{R})$ 取得大于 0 的值,此时容易造成对语音幅度谱的低估。因此,在时频域语音增强网络训练时,可以通过设置不同的分位值 λ,实现对语音估计的控制。

利用式(2)的分位数损失函数,针对 N 个不同的分位值 λ 来训练 N 个 SE-NET,就可以构建图 1 所示的基于网络切换的可控语音增强模型。在使用图 1 模型进行语音增强时,听者可以通过选择使用对应不同分位值的网络,来调控模型的输出在语音损失和噪声残留之间的权衡关系,获得更好的语音增强体验。

1.3　感知条件语音增强网络

图 1 模型中的多个网络是独立训练、各不相同的,带来较高的模型训练时间和空间复杂度;并且,分位值 λ 的取值越多,需要训练的网络越多,模型的空间复杂度越高。针对这一问题,考虑以与听者主观听觉感知相关的 λ 作为网络参数调节的约束条件,利用条件网络结构,训练具有可变输出特性的单个网络模型来实现输出可控的语音增强。

感知条件语音增强网络通过将条件信息与含噪语音特征信息在网络计算中进行融合,从而实现在相同输入下能够获得与条件信息相关的不同输出。本文采用特征线性调制来实现条件信息和特征信息的融合,利用由 λ 产生的条件信息对 SE-NET 中的特征信息进行调制[27],如图 2 所示。假设某一层的输出为特征 $\boldsymbol{f}$,经过线性调制后的特征为:

$$\tilde{\boldsymbol{f}} = \boldsymbol{\alpha}\cdot\boldsymbol{f}+\boldsymbol{\beta} \tag{3}$$

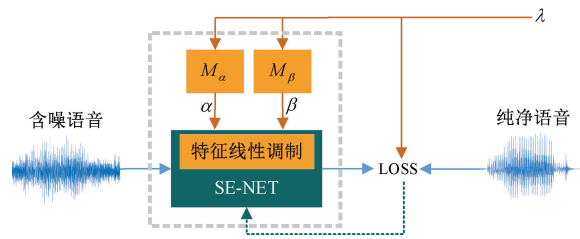


图 2　感知条件语音增强网络结构

Fig. 2　The architecture of perceptual conditional speech enhancement network

其中,$\boldsymbol{\alpha}$ 和 $\boldsymbol{\beta}$ 分别是由 λ 经过网络 M_α 和 M_β 运算得到。

$$\boldsymbol{\alpha} = M_\alpha(\lambda) \tag{4}$$

$$\boldsymbol{\beta} = M_\beta(\lambda) \tag{5}$$

1.4　基于感知条件网络的可控语音增强模型

利用感知条件语音增强网络改进图 1 中的多网络切换模型,得到本文的基于感知条件网络的可控语音增强模型,如图 3 所示。通过设计合理的训练策略,图 3 模型能够获得同图 1 模型一样的 N 种输出特性。相比图 1 模型需要独立训练 N 个不同的 SE-NET,图 3 模型只需要训练单个感知条件语音增强网络即可获得;在采用相同结构的 SE-NET 时,通过合理控制感知条件语音增强网络中 M_α 和 M_β 的参数规模,图 3 模型显然具有更小的空间复杂度。

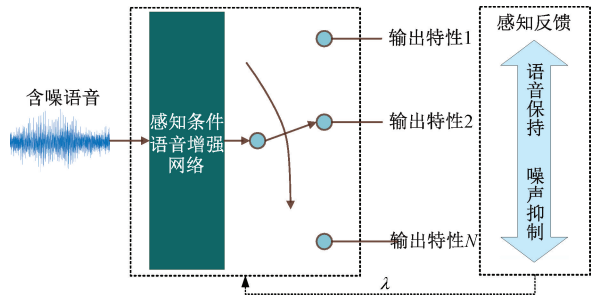


图 3　基于感知条件网络的可控语音增强模型

Fig. 3　Controllable speech enhancement model based on the perceptual conditional network

在利用基于感知条件网络的可控语音增强模型进行语音增强时,听者可以依据自己的主观听觉感受,通过调节可控系数 λ,来调控模型的输出在语音保持和噪声残留之间的关系,获得更好的语音增强体验。

2　实验数据与配置

2.1　实验数据

为了验证基于感知条件网络的可控语音增强模型的有效性,采用语音增强研究中广泛使用的公开数据集来

进行网络的训练与测试,其语音信号来自 Voice Bank 数据集,噪声信号则来自 DEMAND 噪声数据集[28]。训练集由 11 572 段含噪语音及其相应的纯净语音构成,其语音数据来自 28 个说话人(14 男/14 女),整个训练集包含 10 种不同的噪声类型和 0、5、10 和 15 dB 4 种不同的信噪比;测试集由 824 段含噪语音及其相应的纯净语音构成,其语音数据来自与训练集不同的 2 个说话人,整个测试集包含 5 种与训练集不同的噪声和 2.5、7.5、12.5 和 17.5 dB 4 种不同的信噪比。

数据集的原始采样频率为 48 kHz,实验中将其重采样为 16 kHz。短时傅里叶变换使用窗长为 512 点,帧移为 256 点的汉宁窗。

2.2 网络结构

采用一种比较简单的时频域 SE-NET 来验证本文模型的有效性,其隐层为 5 层的 GRU 网络,输出层为 1 个线性层;每个隐层的节点个数为 512,输入为 257 维的含噪语音对数功率谱,输出为 257 维的幅度谱掩蔽值。感知条件语音增强网络中的线性特征调制针对 SE-NET 的 5 个隐层的输出特征进行,其中 M_α 和 M_β 均为 3 个隐层的全连接网络,隐层节点数为 1 024,输出层的节点数为 2 560,对应 SE-NET 的 5 层的输出特征(5×512)。

2.3 评价指标

为了评估模型的语音增强性能,采用如下 4 种客观评价指标,分别为反应语音失真程度的 CSIG,反应背景噪声干扰程度的 CBAK,反应整体语音质量的 COVL[29] 以及语音质量感知评估 PESQ[30]。其中,CSIG、CBAK、COVL 的得分范围为 1 ~ 5,PESQ 的得分范围为 -0.5~4.5;得分越高,代表相应的语音质量越好。

3 实验结果与分析

3.1 分位数损失函数的有效性分析

首先,为了验证分位数损失函数对网络输出特性的调节能力,将式(1)中的分位值 λ 分别赋值为 0.1、0.2、…、0.9,采用相同的网络结构和训练方法,分别训练 9 个不同的 SE-NET。表 1 给出了对应不同分位值的 9 个 SE-NET 的语音增强性能,另外还给出了以平均绝对误差 MAE 训练的相同结构网络的语音增强性能。可见,当 λ 取不同值时,9 个 SE-NET 的各项客观评价指标各不相同,表现出了显著不同的语音增强性能。这表明,通过设置分位数损失函数的分位值,可以训练不同的 SE-NET,达到调节网络输出特性的目的。

综合分析 4 种客观评价指标,当 λ = 0.8 时,所训练的 SE-NET 具有最好的语音增强性能。另外,相比采用平均绝对误差训练的 SE-NET,当分位值 λ 取为 0.6、0.7、0.8 时,分位数损失函数训练的网络在 4 个指标下都要更好,表明通过恰当选择 λ 的值,本文的分位数损失函数还能够有效提高网络的语音增强性能。

表 1 不同损失函数训练的 SE-NET 的性能

Table 1 Performance of SE-NET trained with different loss functions

损失函数	CSIG	CBAK	COVL	PESQ
λ = 0.1	3.59	2.57	2.82	2.07
λ = 0.2	3.81	2.82	3.05	2.31
λ = 0.3	3.90	2.98	3.16	2.43
λ = 0.4	3.98	3.07	3.25	2.53
λ = 0.5	4.04	3.17	3.34	2.64
λ = 0.6	4.07	3.25	3.38	2.69
λ = 0.7	4.11	3.31	3.45	2.78
λ = 0.8	4.10	3.36	3.48	2.84
λ = 0.9	4.00	3.33	3.40	2.81
MAE	4.05	3.16	3.35	2.64

图 4(a)~(i)分别给出了测试集中的一段含噪语音采用 9 个 SE-NET 处理后的增强语音的语谱图,图 4(j)与图 4(k)则分别给出了原始含噪语音和纯净语音的语谱图作为对比。可见,随着分位值 λ 不断增大,增强语音中的噪声成分残留越来越少,但是语音成分损失也越来越明显,直观的表明了本文的分位数损失函数具有调节增强语音中的噪声残留和语音损失水平的能力。综合分析 9 个增强语音的语谱图,可见,λ = 0.8 所对应的增强语音的语谱图与纯净语音的语谱图最接近,表明 λ = 0.8 所对应的网络在噪声抑制和语音保持间达到了较好的平衡,具有最好的语音增强性能,这与客观指标的分析结果是一致的。

3.2 基于感知条件网络的可控语音增强模型性能分析

对于基于感知条件网络的可控语音增强模型的训练,每次迭代给 λ 随机取到{0.1、0.2、…、0.9}中的一个值,同步完成 M_α、M_β 和 SE-NET 的训练。在语音增强时,就可以通过将 λ 设为{0.1、0.2、…、0.9}中的任一值,来调整模型的输出特性。表 2 给出了基于感知条件网络的可控语音增强模型在 λ 取不同值时的语音增强性能。可见,当 λ 取不同值时,模型的各项客观评价指标各不相同,表现出显著不同的语音增强性能。这表明,基于感知条件网络的可控语音增强模型通过结合分位数损失函数与条件语音增强网络,具备了模型输出特性可控的能力。

下面对比基于感知条件网络的可控语音增强模型与 3.1 节中独立训练的语音增强网络的性能,图 5(a)~(d)分别给出了不同 λ 下两者相应的 4 种客观评价指标的

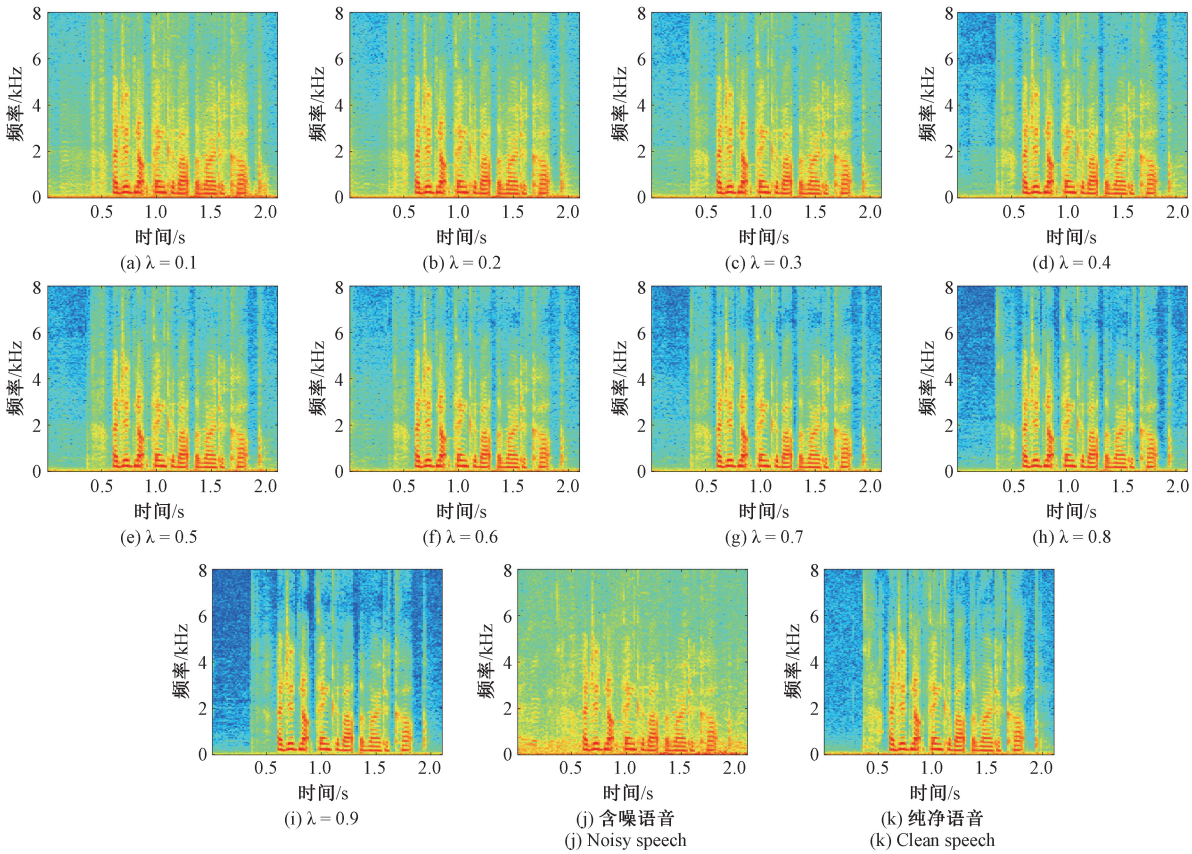


图 4　不同 λ 下训练的 SE-NET 增强后语音的语谱图

Fig. 4　Spectrograms of speech enhanced by SE-NETs trained using different λ

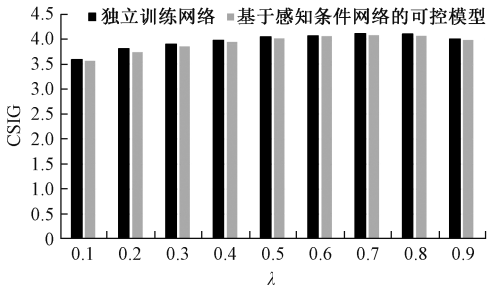
表 2　不同 λ 下基于感知条件网络的可控语音增强模型的性能

Table 2　Performance of the controllable speech enhancement model based on perceptual conditional network under different λ

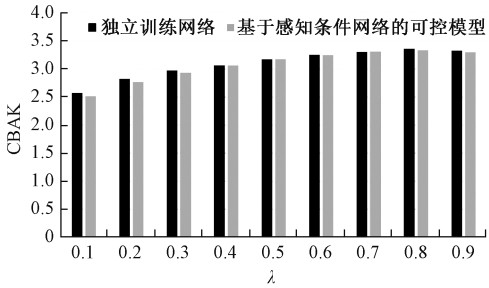
λ	CSIG	CBAK	COVL	PESQ
0. 1	3. 57	2. 51	2. 77	2. 01
0. 2	3. 75	2. 77	2. 97	2. 21
0. 3	3. 86	2. 94	3. 10	2. 36
0. 4	3. 95	3. 07	3. 22	2. 49
0. 5	4. 02	3. 18	3. 31	2. 60
0. 6	4. 06	3. 26	3. 37	2. 69
0. 7	4. 08	3. 32	3. 42	2. 76
0. 8	4. 07	3. 34	3. 44	2. 80
0. 9	3. 98	3. 30	3. 39	2. 81

对比结果。可见基于感知条件网络的可控语音增强模型的性能随 λ 的变化规律与独立训练的语音增强网络的性能随 λ 的变化规律是基本一致的,两者都在 λ = 0. 8 时取得最优结果;综合分析 4 种指标,相同 λ 下基于感知条件网络的可控语音增强模型的性能要略差,但是两者的差距较小,4 种指标的不同分位值下的平均差分别为 0. 03、0. 02、0. 04、0. 04。考虑到基于感知条件网络的可控语音增强模型中的 SE-NET 与独立训练的语音增强网络具有相同的结构,可以得出结论:基于感知条件网络的可控语音增强模型中采用的条件网络结构与训练策略是有效的,能够较好的保持主干 SE-NET 的性能;并且基于感知条件网络的可控语音增强模型在减小空间复杂度的条件下,并没有显著降低语音增强性能。

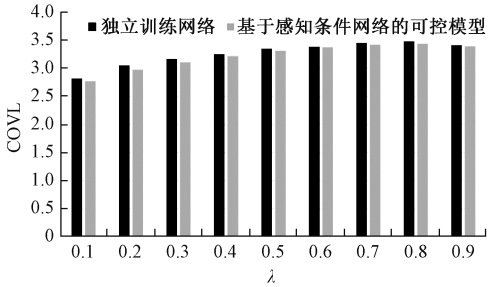
为了进一步验证不同场景下输出特性控制带来的语音增强性能提升,下面以与听者主观听觉感知相关程度较高的 PESQ 指标为例进行语音增强性能的分析。在使用基于感知条件网络的可控语音增强模型对测试集进行语音增强时,由于测试集中的含噪语音包含不同的噪声和不同的信噪比,为了使模型达到最佳的语音增强性能,需要对应每段含噪语音单独设置分位值来控制模型的输出特性。在本实验中,对于一段含噪语音,定义使其增强后语音 PESQ 得分最高的分位值为最佳分位值 λ^*;显然,当对应每段含噪语音的分位值都为最佳分位值 λ^* 时,可以得到 PESQ 指标下模型的最佳性能。表 3 给出了基于感知条件网络的可控语音增强模型在不同分位值



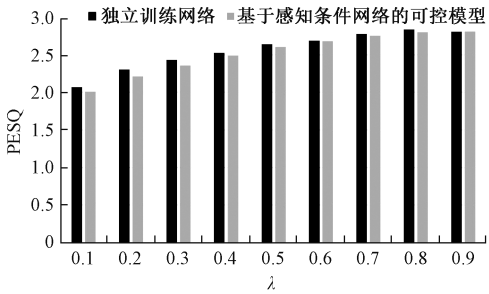
(a) 不同λ下的CSIG对比
(a) Comparison of CSIG under different λ



(b) 不同λ下的CBAK对比
(b) Comparison of CBAK under different λ



(c) 不同λ下的COVL对比
(c) Comparison of COVL under different λ



(d) 不同λ下的PESQ对比
(d) Comparison of PESQ under different λ

图 5 不同 λ 下的独立训练网络与基于感知条件网络的可控模型的语音增强性能比较

Fig. 5 Comparison of speech enhancement performance between independent trained networks and the controllable model based on perceptual conditional network under different λ

下的平均 PESQ 得分。可见,当为每段含噪语音设置最佳分位值时,模型取得明显更高的平均 PESQ 得分,表明通过进行输出特性的控制,本文模型能够为听者在不同噪声条件下提供更好的语音增强性能。

表 3 不同 λ 下的平均 PESQ 得分

Table 3 Average PESQ score under different λ

λ	0.1	0.2	0.3	0.4	0.5
PESQ	2.01	2.21	2.36	2.49	2.60
λ	0.6	0.7	0.8	0.9	λ^*
PESQ	2.69	2.76	2.80	2.81	2.85

4 结 论

为了给不同听者在不同场景下持续提供良好的语音增强主观听觉感受,需要语音增强模型具有可控的多种输出特性。考虑对于语音的高估往往会引入残留噪声成分,而对于语音的低估往往会带来语音成分的损失,本文设计了一种用于时频域语音增强的分位数损失函数来对语音的高估和低估进行权衡,从而间接控制增强语音中的语音损失和噪声残留水平,以达到控制语音增强模型输出特性的目的。以此为基础,为了让单个语音增强网络具有可变的输出特性,本文引入条件网络结构,以分位数损失函数中与听者感知相关的分位值作为条件信息的来源,利用条件信息对含噪语音特征进行线性调制,建立了一种可控的语音增强模型。最后,在公开数据集上通过语音增强实验分别对分位数损失函数的有效性和可控语音增强模型的性能进行了验证分析,实验结果表明:本文设计的分位数损失函数具备调节增强语音中的语音损失和噪声残留水平的能力;基于感知条件网络建立的可控语音增强模型,能够依据听者的主观听觉感受提供不同特性的增强语音输出,使听者获得更好的语音增强体验。

本文仅仅采用了一种比较简单的时频域语音增强网络来验证可控模型的有效性,下一步将采用更加复杂的时频域语音增强网络结构来进行实验验证,并将该模型扩展应用到时域语音增强中。另外,由于实验条件限制,本文对分位值的划分比较粗略,利用训练得到的网络只能实现对输出特性的粗略控制;下一步的研究将对分位值进行更加精细的划分,从而实现对增强语音中语音损失和噪声残留水平更加精确的控制。

参考文献

[1] 李吉祥, 倪旭昇, 颜上取, 等. 基于 A-DResUnet 的语音增强方法[J]. 电子测量与仪器学报, 2022, 36(10): 131-137.

LI J X, NI X SH, YAN SH Q, et al. Speech

enhancement method based on A-DResUnet[J]. Journal of Electronic Measurement and Instrumentation, 2022, 36(10): 131-137.

[2] PASCUAL S, BONAFONTE A, SERRÁ J. SEGAN: Speech enhancement generative adversarial network[C]. Proceedings of the 18th Annual Conference of the International Speech Communication Association, Stockholm: ISCA, 2017: 3642-3646.

[3] RETHAGE D, PONS J, SERRA X. A wavenet for speech denoising [C]. Proceedings of the 43rd International Conference on Acoustics, Speech, and Signal Processing, Alberta: IEEE, 2018: 5069-5073.

[4] QIAN K, ZHANG Y, CHANG S, et al. Speech enhancement using Bayesian wavenet[C]. Proceedings of the 18th Annual Conference of the International Speech Communication Association, Stockholm: ISCA, 2017: 2013-2017.

[5] PANDEY A, WANG D L. A new framework for CNN-based speech enhancement in the time domain [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 27(7): 1179-1188.

[6] PANDEY A, WANG D L. Dense CNN with self-attention for time-domain speech enhancement [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 1270-1279.

[7] DÉFOSSEZ, SYNNAEVE G, ADI Y. Real time speech enhancement in the waveform domain[C]. Proceedings of the 21st Annual Conference of the International Speech Communication Association, Shanghai: ISCA, 2020: 3291-3295.

[8] WANG K, HE B, ZHU W P. TSTNN: Two-stage transformer based neural network for speech enhancement in the time domain [C]. Proceedings of the 46th International Conference on Acoustics, Speech, and Signal Processing, Toronto: IEEE, 2021: 7098-7102.

[9] XU Y, DU J, DAI L, et al. An experimental study on speech enhancement based on deep neural networks[J]. IEEE Signal Processing Letters, 2014, 21(1): 65-68.

[10] XU Y, DU J, DAI L, et al. A regression approach to speech enhancement based on deep neural networks[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2015, 23(1): 7-19.

[11] HUANG P S, KIM M, HASEGAWA-JOHNSON M, et al. Joint optimization of masks and deep recurrent neural networks for monaural source separation [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2015, 23(12): 2136-2147.

[12] CHEN J, WANG D. Long short-term memory for speaker generalization in supervised speech separation [J]. Journal of the Acoustical Society of America, 2017, 141(6): 4705-4714.

[13] TAN K, CHEN J, WANG D. Gated residual networks with dilated convolutions for monaural speech enhancement[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 27 (1): 189-198.

[14] LI Y, LI X, DONG Y, et al. Densely connected network with time-frequency dilated convolution for speech enhancement [C]. Proceedings of the 44th International Conference on Acoustics, Speech, and Signal Processing, Brighton: IEEE, 2019: 6860-6864.

[15] WILLIAMSON D S, WANG Y, 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.

[16] PANDEY A, WANG D L. Exploring deep complex networks for complex spectrogram enhancement [C]. Proceedings of the 44th International Conference on Acoustics, Speech, and Signal Processing, Brighton: IEEE, 2019: 6885-6889.

[17] TAN K, WANG D L. Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 28: 380-390.

[18] HU Y, LIU Y, LYU S, et al. DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement [C]. Proceedings of the 21st Annual Conference of the International Speech Communication Association, Shanghai: ISCA, 2020: 2472-2476.

[19] LV S, HU Y, ZHANG S, et al. DCCRN+: Channel-wise subband dccrn with snr estimation for speech enhancement [C]. Proceedings of the 22nd Annual Conference of the International Speech Communication Association, Brno: ISCA, 2021: 2816-2820.

[20] WANG Z Q, WICHERN G, LE ROUX J. On the compensation between magnitude and phase in speech separation[J]. IEEE Signal Processing Letters, 2021, 28: 2018-2022.

[21] ZHANG L, WANG M, ZHANG Q, et al. PhaseDCN: A phase-enhanced dual-path dilated convolutional network for single-channel speech enhancement[J]. IEEE/ACM

Transactions on Audio, Speech, and Language Processing, 2021, 29: 2561-2574.

[22] LI A, LIU W, ZHENG C, et al. Two heads are better than one: A two-stage complex spectral mapping approach for monaural speech enhancement[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 1829-1843.

[23] NAIR A A, KOISHIDA K. Cascaded time + time-frequency unet for speech enhancement: Jointly addressing clipping, codec distortions, and gaps[C]. Proceedings of the 46th International Conference on Acoustics, Speech, and Signal Processing, Toronto: IEEE, 2021: 7153-7157.

[24] NAREDDULA S K R, GORTHI S, GORTHI R K S S. Fusion-net: Time-frequency information fusion Y-network for speech enhancement[C]. Proceedings of the 22nd Annual Conference of the International Speech Communication Association, Brno: ISCA, 2021: 3360-3364.

[25] ZHANG K, HE S, LI H, et al. DBNet: A dual-branch network architecture processing on spectrum and waveform for single-channel speech enhancement[C]. Proceedings of the 22nd Annual Conference of the International Speech Communication Association, Brno: ISCA, 2021: 2821-2825.

[26] REDDY C K A, DUBEY H, KOISHIDA K, et al. INTERSPEECH 2021 deep noise suppression challenge[C]. Proceedings of the 22nd Annual Conference of the International Speech Communication Association, Brno: ISCA, 2021: 2796-2800.

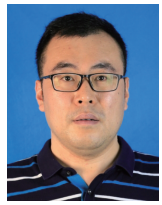
[27] DOSOVITSKIY A, DJOLONGA J. You only train once: Loss-conditional training of deep networks [C]. Proceedings of the 8th International Conference on Learning Representations, Virtual: ICLR, 2020.

[28] VALENTINI-BOTINHAO C, WANG X, TAKAKI S, et al. Speech enhancement for a noise-robust text-to-speech synthesis system using deep recurrent neural networks [C]. Proceedings of the 17th Annual Conference of the International Speech Communication Association, California: ISCA, 2016: 352-356.

[29] HU Y, LOIZOU P C. Evaluation of objective quality measures for speech enhancement [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2007, 16(1): 229-238.

[30] RIX A W, BEERENDS J G, HOLLIER M P, et al. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs [C]. Proceedings of the 26th International Conference on Acoustics, Speech, and Signal Processing, Utah: IEEE, 2001: 749-752.

作者简介



袁文浩(通信作者),2008 年于华东理工大学获得学士学位,2013 年于华东理工大学获得博士学位,现为山东理工大学副教授,主要研究方向为语音信号处理、深度学习。

E-mail: why_sdut@126.com

Yuan Wenhao (Corresponding author) received his B. Sc. degree from East China University of Science and Technology in 2008, and received his Ph. D. degree from East China University of Science and Technology in 2013. He is currently an associate professor at Shandong University of Technology. His main research interests include speech signal processing and deep learning.