doi: 10.3969/j.issn.0490-6756.2020.02.014

多特征全卷积网络的地空通话语音增强方法

高登峰[1,2], 杨　波[1], 刘　洪[1], 杨红雨[1]
(1. 四川大学国家空管自动化系统技术重点实验室, 成都 610065; 2. 四川大学计算机学院, 成都 610065)

摘　要: 为了研究空中交通管理领域中的语音增强问题,并且节约存储资源,提出了一个新的语音增强方法. 在基于全卷积神经网络(FCN)的基础上加入了跳跃连接(Skip Connection),并引入次要特征来进行联合学习. 具体而言,使用语音的对数功率谱(LPS)作为网络的主要训练特征,引入对数梅尔倒谱系数(L-MFCC)作为网络的次要训练特征,来联合优化网络参数. 实验证明,相较于单个 LPS 特征输入的架构,结合 LPS 和 L-MFCC 的多特征网络架构具有更好的语音增强性能表现,且作为次要特征的 L-MFCC 还可以用作其它用途. 实验还证明,跳跃连接的加入可以很好的提高 FCN 的网络性能,且相较于基线的深度神经网络(DNN)模型,新的网络结构在相同参数数量的情况下,要具有更好的性能.

关键词: 语音增强; 语音分离; 全卷积神经网络; 地空通话; 多特征联合学习
中图分类号: TN912.35; TP183　　**文献标识码**: A　　**文章编号**: 0490-6756(2020)02-0289-08

A method of multi-featured full convolutionalneural network based on speech enhancement in air-ground voice communication

GAO Deng-Feng[1,2], *YANG Bo*[1], *LIU Hong*[1], *YANG Hong-Yu*[1]
(1. National Key Laboratory of Air Traffic Control Automation System Technology, Sichuan University, Chengdu 610065, China; 2. College of Computer Science, Sichuan University, Chengdu 610065, China)

Abstract: In order to study speech enhancement in the air traffic control (ATC) and save storage resources, a new speech enhancement method is proposed. Based on Fully Convolutional Networks (FCN), Skip connection is added and secondary features are introduced for joint learning. Specifically, the log-power spectra(LPS) of speech is used as the main training feature, and the logarithmic Mel-Frequency Cepstrum (L-MFCC) is introduced as the secondary training feature to jointly optimizeparameters of FCN. Experiments have shown that the network architecture combining LPS and L-MFCC has better speech enhancement performances than that with single LPS feature, and the L-MFCC as a secondary feature can also be used for other purposes. Experiments also show that the addition of skip connections can improve the FCN network performances, and the new network structure has better performances with the same number of parametersthan the baseline deep neural network (DNN) method.

Keywords: Speech Enhancement; Speech Separation; Full Convolutional Neural Network; Air-Ground Communication; Multi-featured Joint Learning

收稿日期: 2019-03-28
基金项目: 国家自然科学基金委和民航局联合基金(U1833115)
作者简介: 高登峰(1994－), 男, 山西运城人, 硕士研究生, 研究方向为音频处理算法.
通讯作者: 刘洪. E-mail: liuhong@scu.edu.cn

1 引 言

在空管领域当中,由于通讯条件的限制,通信双方的语音信号总是不可避免的会被噪声所干扰,严重损害了语音的可懂度,降低了通信质量.对此,前端语音增强技术仍是最常用和最有效的解决方法之一[1].然而,如何在空管对话环境中,即单声道通信和复杂非平稳噪声较多的条件下取得良好的语音增强表现,仍是一个重要的挑战.

在传统的语音增强算法当中,无监督的方法有谱减法[2]和滤波器法[3],这类方法都是基于语音和噪声之间的数学假设.缺点是语音增强的性能不佳且对未知噪声类型的泛化性较差.随着机器学习技术的发展,语音增强在有监督的方法方面取得了很大的进展,如基于非负矩阵分解[4]的语音增强方法,但是其语音和噪声是分开处理的,无法很好的学习语音和噪声之间的复杂关系.

近年来,基于深度学习的语音增强方法取得了广泛的研究成果[5-7].这类方法的基本原理是通过深度学习技术来建立一种噪声语音到干净语音间的映射函数.在文献[5,6,8]当中,已经证明基于深度神经网络(Deep Neural Networks, DNN)的语音增强方法要比传统的语音增强方法表现更好.此外,基于深度神经网络的语音增强方法还比基于MMF的语音增强方法占用更少的计算资源[9].但是,深度神经网络的语音增强方法仍然存在两个问题没有解决.(1) 语音信号的局部时间谱结构的信息利用;(2) 作为空管语音识别系统的前端之一,采用深度神经网络会占据过多的存储空间和计算资源,不利于部署在小型设备上.

文献[6,8]采用的DNN模型是以全连接的方式来处理语音特征的,这样就无法有效的利用语音信号的局部时间谱结构信息.相反,卷积神经网络(Convolutional Neural Networks, CNN)的体系结构则可以更加关注输入特征的局部时间谱结构.与DNN相比,CNN更加关注每个时频单元(T-F)周围的邻近区域,这样CNN可以更好的拟合信号中空间信息和时间信息的相关性,并减少信号中的平移方差[10].在图像处理领域,CNN已经取得了巨大的研究成果,如图像识别[11],图像分类等.而音频处理领域则借鉴了图像处理的一些思想,也使用CNN取得了广泛的研究成果.例如,文献[12]提出了一种基于CNN的音乐去除模型,与DNN相比,采用CNN获得了更好的识别效果.文献[13]通过估计时频单元的理想比率掩模,采用CNN来分离语音和噪声.此外,CNN通过权值共享还可以大规模的减少需要训练的参数数量从而降低网络规模,节约存储空间和计算资源,使得其可以更好地部署在小型设备上.

CNN的方法是一种数据驱动的方法,其基本原理是对生物神经网络的一种模拟和近似,利用大量的神经元通过互相连接来组成的一种自适应非线性的模拟系统,通过学习和训练来调整其网络神经元的参数,从而构建一种从噪声语音特征到干净语音特征的映射模型,然后通过模型来进行语音增强.但目前还没有一种CNN的语音增强方法在空管语音数据集上取得过成功的表现.在目前的语音增强方法研究当中[6,14],大部分所使用训练集的噪声来源皆为单一噪声,如嘈杂人声,汽车声,雨声等.其噪声特征较为明显,对语音的破坏程度相对较小.而在空管对话系统当中,受到通信条件的限制,其语音信号当中一般包含多种复杂噪声,其噪声类型有复杂加性噪声,信号传播导致的声学混响,加性宽带电子噪声,非线性信号失真引起的噪声,信号传播干扰引起的噪声以及相关仪器引起的噪声等.其噪声特征复杂,对语音的破坏程度更大.此外,之前的大部分研究方法当中所使用的语音特征多为利用短时傅里叶变换到频域后,对每帧进行取绝对值和对数运算的LPS特征[5,6,9].在LPS特征域中,不同频率区间的目标值是独立预测的,没有任何其它的相关约束,并且不容易利用听觉感知中的一些信息[15].因此,当仅利用LPS作为语音特征来进行语音增强处理时,重建后的波形往往会出现部分失真.所以,本文提出了一种多特征联合训练的网络架构,采用多个语音特征来联合优化目标函数.这种特征不同但优化目标相同的架构可以显著的改善重建后语音的鲁棒性.此外,次要特征还可以作为辅助信息来用作其它用途,比如语音质量的评测,语音设备来源检测[16]等.

2 主要流程

本方法的目标是通过卷积神经网络来建立从噪声语音特征到干净语音特征的映射模型.所采用的框架与文献[5]类似.在训练阶段,准备多组噪声语音-干净语音的语音数据对,然后提取其特征,分为一个主要特征和一个次要特征.然后将其特征合并为联合特征一起作为模型的训练特征,以训练回归网络.在语音增强阶段,则将噪声语音的联合特

征输入到训练好的网络模型中，以产生增强后的联合特征. 在语音重建阶段，则从增强得到的联合特征中提取出语音的主要特征，并重建语音波形信息，其具体流程如图 1 所示.

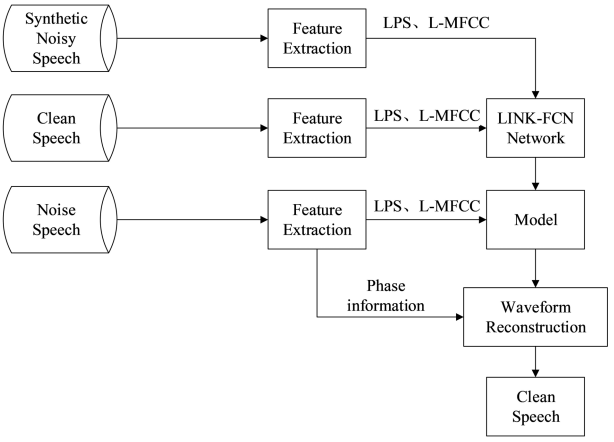

图 1　语音增强基本流程图

Fig. 1　Basic flowchart of speech enhancement

2.1　语音特征的提取

首先，需要确定所采用的语音特征. 在主要特征当中，理想二元掩模，理想比例掩模，短时傅里叶变换幅度谱及其掩模[17-18]，对数功率谱等特征都曾被使用，综合比较其结果后发现，采用对数功率谱作为主要的训练目标，其表现要更为出色[19]. 次要特征选取的是对数梅尔倒谱系数(Logarithmic Mel-Frequency Cepstrum, L-MFCC). MFCC 特征是语音识别[20]，说话人识别[21]和音乐建模中最流行的语音特征之一，应用梅尔滤波可以使处理后的语音信号与人类的听觉感知相一致，次要目标选取 MFCC 可以更好的约束网络. 此外，MFCC 中的离散余弦变换(Discrete Cosine Transform, DCT)操作还可以将不同信道的相关信息合并到每个 MFCC 系数当中，这样可以学习到语音不同频率域的相关信息. 此外，对 DCT 操作进行了尺寸固定，提取出的 MFCC 特征尺寸与梅尔滤波器的个数相同. 同时为了与 LPS 的值域保持一致以便更好地约束网络权重的分配，还对 MFCC 执行了取对数操作(L-MFCC). 特征提取的公式如下.

$$N(t,f)=\log(|STFT(n^u)|) \tag{1}$$

$$M(t,f)=\log(|MFCC(n^u)|) \tag{2}$$

其中，N 表示语音信号 n^u 的对数功率谱；$STFT$ 表示短时傅里叶变换；而 M 表示为语音信号 n^u 的对数梅尔倒谱；$MFCC$ 表示梅尔倒谱滤波；t 和 f 分别代表语音信号的时间和频率.

然后，对语音特征执行扩帧操作. 文献[22]已经证明，特征帧的扩展有利于将语音中的噪声信息更好地反馈到神经网络当中(称之为噪声感知训练)，相较于单帧输入的方式，多帧堆叠可以很好地提高语音增强的性能[6]. 设 n_t 为 $N(t,f)$ 的 t^{th} 帧，将上下文扩展帧表示为 y^t，则

$$y^t=[n_{t-\tau},\cdots,n_{t-1},n_t,n_{t+1},\cdots,n_{t+\tau}] \tag{3}$$

最后，将两种特征组合为联合特征 $S(t,f)$，作为训练网络的输入和输出，即

$$S(t,f)=\text{Concatenate}(N(y^t,f),M(y^t,f)) \tag{4}$$

同时，对联合特征做归一化处理，使其均值为零，方差为一，这样处理可以使得训练出的网络模型具有更好的性能.

2.2　网络模型的构建

本研究并没有使用传统的 CNN 架构[10]，即包含多层卷积层和池化层，输出层则为若干全连接层的架构. 因为实验发现，池化层的加入会导致增强后的语音信息出现严重的失真[9]. 因此，本文提出了一种结合自动编码器原理的卷积网络. 它由若干层重复的卷积层，批标准化层，ReLU 激活层所组成，没有任何的池化层和采样层，在输出层采用的仍然是卷积层，这使得网络成为了一个全卷积网络(FCN). 注意，在本文中所提到和采用的卷积层指的都是一维卷积，卷积方向均为频域方向.

FCN 网络分为编码器和解码器两部分，首先沿着编码器将语音特征逐步编码为较高的维度，之后则沿着解码器将其逐步解码还原，其中编码器和解码器的卷积层数量和维度皆保持对称. 此外，本文还将跳跃连接添加到 FCN 网络当中，以便在训练阶段更好地进行优化并提高性能. 跳跃连接的添加方式是将编码器和解码器中的相同维度的层分别进行连接，这样的网络，本文称为 LINK-FCN 网络网络，其结构如图 2 所示.

另外，还与其它多种网络结构进行了性能对比，其中包括作为基线的全连接网络(DNN)，没有添加跳跃连接的全卷积网络(FCN)以及仅使用 LPS 特征的 LINK-FCN-1f 网络.

在之前的基于神经网络的语音增强研究当中[5-6,19]，都使用了基于 RBM 或者基于自动编码器的预训练技术来用于神经网络的学习. 但实验发现，当给定的训练数据集足够大时，便可以跳过预训练阶段. 所取得的训练结果与采用了预训练的结果相比，几乎没有区别.

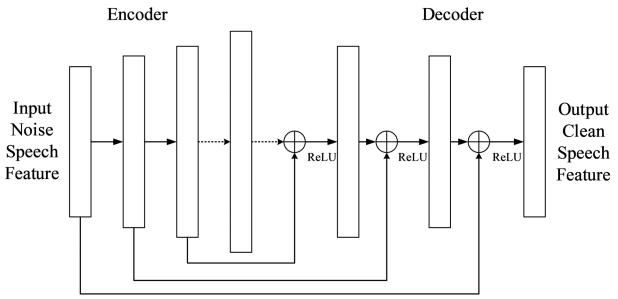

图 2 LINK-FCN 网络结构示意图

Fig. 2 LINK-FCN network structure diagram

所有网络的训练,均采用最小均方误差作为损失函数,采用 Adam 作为优化函数,Batchsize 为 512,卷积层的卷积核大小为 11,其它具体的网络参数如表 1 所示. 而学习率则根据不同的网络采用了不同的大小. 为了保证所有模型训练的充分性,实验会通过调整网络模型的超参数,使其损失值 loss 达到最小,精度值 acc 达到最高. 所有的模型均保证了其训练效果为最佳.

2.3 语音波形的重建

在预测阶段,则是对未知噪声语音进行语音增强并进行波形重建,首先利用已经训练完成的网络模型来对噪声语音进行增强处理,产生增强语音的联合特征帧. 然后从中提取出 LPS 特征,之后由以下公式来进行频谱重构.

$$Y^f(d)=\exp\{\hat{Y}^f(d)/2\}\cdot\exp\{j\cdot\angle X^f(d)\} \tag{5}$$

其中,$\angle X^f(d)$表示噪声语音 X 在 d^{th} 帧处的相位信息. 虽然语音的相位信息在人类的听觉识别当中具有十分重要的作用,但是考虑到人耳对微小的相位信息失真并不十分敏感[23],可以直接利用原始噪声语音当中的相位信息. 所以,利用模型输出的 LPS 特征结合原始噪声语音中的相位信息,然后执行逆傅里叶变换操作将信号转换回时域,再利用文献[19]中语音帧的重叠相加的方法合成整个语音波形. 最后,噪声语音文件便被语音增强成为干净语音文件.

3 实验设置

3.1 实验数据设置

实验所采用的数据皆是从空管地空通话系统中实际通信的语音流当中采集而来,数据来自成都空管局,成都机场,太原机场,上海机场等地. 其中对话人的性别分布均衡,中英文指令分布均衡,语音地区分布均衡. 这样设置的数据集可以增强网络的泛化性且不会对性能产生消极影响.

数据集准备完毕后,还需要进一步进行处理. 首先,利用语音活动检测系统[24](Voice Activity Detection,VAD)对数据进行静音切除,并将处理后的语音数据统一设置为单声道,采样率为 8 kHz 的 WAV 文件. 另外,由于空管对话系统的条件限制,实验很难找到没有任何噪声的干净语音,因此,根据频域信噪比(F-SNR)来对语音数据进行分类,将 F-SNR>10 dB 的语音文件称之为"干净语音",而将 F-SNR<2 dB 的文件称为"噪声语音". 其中,频域信噪比(F-SNR)的计算公式如下.

$$F\text{-}SNR(dB)=20\cdot\log10\frac{\mathrm{std}(|STFT(A_{singal})|)}{\mathrm{mean}(|STFT(A_{\mathrm{singal}})|)+10^{-4}} \tag{6}$$

其中,A_{singal} 为信号谱幅度;$STFT$ 指的是短时傅里叶变换;std 指的是标准偏差;mean 指的是其算数平均值.

接下来,从"干净语音"当中随机选取一批数据作为语音样本,共计 20 339 条语音,时长总计约为 40 h. 而噪声的选取则是从"噪声语音"当中选取某些纯噪声片段而获得的,包括复杂加性噪声,复杂非平稳噪声,声学混响,加性宽带电子噪声,信号失真噪声,仪器噪声等,大致 200 余类噪声. 所有的语音样本都通过随机选取噪声和平滑扩展的方式被添加上了噪声. 即

$$X(t)=S(t)+\alpha\cdot N(t) \tag{7}$$

其中,系数为

$$\alpha=\frac{10^{SNR\langle dB\rangle/20}}{A_{\mathrm{speech}}/A_{\mathrm{noise}}} \tag{8}$$

$X(t)$代表合成后的噪声语音信号;$S(t)$表示干净语音信号;$N(t)$表示噪声信号. 通过调节参数 α 来调节噪声大小,使得合成后的语音信号信噪比(SNR)均匀分布在 0 dB 到 10 dB 的区间之内,注意,用于合成的"干净语音"并非真正的干净语音,所以合成后语音的真实信噪比会比实际的更小,但并不影响将其作为噪声大小程度的依据.

最终用于训练的数据集共有合成的语音文件对 40 678 个,语音时长共计 80 h 左右. 并用同样的方法,采用不同的数据构建测试集,共有合成语音文件对 1 000 个,语音时长共计 2 h 左右.

验证集的构造则分为两种,验证集 1 是与上述方法相同的合成语音,每个语音都会根据式(7)合成 SNR 为 0 dB,2 dB,5 dB,10 dB 的 4 组数据,每

组间的语音相同. 验证集 2 则是非合成的噪声语音，即从真实的噪声语音当中随机选取的一批数据.

此外，进行特征提取时，语音帧的长度为 256（即 32 ms），帧位移长度为 128. 使用短时傅里叶变换（DFT）来将语音信号转换到频域. 进行 MFCC 变换时，采用的梅尔滤波器的个数为 78 个，窗口函数选用的为 Hamming 窗，扩帧时的位移为 7.

3.2　实验设置

3.2.1　DNN vs FCN　**实验 1** 是将作为基线的 DNN 网络与本文提出的 FCN 网络来进行性能对比，以证明 FCN 网络进行语音增强的优越性. 为了比较其性能表现，实验将两个网络的参数数量都维持在相同的数量级，且 FCN 没有添加跳跃连接. 采用的语音特征皆为 LPS 和 L-MFCC 的联合特征. 具体的结构参数如表 1 所示.

3.2.2　FCN vs LINK-FCN　**实验 2** 是将 FCN 与添加了跳跃连接的 LINK-FCN 进行了性能对比，以证明跳跃连接的添加可以切实提高 FCN 中语音增强的性能表现. 其网络结构和参数完全相同，只是 LINK-FCN 在对应的网络层中添加了跳跃连接，采用的语音特征皆为 LPS 和 L-MFCC 的联合特征. 其网络结构如图 2 所示.

3.2.3　LINK-FCN-1f vs LINK-FCN　**实验 3** 是将仅使用 LPS 作为特征的 LINK-FCN-1f 与使用 LPS 和 L-MFCC 联合特征的 LINK-FCN 进行了性能对比，以证明多特征联合训练的架构可以很好的提高语音增强的性能. 实验所采用的网络结构相同，仅训练特征不同.

表 1 中，Flatten 指的是扁平层；Dense 指的是全连接层；Conv1D 指的是一维卷积层；BN 指的是批标准化层（Batch Normalization，BN）；Nodes 指的是每层卷积核的数目.

3.3　实验评价指标

为了评估和对比各个网络的语音增强的性能表现，实验采用了 4 种常用的客观测量指标和一种主观评价指标来进行性能评判. 分别是平均绝对误差比对（Mean Absolute Deviation，MAD），频域信噪比（Frequency Signal-Noise ratio，F-SNR），语音质量感知评估（Perceptual Evaluation of Speech Quality，PESQ）和短时客观可懂度（Short-Time Objective Intelligibility，STOI）.

平均绝对误差比对（MAD）是用来评判模型表现常用的度量之一，表示模型预测的估计值与真实值之间的差异化程度，其值越低越好. 计算公式为

$$MAD = \frac{1}{N}\left(\sum_{i}^{N} |f(X_i) - Y_i|\right) \tag{9}$$

其中，$f(X_i)$为根据模型得出的估计语音特征；Y_i 为真实干净语音特征，用来对比的特征为 LPS.

信噪比一直是衡量语音失真程度的常用指标之一，但由于真实的干净语音并不可见，所以采用频域信噪比（F-SNR）作为比较标准，其计算方法如式（6）所示.

语音质量感知评估（PESQ）是 ITU-T 在 2001 年推出的 P.862 标准中建议使用的语音质量评价指标. 其得分范围在 0.5～4.5 之间，得分越高则表示语音质量相对越好.

表 1　网络配置参数

Tab.1　Network config

Networks	Layers	Nodes	Parameters
DNN	Flatten×1 +(Dense,ReLU)×5	1150-1150-1150-1150-1150	8 378 085
FCN	(Conv1D,ReLU,BN)×16 +Conv1D	8-16-32-64-128-256-384-512- 384-256-128-64-32-16-8-1	8 353 933
LINK-FCN	(Conv1D,ReLU,BN)×16 +Conv1D	8-16-32-64-128-256-384-512- 384-256-128-64-32-16-8-1	8 352 909
LINK-FCN-1f	(Conv1D,ReLU,BN)×16 +Conv1D	8-16-32-64-128-256-384-512- 384-256-128-64-32-16-8-1	7 983 424

短时客观可懂度（STOI）是音频处理领域常用的评价指标之一，通过与原始的干净语音进行时频对比而得出评分，其值在 0～1.0 之间，分值越高表明越接近干净语音，其语音质量越好. 实验采用了两种 STOI 算法来进行评分，一种是传统的 STOI 算法，一种是 Jesper Jensen 提出的 ESTOI（Extended STOI）算法[25].

主观评价方法则采用常用的平均意见分法（Mean Opinion Score，MOS）. 参与评分的人员共有 20 人，在不告知数据来源的情况下，听取原始语

音和增强后的语音,并对该语音做出 0～5.0 分的评价.每个网络模型产生的数据为 5 条,然后随机分发给测试人员.

4 实验结果与分析

4.1 客观测量指标分析

在不同 SNR 级别下的测试结果如表 2～表 5 所示,其中的值均从验证集 1 当中测得数据的平均值,数值加粗的是在当前组内表现最优的数据.

表 2 在 0 dB 下的对比结果

Tab. 2 Comparison results at 0 dB

Networks	MAD	STOI	ESTOI	PESQ	F-SNR	Parameters
Noisy	——	0.678 4	0.460 4	1.770 3	0.782 8	——
DNN	0.702 6	0.794 9	0.604 8	2.476 9	9.315 4	8 378 085
FCN	0.655 2	0.807 2	0.627 5	2.544 1	9.866 1	8 353 933
LINK-FCN	0.654 7	**0.821 2**	**0.656 4**	**2.612 0**	**9.932 4**	8 352 909
LINK-FCN-1f	**0.642 7**	0.778 2	0.592 1	2.455 6	9.438 6	**7 983 424**

表 3 在 2 dB 下的对比结果

Tab. 3 Comparison results at 2 dB

Networks	MAD	STOI	ESTOI	PESQ	F-SNR	Parameters
Noisy	——	0.716 1	0.506 8	1.852 8	1.596 2	——
DNN	0.702 6	0.815 7	0.633 6	2.560 9	9.548 4	8 378 085
FCN	0.655 2	0.827 3	0.660 1	2.642 1	9.957 8	8 353 933
LINK-FCN	0.654 7	**0.844 4**	**0.693 4**	**2.719 3**	**10.054 6**	8 352 909
LINK-FCN-1f	**0.642 7**	0.809 4	0.635 4	2.587 1	9.884 1	**7 983 424**

由表 2～表 5 可以看出,在不同 SNR 级别的噪声语音环境当中,本文提出的方法均能将语音提升到 F-SNR>9 dB 的程度,说明提出的方法基本都具有很好的语音增强表现.在训练数据当中出现的噪声类型都取得了很好的滤除效果.且对于复杂加性噪声和复杂平稳噪声的滤除性能更好,其原理可能是 CNN 可以更充分地利用语音的时频相关性,使得其对于在整个时域上分布规律的噪声类型拥有更好的滤除性能.但是,无论噪声数据的 SNR 级别是多少,提出的方法将其 F-SNR 提升的上限基本相同,这也证明了深度学习是一种数据驱动的方法,其所能达到的最大性能与所使用的数据相关.

另外,从表 2～表 5 还可以看出,在任何 SNR 级别的验证数据下,无论是否添加跳跃连接,在相同参数级别的条件下,FCN 网络结构总是要优于 DNN 网络结构,这也证明 FCN 可以更好的利用语音的时频信息,拥有更好的语音增强表现.

表 4 在 5 dB 下的对比结果

Tab. 4 Comparison results at 5 dB

Networks	MAD	STOI	ESTOI	PESQ	F-SNR	Parameters
Noisy	——	0.779 6	0.593 5	2.025 0	3.393 7	——
DNN	0.702 6	0.838 7	0.677 4	2.680 9	9.579 3	8 378 085
FCN	0.655 2	0.852 8	0.705 6	2.758 3	9.874 1	8 353 933
LINK-FCN	0.654 7	**0.870 6**	**0.741 2**	**2.856 8**	**10.089 5**	8 352 909
LINK-FCN-1f	**0.642 7**	0.843 3	0.695 7	2.739 9	9.568 6	**7 983 424**

表 5 在 10 dB 下的对比结果

Tab. 5 Comparison results at 10 dB

Networks	MAD	STOI	ESTOI	PESQ	F-SNR	Parameters
Noisy	——	0.851 7	0.697 2	2.325 5	5.641 4	——
DNN	0.702 6	0.854 9	0.711 0	2.822 5	9.691 1	8 378 085
FCN	0.655 2	0.869 6	0.737 9	2.883 9	9.792 4	8 353 933
LINK-FCN	0.654 7	**0.890 4**	**0.779 6**	**3.014 5**	**10.322 8**	8 352 909
LINK-FCN-1f	**0.642 7**	0.863 8	0.736 4	2.873 2	10.077 1	**7 983 424**

通过 FCN 和 LINK-FCN 的对比,可以发现,跳跃连接的添加可以有效地提高网络的性能,每个 SNR 级别的验证集下,LINK-FCN 相比于 FCN,PESQ 的提升都在 0.1 以上,ESTOI 的提升也普遍都在 0.03 以上,是一个不小的性能提升.分析其原因是,FCN 网络中的解码器解码时会丢失部分编码器处的语音信息,而跳跃连接的添加则为解码器提供了部分编码器处的信息,使得增强后的语音鲁棒性更好.

而通过 LINK-FCN 和 LINK-FCN-1f 的对比,可以发现,多特征联合训练的网络性能,要远远好于单特征训练的网络性能,其中 PESQ 的提升达到了 0.15 以上,ESTOI 的提升也有 0.04 以上.还可以看出,在 SNR 为 0 dB 的验证集当中,单特征的 LINK-FCN-1f 表现甚至还差于多特征的 DNN 网络.这也证明了,多特征联合训练的网络可以很大程度地提高语音增强的性能表现.另外,在 MAD 项的表现上,LINK-FCN-1f 的表现则优于其他网络,原因是 MAD 进行分析对比时仅采用了 LPS 特征,这也证明了单特征训练的网络往往会陷入该特征过拟合的状态,使得网络的鲁棒性变差.

4.2 主观性能分析

因为现实中面对的噪声语音并非合成的噪声

语音，所以实验也准备了真实的噪声语音来进行评测(即验证集 2)，但由于没有真实的干净语音来进行对比分析，所以采用主观评测的方式来进行，测试结果如表 6 所示，其结果为平均值.

表 6　主观性能对比

Tab. 6　Subjective performance comparison of networks

Networks	MOS	F-SNR
Noisy	2.5	1.3502
DNN	3.3	6.9601
FCN	3.4	7.0846
FCN-LINK	**3.6**	**7.2291**
FCN-LINK-1f	3.3	7.0731

从表 6 中可以看出，实验结果与验证集 1 的表现基本一致，本文所提出的方法均能对真实的噪声语音进行很好的噪声滤除效果，尤其是对于在训练数据集当中出现过的噪声类型. 而对于未出现在训练数据集中的未知噪声类型，模型仍旧可以对其中的复杂加性噪声类型和复杂平稳噪声类型以及分布接近的复杂非平稳噪声进行一定的滤除，而其它噪声类型的滤除性能则有所下降. 其原理是 CNN 的局部连接和权值共享特性，使得网络模型对未知噪声类型拥有更好的泛化能力.

4.3　频谱图的对比

此外，还选取了增强前后的语音频谱图作为对比，由于篇幅限制，仅选取一例来进行分析，结果如图 3 所示，从上到下依次为噪声语音，干净语音，以及对应网络语音增强后的语音，其中虚线方框处表示信息丢失的部分.

从图 3 中可知，本文方法基本取得了不错的语音增强效果，可很好地在保留原有语音信息的基础上去除噪声，但相较于原有的干净语音，都存在不同程度的信息失真. 图 3 中，DNN 与 FCN 相比，丢失了更多的语音高频信息，而 FCN 则普遍失真较少. 这也证实了 DNN 对于语音信号的局部时间谱结构信息无法有效的利用. 而 FCN 和 FCN-LINK 的对比则证明，跳跃连接的添加可有效地减少语音信息的丢失，使得网络的鲁棒性更好. FCN-LINK 和 FCN-LINK-1f 的对比则可以看出，FCN-LINK-1f 在中低频域丢失了很多的信息，而 L-MFCC 特征的加入则使得中低频语音失真大大减少且更加一致，分析其原因是 L-MFCC 中的 Mel 滤波强调了低频信息，这也证明了多特征联合训练的方式可以显著的提高增强后语音信息的鲁棒性.

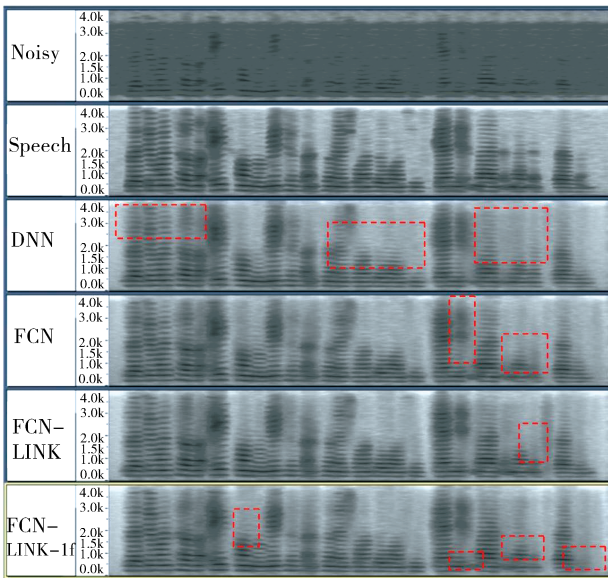

图 3　语音频谱图对比

Fig. 3　Speech spectrum comparison

综上所述，本文所提出的基于多特征的全卷积神经网络的语音增强方法在空管语音数据集上取得了最优秀的表现.

5　结　论

在本文中，研究了在复杂噪声条件下的空管对话语音的语音增强技术. 提出了一种基于多特征的全卷积神经网络的语音增强方法，同时在网络中添加跳跃连接来获得更好的性能表现. 还通过在目标函数当中添加 L-MFCC 特征来约束网络，显著增强了增强语音的鲁棒性. 实验证明，本文所提出的方法在空管对话语音数据集当中取得了十分优秀的表现，可以显著地减少语音信息的失真.

参考文献：

[1]　Wang D, Chen J. Supervised speech separation based on deep learning: an overview[J]. IEEE T Audio Speech, 2018, 26: 1702.

[2]　Karam M, Khazaal H F, Aglan H, *et al*. Noise removal in speech processing using spectral subtraction [J]. J Signal Inf Process, 2014, 5: 32.

[3]　Elfattah M A, Dessouky M I, Abbas A M, *et al*. Speech enhancement with an adaptive Wiener filter [J]. Int J Speech Technol, 2014, 17: 53.

[4]　Nasser M, Paris S, Arne L. Supervised and unsupervised speech enhancement using nonnegative matrix factorization [J]. IEEE T Audio Speech, 2013, 21: 2140.

[5]　Xu Y, Du J, Dai L R, *et al*. An experimental study

on speech enhancement based on deep neural networks [J]. IEEE Signal Process Lett, 2014, 21: 65.

[6] Xu Y, Du J, Dai LR, *et al*. A regression approach to speech enhancement based on deep neural networks [J]. IEEE/ACM T Audio, Speech, 2015, 23: 7.

[7] 王怡斐, 韩俊刚, 樊良辉. 基于 WGAN 的语音增强算法研究[J]. 重庆邮电大学学报: 自然科学版, 2019, 31: 140.

[8] 高登峰, 郭东岳, 杨波. 基于深度神经网络的地空通话语音增强方法[C]//第一届空中交通管理系统技术学术年会论文集. 北京: 电子工业出版社, 2018.

[9] Fu S, Tsao Y, Lu X, *et al*. SNR-aware convolutional neural network modeling for speech enhancement [C]//Proceedings of The Conference of The International Speech Communication Association. San Francisco: IEEE, 2016.

[10] Sainath T N, Mohamed A, Kingsbury B, *et al*. Deep convolutional neural networks for LVCSR [C]//Proceedings of The International Conference on Acoustics, Speech, and Signal Processing. Vancouver, BC, Canada: IEEE, 2013.

[11] 姚礼垚, 熊浩, 钟依健, 等. 基于深度网络模型的牛脸检测算法比较[J]. 江苏大学学报: 自然科学版, 2019, 40: 197.

[12] Zhao M, Wang D, Zhang Z, *et al*. Music removal by convolutional denoising autoencoder in speech recognition [C]//Proceedings of The ASIA Pacific Signal and Information Processing Association Annual Summit and Conference. Hong Kong, China: IEEE, 2015.

[13] Hui L, Cai M, Guo C, *et al*. Convolutional maxout neural networks for speech separation [C] //Proceedings of the International Symposium on Signal Processing and Information Technology. Abu Dhabi, UAE: IEEE, 2015.

[14] Zhang X, Wang D. A deep ensemble learning method for monaural speech separation [J]. IEEE T Audio Speech, 2016, 24: 967.

[15] Alain C, Bernstein L J. Auditory scene analysis [J]. Music Perception: Interdiscip J, 2015, 33: 70.

[16] 邹领, 贺前华, 邝细超, 等. 基于设备噪声估计的录音设备源识别[J]. 吉林大学学报:工学版, 2017, 47: 274.

[17] Wang Y, Narayanan A, Wang D, *et al*. On training targets for supervised speech separation [J]. IEEE T Audio Speech, 2014, 22: 1849.

[18] Narayanan A, Wang D. Ideal ratio mask estimation using deep neural networks for robust speech recognition [C]//Proceedings of the International Conference on Acoustics, Speech, and Signal Processing. Vancouver, BC, Canada: IEEE, 2013.

[19] Xia B, Bao C. Wiener filtering based speech enhancement with weighted denoising auto-encoder and noise classification [J]. Speech Commun, 2014, 60: 13.

[20] Desai D, Joshi M. Speaker recognition using MFCC and hybrid model of VQ and GMM [J]. Ingénierie Des Systèmes D'information, 2014, 13: 53.

[21] Bharti R, Bansal P. Real time speaker recognition system using MFCC and vector quantization technique [J]. Int J Comput Appl, 2015, 117: 25.

[22] Seltzer M L, Yu D, Wang Y, *et al*. An investigation of deep neural networks for noise robust speech recognition [C]//Proceedings of the International Conference on Acoustics, Speech, and Signal processing. Vancouver, BC, Canada: IEEE, 2013.

[23] Lu X, Tsao Y, Matsuda S, *et al*. Ensemble modeling of denoising autoencoder for speech spectrum restoration [C]//Proceedings of the Conference of the International Speech Communication Association. Singapore: IEEE, 2014.

[24] 郭东岳, 高登峰, 杨波, 等. 基于 CNN 的空管地空通话自动切分[C]//第一届空中交通管理系统技术学术年会论文集. 北京: 电子工业出版社, 2018.

[25] Jensen J, Taal C H. An algorithm for predicting the intelligibility of speech masked by modulated noise maskers [J]. IEEE T Audio Speech, 2016, 24: 2009.

引用本文格式:

中 文: 高登峰, 杨波, 刘洪, 等. 多特征全卷积网络的地空通话语音增强方法[J]. 四川大学学报: 自然科学版, 2020, 57: 289.

英 文: Gao D F, Yang B, Liu H, *et al*. A method of multi-featured full convolutional neural network based on speech enhancement in air-ground voice communication[J]. J Sichuan Univ: Nat Sci Ed, 2020, 57: 289.