A Real-Time Speech Enhancement Method Based on Complex Masks and Dilated Convolution Networks*

ZHU Ming[1*], *SUN Shiruo*[2]

(1.*School of Information Engineering, Yancheng Institute of Technology, Yancheng Jiangsu* 224051, *China*;
2.*School of Information Science and Engineering, Southeast University, Nanjing Jiangsu* 210096, *China*)

Abstract: Traditional speech enhancement usually operates on the amplitude spectrum of noisy speech and reconstructs the enhanced speech by using the phase of original noisy speech. In order to further improve the quality of speech enhancement in real environment, a complex dilated convolution network for real-time application is proposed. The results show that the proposed method can significantly improve the intelligibility and quality of speech while keeping the real-time performance. Compared with the baseline model RNNoise, PESQ and STOI are increased by 0.32 and 0.09 respectively.

Key words: speech enhancement; dilated convolutions; gated linear units; phase
EEACC: 6130 **doi**: 10.3969/j.issn.1005-9490.2021.03.019

基于复值掩蔽与扩张卷积的实时语音增强方法*

朱 明[1*], 孙世若[2]

(1.盐城工学院信息工程学院, 江苏 盐城 224051; 2.东南大学信息科学与工程学院, 江苏 南京 210096)

摘 要: 传统的语音增强算法通常是对含噪语音的幅值谱进行处理, 并利用原始含噪语音的相位对增强后的语音进行重构。为改善现有算法的语音质量和实时性, 本文提出一种基于复值掩蔽的扩张卷积网络对含噪语音进行实时增强处理。实验结果表明, 本文提出的方法在保证算法实时性的同时可以显著提高语音的可懂度与质量。相比基线模型 RNNoise, PESQ 和 STOI 分别提升了 0.32 和 0.09。

关键词: 语音增强; 扩张卷积; 线性门控单元; 相位
中图分类号: TN912.3 **文献标识码**: A **文章编号**: 1005-9490(2021)03-0612-04

现实中的语音信号总是会受到各种干扰与污染, 语音增强是从被噪声污染的语音信号中提取干净语音的技术, 目的是提高语音的可懂度与质量, 其在语音通信、自动语音识别系统前端等有着广泛的应用。近些年, 随着深度学习的发展, 更多前沿的有监督深度学习方法被引入到语音增强技术中。

然而, 深度学习的网络模型通常包含大量的矩阵运算, 这使得它必须依靠 GPU 进行推理计算。另一方面, 许多基于深度学习方法的语音增强模型都是非因果系统, 使得它们无法在实时系统当中应用。Valin[1] 提出了一种实时 DSP 和深度学习的混合方法, 使用循环神经网络估计理想频带增益, 可以满足实时因果系统的要求。然而, 这种方法只对语音的幅值谱进行了处理而忽略了相位的作用。受此启发, 本文提出了一种基于复值掩蔽与扩张卷积的语音增强方法对含噪语音进行实时增强处理。通过与基线模型对比来验证模型对语音增强性能的改善。

1 网络结构与优化

本文所提出的网络主要由编码器-解码器和门控扩张卷积模块构成, 如图 1 所示, 其中编码器包含 5 层卷积层, 解码器分为实部和虚部解码器, 分别包含 5 层反卷积层。编码器和解码器中间包含 5 层门控扩张卷积。

项目来源: 国家自然科学基金项目(61673108)
收稿日期: 2020-10-03 **修改日期**: 2020-11-03

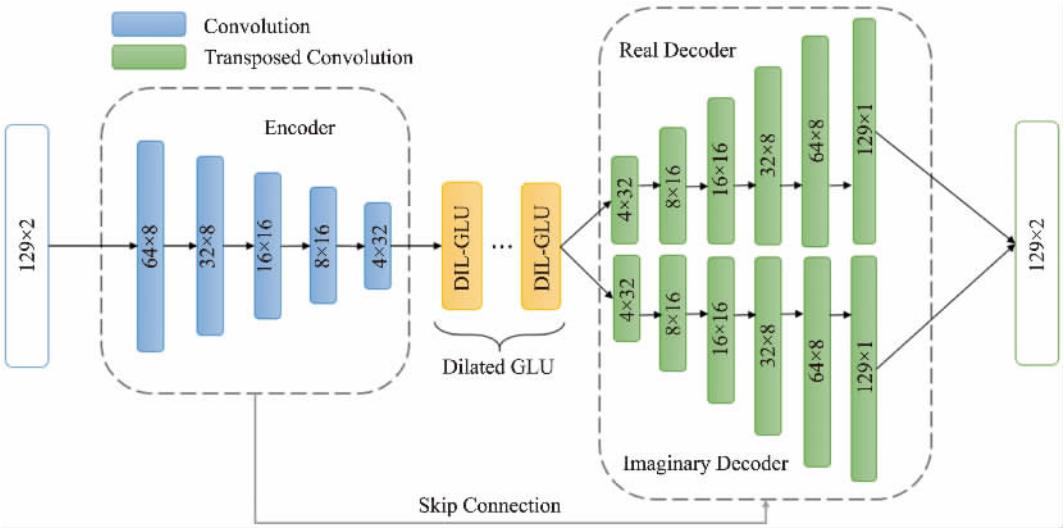

图 1　网络结构

1.1　编码器-解码器

编码器-解码器网络只包含卷积和反卷积操作，并且为了满足因果系统的要求，所有卷积核的尺寸设置为 3×1，即频率维度对应的卷积核尺寸为 3，时间维度对应的卷积核尺寸为 1。这样在保证帧与帧之间独立的同时又可以提取到相邻频点之间的特征信息。

编码器包含 5 层卷积层，输入特征是语音短时傅里叶频谱的实部与虚部。解码器网络采用双流结构，采用 2 个相同的解码器结构分别估计增强语音的实部和虚部。解码器同样包含 5 层与编码器镜像的反卷积层，从而保证输出特征与输入特征的维度一致。所有卷积-反卷积层均采用 ReLU 激活函数，解码器最后一层反卷积采用线性激活以保证输出复值掩蔽取值范围是无界的。

1.2　门控扩张卷积网络

门控扩张卷积网络是连接编码器和解码器的中间级，目的是充分利用过去语音的时频信息，提高网络的性能。它将线性门控单元(GLU) 中的卷积替换为时域维度上的因果扩张卷积并把多个线性门控单元堆叠(共 5 层) ，通过这种方法可以让网络的感知视野达到 64 帧。下面分别介绍因果扩张卷积和线性门控单元。

传统的卷积神经网络为了增加网络的感知视野，通常采用增加网络的深度或者扩大卷积核的尺寸，但这往往导致梯度消失或过大的计算量，导致整体网络性能的下降。为了有效地解决这一问题，Yu 等[2] 提出了扩张卷积的概念，其最大的特点是网络的感知视野随网络深度指数增加。如图 2 所示，扩张卷积通过在卷积核每个元素之间填充 0 元素来增加感知视野大小。因果关系是通过限制卷积核只在过去的样本上扩张实现的。

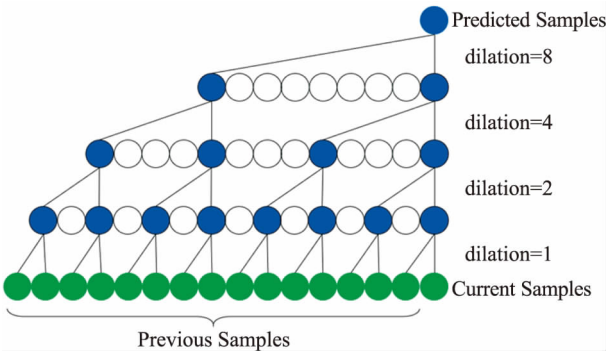

图 2　一维因果扩张卷积(卷积核大小为 2)

门控机制[3] 最早是用于缓解 RNN 在传播过程中的梯度消失问题，从而有效保留时间序列的信息。Dauphin 等[4] 改进了 LSTM 中的门控机制并用于卷积神经网络[5] 中，称为线性门控单元，其基本结构如图 3 所示。

$$gate = \mathrm{conv}_1(x) \odot \sigma(\mathrm{conv}_2(x)) \tag{1}$$

式(1) 表示门控机制的输出，其中 conv(x) 表示

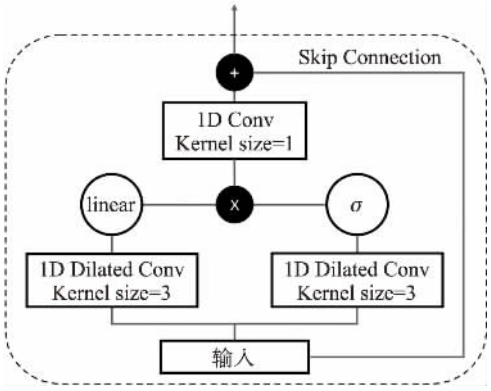

图 3　线性门控单元

卷积,tanh 表示 tanh 激活函数,σ 表示 sigmoid 激活函数,⊙表示元素相乘。线性门控单元输出的梯度表示为:

$$\nabla gate = \nabla \mathrm{conv}_1(x) \odot \sigma(\mathrm{conv}_2(x)) + \mathrm{conv}_1(x) \odot \sigma'(\mathrm{conv}_2(x)) \nabla \mathrm{conv}_2(x) \quad (2)$$

通常,梯度消失是由于反向传播过程中存在 tanh 和 sigmoid 激活函数的导数,而上述梯度公式中的第一项 $\nabla \mathrm{conv}_1(x) \odot \sigma(\mathrm{conv}_2(x))$ 不包含激活函数的导数,因而可以将梯度流在层与层之间有效传递。

1.3 网络的训练目标和优化

传统的理想比率掩模[6](IRM)是在幅值域中定义的,而本文采用复值理想比率掩模[7](cIRM),它可以有效利用频域中的相位信息,更加有效地重构语音波形。cIRM 同时利用了复数频谱的实部和虚部对语音波形进行重构,便于网络学习语音的相位信息,具体计算方法如式(3):

$$M = M_r + iM_i = \frac{Y_r S_r + Y_i S_i}{Y_r^2 + Y_i^2} + i\frac{Y_r S_i + Y_i S_r}{Y_r^2 + Y_i^2} \quad (3)$$

式中:Y 表示含噪语音的频谱,S 表示干净语音的频谱,下标 r 表示实部,i 表示虚部。解码器的双流结构分别预测 cIRM 的实部 M_r 和虚部 M_i。为了进一步提高语音的可懂度,网络还联合优化了 PASE[8] 特征损失。PASE 是 Koyama 等提出的一种基于 DNN 的语音编码器,利用 PASE 提取增强语音和干净语音的特征并计算两者的均方误差(MSE)作为联合优化损失项。则对于本文所述的网络损失函数为:

$$L_{\mathrm{joint}} = L_{\mathrm{cIRM}} + L_{\mathrm{PASE}} = \frac{1}{N}\sum_{n=0}^{N-1} \| \hat{M}_n - M_n \|^2 + \frac{1}{N}\sum_{n=0}^{N-1} \| \mathrm{pase}(\hat{y}_n) - \mathrm{pase}(y_n) \|^2 \quad (4)$$

式中:$\hat{y}_n$ 和 y_n 分别代表估计的增强语音和干净语音,N 表示网络训练时一个批次的大小。训练和测试均在 16 kHz 采样率下进行,分帧长度为 256 点(16 ms),帧移为 160 点(10 ms),窗函数选择汉明窗,并计算 256 点 FFT 得到 129×2 维复数频谱作为输入特征,其中包含 129 维实部特征和 129 维虚部特征。利用网络输出的 cIRM 和含噪语音,通过下式重构增强语音谱。

$$s = \mathrm{ifft}[S_r + iS_i] = \mathrm{ifft}[(M_r X_r - M_i X_i) + i(M_r X_i + M_i X_r)] \quad (5)$$

式中:s 代表语音的时域波形,S 表示语音的频谱,X 表示含噪语音频谱。

2 实验与结果

2.1 实验设置

实验在公开数据集 Voice Bank[9] 和 DEMAND[10] 上进行,其中 Voice Bank 为干净语音数据集,训练集包含 28 个不同的说话人,测试集包含 2 个不同的说话人。DEMAND 为噪声数据集,选取其中 10 种不同类型的噪声进行合成,训练集的合成信噪比为 0 dB、5 dB、10 dB、15 dB,测试集的合成信噪比为 2.5 dB、7.5 dB、12.5 dB、17.5dB,所有音频的采样率均为 16 kHz,总共包含 11 572 条训练语音和 824 条测试语音。模型学习率设置为 0.000 1,一个批次的大小为 64,总共训练 30 轮。实验选择 PESQ、STOI 指标来评估本文所述方法的性能并计算模型的参数量和计算复杂度来评估模型的实时性。本文选择了 RNNoise[1]、NSNet2[13] 作为基线模型进行对比。

2.2 性能对比与分析

实验结果见表 1。

表 1 模型性能对比

方法	PESQ	STOI	参数量	计算复杂度(FLOPs)
Noisy	1.97	0.921	—	—
RNNoise	2.29	0.923	0.61M	**17.5M**
NSNet2	2.27	0.905	2.69M	53.6M
Proposed	**2.61**	**0.932**	**0.09M**	21.7M

粗体表示不同指标下的最优结果。由表 1 可以看出,本文所述模型在两项常用语音可懂度与质量的评价准则下均实现了最优结果。相比原始含噪语音,本文所述模型使平均 PESQ 和 STOI 分别提升了 0.64 和 0.011,远高于 RNNoise 和 NSNet2,这说明有效地估计语音的相位信息可以提升语音重构的质量。

另一方面,得益于全卷积操作,模型参数量只有 RNNoise 和 NSNet2 的 15%和 3%。而模型的计算复杂度也接近 RNNoise,且远低于 NSNet2。在实际测试中,利用 CPU(Intel i7-8700)推理一帧(10 ms)语音所需要的时间为 3.8ms,这说明模型在满足实时性要求的情况下,可以有效提高语音质量与可懂度。

3 总结与展望

本文研究了一种基于复值掩蔽与扩张卷积的实时语音增强方法,模型在保证较小参数量的基础上,提高了算法的实时性能。此外,通过估计 cIRM,模型充分利用语音的相位信息,提高了语音重构的质量。

为了进一步改善实时系统中语音增强的性能,可以采用更加符合人耳听觉感知的损失函数作为优

化目标,由于损失函数的优化只会增加模型训练过程的计算复杂度,因而不会影响实时系统中推理计算的复杂度。除此之外,还可以使用更大规模的噪声数据集来训练模型,使模型在真实噪声环境中有更好的泛化性能。

参考文献:

[1] Valin J M. A Hybrid DSP/Deep Learning Approach to Real-Time Full-Band Speech Enhancement [C]//2018 IEEE 20th International Workshop on Multimedia Signal Processing(MMSP) . Vancouver,Canada: IEEE,2018: 1-5.

[2] Yu,Fisher,Koltun V. Multi-Scale Context Aggregation by Dilated Convolutions[C]//International Conference on Learning Representations(ICLR) . San Juan,Puerto Rico,2016. arXiv: 1511.07122.

[3] Hochreiter S,Schmidhuber J. Long Short-Term Memory[J]. Neural Computation,1997,9(8) : 1735-1780.

[4] Dauphin Y N,Fan A,Auli M,et al. Language Modeling with Gated Convolutional Networks[C]//Proceedings of the 34th International Conference on Machine Learning,Mountain View,California,USA, CoRL 2017: 933-941.

[5] Tan K,Chen J,Wang D L. Gated Residual Networks with Dilated Convolutions for Monaural Speech Enhancement[J]. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP) ,2018,27: 189-198.

[6] Wang Y,Narayanan A,Wang D L. On Training Targets for Supervised Speech Separation[J]. IEEE/ACM Transactions on Audio Speech & Language Processing,2014: 1849-1858.

[7] Williamson D S,Wang Y,Wang D L. Complex Ratio Masking for Monaural Speech Separation[J]. IEEE/ACM Transactions on Audio Speech & Language Processing,2016,24(3) : 483-492.

[8] Koyama Y,Vuong T,Uhlich S,et al. Exploring the Best Loss Function for DNN-Based Low-Latency Speech Enhancement with Temporal Convolutional Networks[J]. arXiv preprint arXiv: 2005.11611,2020.

[9] Veaux C,Yamagishi J,King S. The Voice Bank Corpus: Design, Collection and Data Analysis of a Large Regional Accent Speech Database[C]//Oriental Cocosda Held Jointly with Conference on Asian Spoken Language Research & Evaluation. Cape Panwa Hotel,Phuket,Thailand: IEEE,2014: 1-4.

[10] Thiemann J,Ito N,Vincent E. The Diverse Environments Multi-Channel Acoustic Noise Database: A Database of Multichannel Environmental Noise Recordings[J]. Journal of the Acoustical Society of America,2013,133: 5591.

[11] Rix A, Beerends J, Hollier M, et al. Perceptual Evaluation of Speech Quality(PESQ) —A New Method for Speech Quality Assessment of Telephone Networks and Codecs[C]//Proceedings of ICASSP,Salt Lake City,Utah,USA: 2001: 749-752.

[12] Taal C H, Hendriks R C, Heusdens R, et al. A Short-Time Objective Intelligibility Measure for Time-Frequency Weighted Noisy Speech[C]//IEEE International Conference on Acoustics Speech & Signal Processing. Dallas, Texas, USA: IEEE, 2010: 4214-4217.

[13] Sebastian B,Ivan T. Data Augmentation and Loss Normalization for Deep Noise Suppression[M]//22nd International Conference on Speech and Computer(SPECOM) ,St. Petersburg,Russia: 2020: 79-86.

朱　明(1971—) ,男,江苏盐城人,讲师,主要研究领域为语音识别和图像信号处理等,zhum@ ycit.cn;

孙世若(1996—) ,男,江苏南京人,东南大学信息科学与工程学院硕士研究生,主要研究领域为基于深度学习的语音信号处理,220190786@ seu.edu.cn。