计算机工程
Computer Engineering
ISSN 1000-3428,CN 31-1289/TP

《计算机工程》网络首发论文

题目：一种基于时频域特征融合的语音增强方法
作者：袁文浩，时云龙，胡少东，娄迎曦
DOI：10.19678/j.issn.1000-3428.0059354
网络首发日期：2020-10-23
引用格式：袁文浩，时云龙，胡少东，娄迎曦．一种基于时频域特征融合的语音增强方法．计算机工程. https://doi.org/10.19678/j.issn.1000-3428.0059354

网络首发时间：2020-10-23 15:42:51
网络首发地址：https://kns.cnki.net/kcms/detail/31.1289.tp.20201023.1428.003.html

一种基于时频域特征融合的语音增强方法

袁文浩* 时云龙 胡少东 娄迎曦
（山东理工大学计算机科学与技术学院 淄博 255000）

摘要：为了充分利用含噪语音特征来提高基于深度神经网络的语音增强方法的性能，提出了一种基于时频域特征融合的语音增强方法。首先以含噪语音的波形作为训练特征，以纯净语音的对数功率谱作为训练目标，设计了一种能够建模含噪语音时域特征和纯净语音频域特征之间映射关系的语音增强网络；然后在此基础上，将含噪语音的波形和对数功率谱一起作为训练特征，构建了一种能够充分融合含噪语音时域和频域特征的语音增强网络。实验结果表明，相比单纯使用频域特征的方法，该方法能够显著提升增强语音的语音质量和可懂度，具有更好的语音增强性能。

关键词：语音增强；深度神经网络；特征融合；时域特征；频域特征
开放科学标识符（OSID）：

A Speech Enhancement Approach Based on Fusion of Time-domain and Frequency-domain Features

YUAN Wen-hao*, SHI Yun-long, HU Shao-dong, LOU Ying-xi
(College of Computer Science and Technology, Shandong University of Technology, Zibo 255000, China)

Abstract: To improve the performance of speech enhancement methods based on deep neural network by making full use of noisy speech features, a speech enhancement method based on the fusion of time-domain and frequency-domain features is proposed. First, by using the waveform of noisy speech and the log power spectrum of clean speech as the training feature and the training target respectively, a speech enhancement network that can model the mapping between the time-domain features of noisy speech and the frequency-domain features of clean speech is designed. Then, on this basis, the waveform and log power spectrum of noisy speech are used as training features together, and a speech enhancement network that can fully integrate the time-domain and frequency-domain features of noisy speech is constructed. Experimental results show that, compared to the methods using frequency-domain features alone, the proposed method can significantly improve the speech quality and intelligibility of enhanced speech, and has better speech enhancement performance.

Key words: Speech enhancement; Deep neural network; Feature fusion; Time-domain features; Frequency-domain features
DOI: 10.19678/j.issn.1000-3428.0059354

0 概述

基于深度神经网络的语音增强方法相比传统的统计方法显著提高了非平稳噪声条件下的语音增强性能，近年来成为语音增强领域的研究热点[1-3]。为了提高网络的语音增强性能，现有的研究工作主要针对训练特征和训练目标的设计以及网络结构的改进展开。根据训练特征和训练目标的设计方法，基于深度神经网络的语音增强方法可以分为频域和时域两类。

频域的语音增强一般采用含噪语音经过短时傅里叶变换得到的幅度谱或对数功率谱作为训练特征，而训练目标除了纯净语音的幅度谱或对数功率谱，还可以是由幅度谱计算得到的掩蔽特征。文献[4-5]采用全

基金项目：国家自然科学基金(61701286)
作者简介：袁文浩(1985－)，男，副教授、博士，主要研究方向为语音增强、语音信号处理，本文通信作者；时云龙，硕士研究生；胡少东，硕士研究生；娄迎曦，硕士研究生。**E-mail：**why_sdut@126.com

连接神经网络建立了一个从含噪语音对数功率谱到纯净语音对数功率谱的映射关系。基于语音在时间的序列性，文献[6]和文献[7]分别采用循环神经网络(Recurrent Neural Network，RNN)和长短时记忆(Long Short-Term Memory，LSTM)网络来估计含噪语音的掩蔽特征。而当采用多帧的含噪语音幅度谱或对数功率谱作为训练特征时，网络输入将在时间和频率两个维度上都具有相关性，因此文献[8-12]采用卷积神经网络(Convolutional Neural Network，CNN)来进行语音增强，其中：文献[8]采用一个全卷积神经网(Fully Convolutional Network，FCN)络构来建立含噪语音幅度谱和纯净语音幅度谱之间的映射关系；文献[9]的网络结构则结合了卷积层、池化层和全连接层三种不同的网络层；为了提高网络在时间和频率两个维度上的感受野，文献[10-12]将空洞卷积引入到语音增强网络中，其中文献[10-11]还在网络结构中引入了门控机制和残差学习，而文献[12]中的网络结构则是基于密集连接卷积网络进行设计的。得益于卷积神经网络中的参数共享机制，基于卷积神经网络的语音增强方法能够大大减少网络中需要训练的参数量。通过结合卷积神经网络和循环神经网络两种不同的网络结构，文献[13]提出了一种用于语音增强的卷积循环网络(Convolutional Recurrent Network，CRN)，文献[14]将两层的 LSTM 嵌入到一个全卷积的编码器-解码器(Convolutional Encoder-Decoder，CED)中，提出了另外一种形式的 CRN；实验结果表明 CRN 相比 LSTM 等循环神经网络进一步提高了语音增强性能。

时域的语音增强则采用含噪语音和纯净语音的时域波形分别作为训练特征和训练目标。文献[15]采用 FCN 建立了一个含噪语音帧波形到纯净语音帧波形的映射关系，并指出采用全连接层的 DNN 不适用于时域语音增强；在此基础上，文献[16]采用一个基于 STOI 的损失函数代替均方误差函数来进行网络的训练，进一步提高了增强语音的可懂度。文献[17]将因果卷积嵌入到一个由全卷积网络构成的编码器-解码器中，构建了一个实时的时域语音增强网络。文献[18-19]对时域语音增强中不同损失函数的性能进行了评估，指出：对于时域语音增强，先将时域信号转换到频域然后在频域设计损失函数相比直接在时域设计损失函数具有更好的性能，并基于此提出了一种新的时域语音增强网络的设计框架。

在频域的语音增强中，通常使用含噪语音和纯净语音的对数功率谱作为训练特征和训练目标，对数功率谱的计算是基于信号的短时傅里叶变换，其对于含噪语音特征的表达具有一定的局限性。时域语音增强直接使用含噪语音和纯净语音的波形作为训练特征和训练目标，虽然相比频域语音增强能够更好的利用含噪语音特征，但是其性能非常依赖于损失函数的设计，而设计复杂的损失函数会大大提高网络的训练难度。基于以上分析，为了设计一种易于训练且能够充分利用含噪语音的特征来提高语音增强性能的网络结构，本文采用含噪语音的时域波形作为训练特征，同时采用纯净语音的频域对数功率谱作为训练目标，依靠深度神经网络强大的特征计算能力来建立含噪语音时域波形和纯净语音频域对数功率谱之间的映射关系，并进一步通过进行网络结构的设计将含噪语音的时域特征与频域特征在网络的深层中进行融合，提出了一种基于时频域特征融合的深度神经网络语音增强方法，最后通过语音增强实验从增强语音的语音质量和可懂度两方面对本文所提语音增强方法的性能进行了客观评估。

1 基于深度神经网络的语音增强模型

在基于深度神经网络的语音增强方法中，为了训练语音增强网络，将语音增强问题采用回归模型进行解决。通过网络的训练学习构造一个非线性映射函数 f_θ 来表达含噪语音帧和增强语音帧之间的回归关系，其中 θ 是网络的参数集合。网络的训练采用均方误差损失函数，

$$L(\theta)=\frac{1}{M}\sum_{l=1}^{M}\left\| f_\theta\left(X_l\right)-T_l\right\|_2^2 \tag{1}$$

其中，X_l 是网络输入的训练特征，T_l 是网络的训练目标，M 是网络训练采用的 Mini-batch 的大小。根据 X_l 和 T_l 在时域或频域的不同设计方法，该模型可以作为频域的语音增强模型，也可以作为时域的语音增强模

型。需要注意的是，为了减小网络的训练参数规模，并保证语音增强模型的因果性，本文中的 X_l 和 T_l 均为单帧的频域或时域特征。

1.1 频域语音增强模型

对于频域的语音增强，在网络训练阶段，通常采用的训练特征和训练目标分别为含噪语音和纯净语音的对数功率谱，

$$X_l=\left[\log_e\left(|Y_{l,1}|^2\right),\log_e\left(|Y_{l,2}|^2\right),\cdots,\log_e\left(|Y_{l,k}|^2\right),\cdots,\log_e\left(|Y_{l,K}|^2\right)\right] \tag{2}$$

$$T_l=\left[\log_e\left(|S_{l,1}|^2\right),\log_e\left(|S_{l,2}|^2\right),\cdots,\log_e\left(|S_{l,k}|^2\right),\cdots,\log_e\left(|S_{l,K}|^2\right)\right] \tag{3}$$

其中，$|Y_{l,k}|^2$ 和 $|S_{l,k}|^2$ 分别是含噪语音和纯净语音通过短时傅里叶变换计算得到的第 l 帧的第 k 个频带的功率。

在利用训练得到的网络进行语音增强时，首先根据含噪语音第 l 帧的对数功率谱计算增强语音第 l 帧的对数功率谱，

$$\hat{T}_l=f_\theta\left(X_l\right) \tag{4}$$

然后结合含噪语音第 l 帧的相位谱 α_l 进行短时傅里叶逆变换(ISTFT)，得到增强语音第 l 帧的时域信号

$$\hat{s}_l=\mathrm{ISTFT}(\exp(\hat{T}_l/2)\bullet\exp(j\alpha_l)) \tag{5}$$

1.2 时域语音增强模型

对于时域的语音增强，在网络训练阶段，采用的训练特征和训练目标分别为含噪语音和纯净语音的波形，

$$X_l=\left[y_{l,1},y_{l,2},\cdots,y_{l,n},\cdots,y_{l,N}\right] \tag{6}$$

$$T_l=\left[s_{l,1},s_{l,2},\cdots,s_{l,n},\cdots,s_{l,N}\right] \tag{7}$$

其中，$y_{l,n}$ 和 $s_{l,n}$ 分别是含噪语音和纯净语音波形经过分帧后得到的第 l 帧的第 n 个采样点。

在利用训练得到的网络进行语音增强时，可以直接根据含噪语音第 l 帧的波形计算增强语音第 l 帧的时域信号，

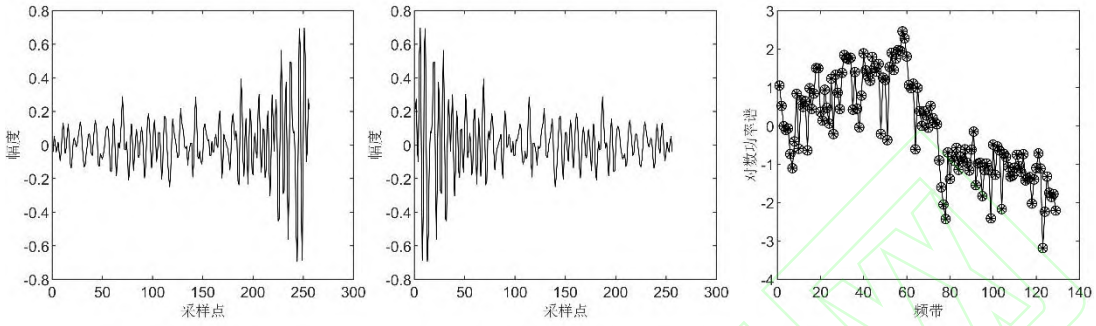
$$\hat{s}_l=\hat{T}_l=f_\theta\left(X_l\right) \tag{8}$$

2 基于时频域特征融合的语音增强模型

2.1 时频域特征融合语音增强模型

在频域的深度神经网络语音增强方法中，如式(2)所示，通常采用含噪语音的对数功率谱作为网络的输入，而对数功率谱对于含噪语音特征的表达是具有局限性的，这种局限性主要表现在两个方面。一方面，对数功率谱忽略了信号的相位信息，例如，图 1(a)给出了 256 点的采样频率为 8kHz 的噪声信号的波形；图 1(b)给出了另外一段噪声信号的波形，该噪声段为图 1(a)中噪声信号的倒序排列；图 1(c)则给出了上述两种噪声信号的对数功率谱，可见：虽然图 1(a)和 1(b)中的噪声具有截然不同的变化趋势，但是两种噪声却具有完全相同的对数功率谱特征，这表明由于忽略了相位信息，对数功率谱不能完整表达含噪语音的特征。另一方面，式(2)中对数功率谱的特征维度 K 受限于短时傅里叶变换窗长的选择，而窗长的选择受海森堡不确

定性原理限制，难以同时满足时间分辨率和频率分辨率的需求。在对数功率谱的计算中窗长一般是按照语音信号的短时平稳特性进行选择，因此对数功率谱能够较好的表达语音信号的短时特性；而含噪语音信号是由语音信号和噪声信号叠加得到，因为噪声信号的来源不同，不同噪声信号的特性复杂多样，为了更好的提取含噪语音的短时特征，理论上在进行短时傅里叶变换时应该选用不同的窗长，因此采用相同窗长计算得到的对数功率谱并不能很好的表达含噪语音的短时特性。例如，对于采样频率为 8kHz 的信号，通常采用的窗长为 256 点[4-5]，但是对于图 1(a)中的噪声，256 点的窗长显然不能反映该噪声随时间的快速变化趋势，这表明采用固定窗长的对数功率谱不能完整表达含噪语音中不同类型噪声的变化特性。



(a) 噪声信号 1 的波形　　(b) 噪声信号 2 的波形　　(c) 两种噪声的对数功率谱

图 1 两种噪声信号的波形与对数功率谱

Figure 1 Waveform and log power spectrum of two noises

在时域的深度神经网络语音增强方法中，直接采用式(7)中的纯净语音波形作为训练目标，由于波形特征中的采样点具有快速变化的特性，使用式(1)中的均方误差损失函数进行网络的训练，并不能得到相比频域方法更好的语音增强性能，因此在时域的语音增强方法需要充分考虑损失函数的设计问题，网络结构通常比较复杂且难于训练。

为了充分利用含噪语音特征来提高深度神经网络的语音增强性能，且同时避免复杂损失函数的设计难题，保证网络易于训练，本文将含噪语音的频域对数功率谱和时域波形共同作为训练特征，同时采用纯净语音的频域对数功率谱作为训练目标，设计了一种时频域特征融合的语音增强网络结构。因为含噪语音的时域特征与频域特征具有较大的差别，直接组合两种特征作为网络的输入是不合适的。为了在网络中深度融合含噪语音的时域和频特征来进行语音增强，本文基于时频域特征融合的语音增强网络结构包括三个模块，分别为：以含噪语音时域波形特征作为输入的时域特征计算模块，以含噪语音对数功率谱特征作为输入的频域特征计算模块，结合两个特征计算模块的输出作为输入的语音增强模块，三个模块对应的参数集合分别为θ_t、θ_f和θ_{tf}，整个网络的参数集合可以表示为

$$\theta=\left\{\theta_t,\theta_f,\theta_{tf}\right\} \tag{9}$$

网络的训练目标与式(3)相同，网络的训练特征为

$$X_l=\left[X_l^t,X_l^f\right] \tag{10}$$

其中X_l^f与式(2)一致，是频域的对数功率谱特征

$$X_l^f=\left[\log_e\left(\left|Y_{l,1}\right|^2\right),\log_e\left(\left|Y_{l,2}\right|^2\right),\cdots,\log_e\left(\left|Y_{l,k}\right|^2\right),\cdots,\log_e\left(\left|Y_{l,K}\right|^2\right)\right] \tag{11}$$

X_l^t与式(6)一致，是时域的波形特征

$$X_l^t = \left[y_{l,1}, y_{l,2}, \cdots, y_{l,n}, \cdots, y_{l,N} \right] \tag{12}$$

在利用训练得到的网络进行语音增强时，首先根据含噪语音第 l 帧的波形和对数功率谱计算增强语音第 l 帧的对数功率谱，

$$\hat{T}_l = f_\theta(X_l) = f_{\theta_f}\left(\left[f_{\theta_t}\left(X_l^t\right), f_{\theta_f}\left(X_l^f\right) \right] \right) \tag{13}$$

然后按照式(5)同样的计算方法得到增强语音第 l 帧的时域信号 $\hat{s}_l$。

2.2 网络结构

本文的时频域特征融合网络基于卷积循环神经网络结构进行设计，卷积循环神经网络结合了卷积神经网络的特征提取能力和循环神经网络对于长期依赖信息的建模能力，在语音增强中表现出了良好的性能[13-14]。时频域特征融合网络中的时域特征计算模块和频域特征计算模块均为多层的卷积网络结构，两者具有类似的结构，唯一的区别是，由于时域特征和频域特征的维度不同，最后一个卷积层所用的卷积滤波器的大小不同，语音增强模块则由两层的 LSTM 构成，该网络记为 TF-CRN，网络的具体结构如图 2 所示。

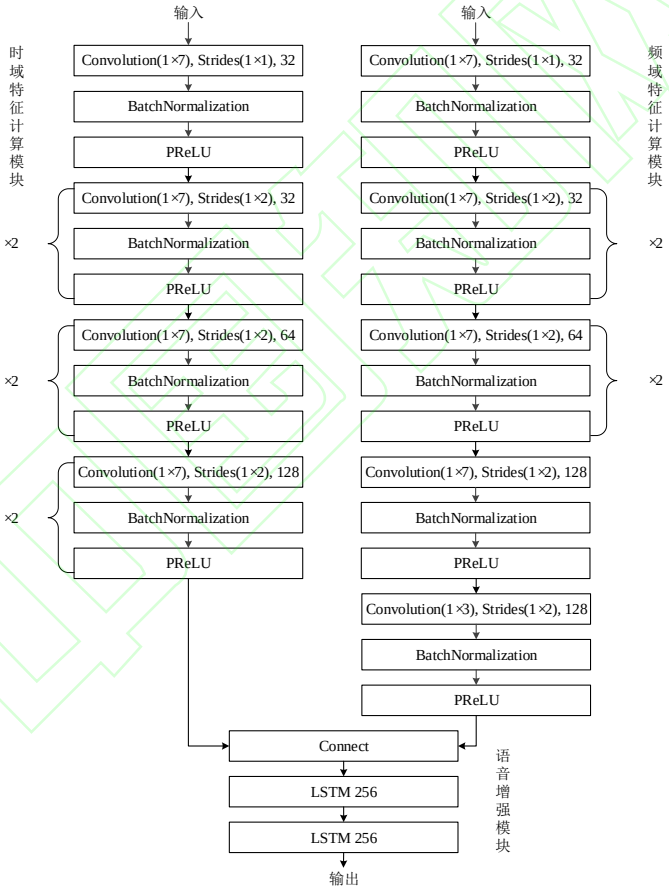


图 2 时频域特征融合网络结构图

Figure 2 Structure diagram of time and frequency feature fusion network

另外，如果只保留时域特征计算模块和语音增强模块，训练得到的网络可以建立含噪语音时域波形和纯净语音频域对数功率谱之间的映射关系，该网络记为 T-CRN；而如果只保留频域特征计算模块和语音增强模块，训练得到的网络可以建立含噪语音对数功率谱和纯净语音对数功率谱之间的映射关系，是一种频域的语音增强方法，该网络记为 F-CRN。

3 实验与结果分析

3.1 训练集与测试集

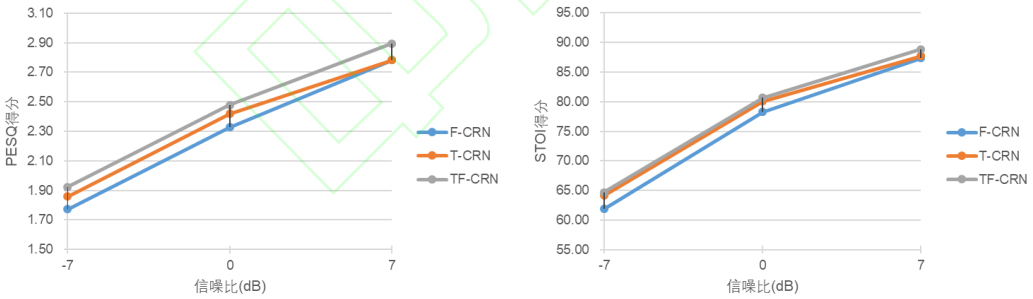
训练集和测试集均基于 TIMIT 语音库构造，其中训练集中的纯净语音来自 TIMIT 语音库的训练集，测试集中的纯净语音来自 TIMIT 语音库的测试集[20]，纯净语音的采样频率均转换为 8kHz。为了合成训练集中的含噪语音，选取文献[21]中的 100 个真实噪声段，将其采样频率同样转换为 8kHz，然后按照-10dB、-5dB、0dB、5dB 和 10dB 五种信噪比与纯净语音进行合成，从所有合成得到的含噪语音中随机选取 50000 段，与其相应的纯净语音一起构成训练集。为了合成测试集中的含噪语音，选取 Noisex92 噪声库中 Factory2、Buccaneer1、Destroyer engine 和 HF channel 噪声 [22]，将其采样频率转换为 8kHz，按照-7dB、0dB 和 7dB 三种信噪比与 192 段纯净语音进行合成，选取全部 2304(192×3×4)段含噪语音，与相应的纯净语音一起构成测试集。需要注意的是，为了检验语音增强方法对不同噪声条件的泛化能力，测试集选取的 4 类噪声是与训练集完全不同的未知噪声，测试集中的-7dB 和 7dB 是不同于训练集的未知信噪比。

频域对数功率谱特征计算所用的短时傅里叶变换的帧长为$32ms$(256 点)，帧移为$16ms$(128 点)，相应的频域特征维度为 129；时域波形按照语音段均标准化到-1 到 1 之间，分帧的帧长同样为 256 点，帧移为 128 点，相应的时域特征维度为 256。

3.2 语音增强性能比较

为了客观评价不同网络的语音增强性能，分别采用不同网络对测试集含噪语音进行语音增强，并比较不同网络增强后语音的平均语音质量和平均可懂度。其中，语音质量的评价指标为 PESQ(Perceptual Evaluation of Speech Quality，PESQ)，其得分范围为-0.5 到 4.5，得分越高代表语音质量越好[23]；语音可懂度的评价指标为(Short Time Objective Intelligibility，STOI)[20]，其得分范围为 0 到 1，得分越高代表语音可懂度越好[24]。

首先对 F-CRN、T-CRN 和 TF-CRN 三种网络的语音增强性能进行比较，图 3 给出了与三种网络相应的不同信噪比下增强语音的平均 PESQ 得分和平均 STOI 得分，其中平均 STOI 得分以百分比的形式进行表示。可见，三种网络中，TF-CRN 在两种指标的不同信噪比下都取得了最好的结果，表明 TF-CRN 能够充分融合时域和频域特征，相比单纯采用频域或时域特征的网络提高了语音增强性能；在两种指标的低信噪比(-7dB 和 0dB)条件下，T-CRN 相比 F-CRN 取得了更好的结果，表明在相同的 CRN 网络结构下，采用时域波形特征能够带来更好的语音增强性能。



(a) 平均 PESQ 得分　　(b) 平均 STOI 得分

图 3 不同信噪比下 F-CRN、T-CRN 和 TF-CRN 的语音增强性能比较

Figure 3 Comparison of speech enhancement performance of F-CRN, T-CRN and TF-CRN under different SNRs

为了验证基于时频域特征融合的语音增强方法的有效性，将其与其它四种网络进行语音增强性能的比较，作为对比的第一个网络是与文献[4]相同的具有 3 层结构的 DNN，每个隐层的节点个数为 2048；第二个网络则采用 3 层 LSTM，每层的 Cell 维度为 512；第三个网络则基于文献[14]设计；第四个网络是与文献[15]一致的 FCN。前三个网络均为频域的语音增强，均采用单帧的含噪语音和纯净语音的对数功率谱分别作为训练特征和训练目标，分别记为 F-DNN、F-LSTM 和 CED-CRN；第四个网络则为时域的语音增强，采用单

帧的含噪语音和纯净语音的时域波形分别作为训练特征和训练目标，记为 T-FCN。

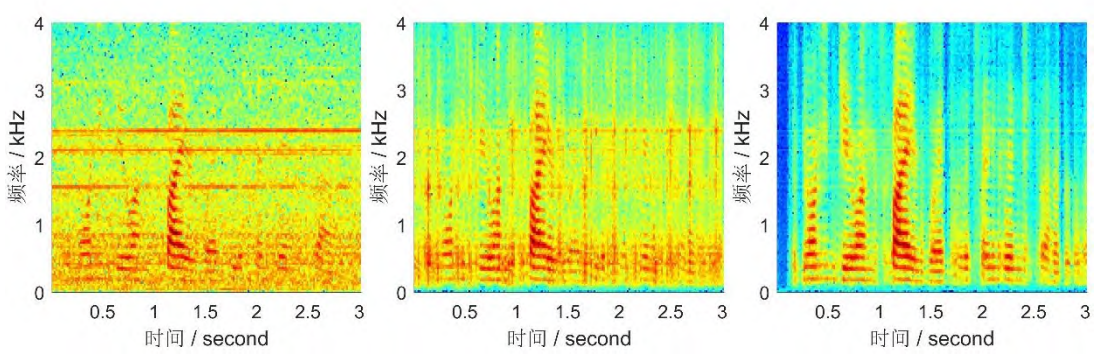
表 1 给出了与五种网络相应的不同信噪比下增强语音的平均 PESQ 得分和平均 STOI 得分，其中平均 STOI 得分以百分比的形式进行表示。可见，在不同信噪比下，相比含噪语音，F-DNN 增强后语音的平均 PESQ 得分提升非常有限，平均 STOI 得分甚至出现了下降，表明当采用单一帧的含噪语音作为输入时，DNN 由于缺乏对时间依赖信息建模的能力，不能进行有效的语音增强；其它频域方法包括 F-LSTM 和 CED-CRN 增强后语音的平均 PESQ 得分和平均 STOI 得分相比含噪语音均有明显提升，表明它们能够进行有效的语音增强，结合平均 PESQ 得分和平均 STOI 得分的结果整体来看，CED-CRN 相比 F-LSTM 具有更好的性能；时域方法 T-FCN 增强后语音的平均 PESQ 得分和平均 STOI 得分相比含噪语音在绝大多数条件下都有所提升，但是提升幅度相比频域的 F-LSTM 和 CED-CRN 有明显差距，表明与这两种频域方法相比，T-FCN 的语音增强性能较差；在五种网络中，除了 7dB 下的平均 STOI 得分，TF-CRN 在其它噪声条件下的两种指标都取得了最好的结果，表明 TF-CRN 通过融合时域和频域特征，具备了最好的语音增强性能。

表 1 不同信噪比下不同网络的语音增强性能比较

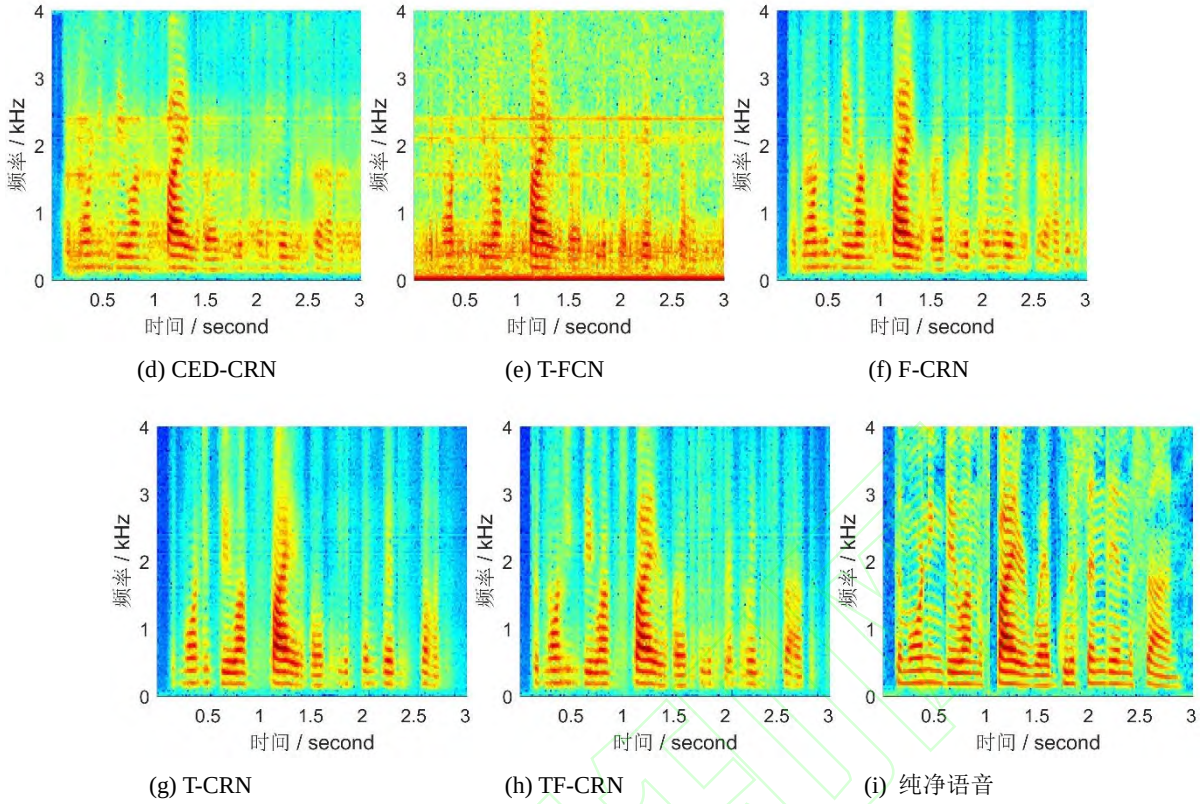
Table 1 Comparison of speech enhancement performance of different networks under different SNRs

评价指标	平均 PESQ 得分			平均 STOI 得分		
信噪比(dB)	-7	0	7	-7	0	7
含噪语音	1.43	1.77	2.20	53.43	69.15	83.79
F-DNN	1.39	1.91	2.40	48.14	65.81	79.04
F-LSTM	1.75	2.28	2.75	61.61	78.02	87.94
CED-CRN	1.77	2.30	2.77	61.26	78.48	89.45
T-FCN	1.57	1.93	2.27	54.89	70.99	81.11
TF-CRN	1.92	2.48	2.90	64.72	80.66	88.83

下面通过对比不同网络增强语音的语谱图来更加直观的比较不同网络的语音增强性能。图 4(a)给出了一段含有 N3 噪声信噪比为 0dB 的含噪语音的语谱图，图 4(b)-(h)则分别给出了采用 F-DNN、F-LSTM、CED-CRN、T-FCN、F-CRN、T-CRN 和 TF-CRN 进行处理后增强语音的语谱图，图 4(i)则给出了相应的纯净语音的语谱图作为对比。通过对比增强语音和纯净语音的语谱图，可见：F-DNN 和 T-FCN 增强后的语音仍然存在大量的噪声成分，表明 F-DNN 和 T-FCN 的语音增强性能较差；F-LSTM、CED-CRN 和 F-CRN 三种频域语音增强方法虽然能够抑制大部分的噪声成分，但是增强后的语音中仍然存在明显可见的噪声成分；T-CRN 具有最好的噪声抑制能力，但是对于语音成分的保留能力稍逊于 TF-CRN；七种方法中，TF-CRN 在噪声成分抑制和语音成分保留上取得了最好的折衷效果，具有最好的语音增强性能。非正式的试听实验也进一步验证了上述结论。



(a) 含噪语音　(b) F-DNN　(c) F-LSTM



(d) CED-CRN (e) T-FCN (f) F-CRN

(g) T-CRN (h) TF-CRN (i) 纯净语音

图 4 不同网络增强语音的语谱图比较

Figure 4 Comparison of spectrograms of speech enhanced by different networks

为了说明时频域特征融合网络在结构设计上的有效性，将其与其它四种网络进行空间复杂度的比较。表 2 给出了不同网络的参数规模，可见：由于采用了全卷积网络，T-FCN 的参数量要远低于其他网络；在具有较好语音增强性能的 F-LSTM、CED-CRN、F-CRN、T-CRN 和 TF-CRN 五种网络中，CED-CRN 由于网络结构复杂，参数量最大，F-LSTM 次之；而本文三种网络参数量则明显小于两者，其中 TF-CRN 的参数量约为 F-LSTM 的 38.35%，约为 CED-CRN 的 21.75%，表明本文的时频域特征融合网络利用更小的网络参数规模取得更好的语音增强性能。

表 2 不同网络的参数规模比较

Table 2 Comparison of parameter scales of different networks

网络	F-DNN	F-LSTM	CED-CRN	T-FCN	F-CRN	T-CRN	TF-CRN
参数量(M)	8.92	5.58	9.84	0.01	1.38	1.58	2.14

4 结束语

在基于深度神经网络的语音增强方法中，采用时域波形作为训练特征和训练目标的时域方法非常依赖于损失函数的设计；而采用对数功率谱作为训练特征和训练目标的频域方法则受限于短时傅里叶变换的特征表达能力，无法充分利用含噪语音的特征。为了解决上述问题，本文采用含噪语音的时域波形作为训练特征，同时采用纯净语音的对数功率谱作为训练目标，通过设计合理的网络结构 T-CRN 建立含噪语音时域波形和纯净语音对数功率谱之间的映射关系，并进一步设计了一种能够融合含噪语音的时域波形特征和频域对数功率谱特征的网络 TF-CRN。实验结果表明，与采用对数功率谱作为训练特征的频域语音增强方法相比，T-CRN 具有更好的语音增强性能，而融合时域和频域特征的 TF-CRN 更是相比其它网络显著提高了增强语音的语音质量和可懂度。在本文工作中，对于网络结构的设计主要依赖于实验结果的验证，缺乏详尽的理论分析与证明，下一步的研究工作将结合卷积神经网络的特征可视化技术，对网络中的特征计算过程进行深入研究，从而进一步优化网络结构。

参考文献

[1] 刘文举, 聂帅, 梁山, 等. 基于深度学习语音分离技术的研究现状与进展[J]. 自动化学报, 2016, 42(6): 819-833.
LIU Wenju, NIE Shuai, LIANG Shan, et al. Deep learning based speech separation technology and its developments[J]. Acta Automatica Sinica, 2016, 42(6): 819-833.

[2] WANG D L, CHEN J. Supervised speech separation based on deep learning: An overview[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 26(10): 1702-1726.

[3] WANG Y, NARAYANAN A, WANG D L. On training targets for supervised speech separation[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014, 22(12): 1849-1858.

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

[5] XU Y, DU J, DAI L R, 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.

[6] 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.

[7] WENINGER F, ERDOGAN H, WATANABE S, et al. Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR[C]// Proceedings of International Conference on Latent Variable Analysis and Signal Separation. Liberec: Springer International Publishing, 2015: 91-99.

[8] PARK S R, LEE J. A fully convolutional neural network for speech enhancement[C]// Proceedings of the Eighteenth Annual Conference of the International Speech Communication Association. Stockholm: ISCA, 2017: 1993-1997.

[9] FU S W, TSAO Y, LU X. SNR-aware convolutional neural network modeling for speech enhancement[C]// Proceedings of the Seventeenth Annual Conference of the International Speech Communication Association. California: ISCA, 2016: 3768-3772.

[10] TAN K, CHEN J, WANG D. Gated residual networks with dilated convolutions for supervised speech separation, [C]// Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing. Alberta: IEEE, 2018: 21-25.

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

[12] LI Y, LI X, DONG Y, LI M, XU S, XIONG S. Densely connected network with time-frequency dilated convolution for speech enhancement [C]// Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing. Brighton: IEEE, 2019: 6860-6864.

[13] ZHAO H, ZARAR S, TASHEV I, et al. Convolutional-recurrent neural networks for speech enhancement[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018: 2401-2405.

[14] TAN K, WANG D L. A Convolutional Recurrent Neural Network for Real-Time Speech Enhancement[C]//Interspeech. 2018: 3229-3233.

[15] FU S W, TSAO Y, LU X, et al. Raw waveform-based speech enhancement by fully convolutional networks[C]//2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE, 2017: 006-012.

[16] FU S W, WANG T W, Tsao Y, et al. End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks[J]. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), 2018, 26(9): 1570-1584.

[17] PANDEY A, WANG D L. TCNN: Temporal Convolutional Neural Network for Real-Time Speech Enhancement in The Time Domain[C]//ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019: 6875-6879.

[18] PANDEY A, WANG D. A New Framework for Supervised Speech Enhancement in the Time Domain[C]//Interspeech. 2018: 1136-1140.

[19] 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 (TASLP), 2019, 27(7): 1179-1188.

[20] GAROFOLO J S, LAMEL L F, FISHER W M, et al. TIMIT acoustic-phonetic continuous speech corpus[J]. Linguistic data consortium, Philadelphia, 1993, 33.

[21] HU G, “100 nonspeech environmental sounds, 2004” [OL]. http://web.cse.ohiostate.edu/pnl/corpus/HuNonspeech/HuCorpus.html.

[22] VARGA A, STEENEKEN H J M. Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems [J]. Speech Communication, 1993, 12(3): 247-251.

[23] 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 IEEE International Conference on Acoustics, Speech, and Signal Processing. Utah: IEEE, 2001: 749-752.

[24] TAAL C H, HENDRIKS R C, HEUSDENS R, et al. An algorithm for intelligibility prediction of time–frequency weighted noisy speech[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2011, 19(7): 2125-2136.