引用格式：康峥, 黄志华, 赖惠成. 基于深度压缩感知的语音增强模型[J]. 声学技术, 2022, **41**(6): 862-870. [KANG Zheng, HUANG Zhihua, LAI Huicheng. Speech enhancement model based on deep compressed sensing[J].Technical Acoustics, 2022, **41**(6): 862-870.] **DOI**: 10.16300/j.cnki.1000-3630.2022.06.011

基于深度压缩感知的语音增强模型

康 峥，黄志华，赖惠成
(新疆大学信息科学与工程学院，新疆信号检测与处理重点实验室，新疆乌鲁木齐 830017)

摘要：随着压缩感知的深入研究，压缩感知在语音增强方面的应用也备受关注。针对传统压缩感知语音增强算法中存在的不足，将压缩感知与深度学习结合构建名为基于深度压缩感知的语音增强模型(Speech Enhancement based on Deep Compressed Sensing, SEDCS)。基于压缩感知原理使用编解码模型代替压缩感知中语音信号稀疏过程，使用卷积神经网络代替测量矩阵实现语音信号观测降维过程，通过联合训练的方式实现语音增强。实验结果表明：该模型能够完成语音增强任务，并且与现有的压缩感知语音增强算法相比，该模型能取得较好的语音增强效果；相比利用深度学习的语音增强算法，该模型虽性能一般，但在模型泛化性能和测试阶段的增强时间效率上有一定提升。

关键词：语音增强；压缩感知；深度学习；卷积神经网络
中图分类号：TN912.35 **文献标志码：**A **文章编号：**1000-3630(2022)-06-0862-09

Speech enhancement model based on deep compressed sensing

KANG Zheng, HUANG Zhihua, LAI Huicheng
(*School of Information Science and Engineering, Key Laboratory of Signal Detection and Processing in Xinjiang Uygur Autonomous Region, Xinjiang University, Urumqi* 830017*, Xinjiang, China*)

Abstract: With the further research of compressed sensing, the application of compressed sensing in speech enhancement has attracted much attention. Aiming at the shortcomings of traditional compressed sensing speech enhancement algorithms, a speech enhancement model based on deep compressed sensing (SEDCS) is built by combining compressed sensing and deep learning. Based on the principle of compressed sensing, the codec model is used to replace the sparse process of speech in compressed sensing, and the convolutional neural network is used to replace the measurement matrix to realize the measurement and dimension reduction of speech. The speech enhancement of the model is obtained by jointly training. The experimental results show that the proposed model can complete the speech enhancement task and achieve good speech enhancement effect compared with the existing compressed sensing speech enhancement algorithm. Compared with the speech enhancement algorithm using deep learning, the performance of the model is general, but it is improved in the model generalization ability and the enhancement time efficiency in the test stage.

Key words: speech enhancement; compressed sensing; deep learning; convolutional neural network

0 引 言

语音增强的目的是提高被噪声所干扰的语音质量与可懂度[1]。目前，语音增强在电话通信、助听设备以及语音识别等领域应用广泛。语音增强的传统方法有谱减法、子空间法、维纳滤波法等[1]。这些算法一般都基于特定的假设，如噪声是平稳的，但对于低信噪比和非平稳噪声情况下语音增强的效果较差。

随着深度学习的发展，深度神经网络被用于构建语音增强模型，以解决传统语音增强算法中对非平稳噪声增强效果差的问题。2014 年 Goodfellow 等[2]提出生成对抗网络(Generative Adversarial Nets, GAN)并证明了能够通过 GAN 生成图像样本。Pascual 等[3]将 GAN 应用于语音增强(Speech Enhancement Generative Adversarial Network, SEGAN)，实现了语音信号端到端快速增强，为语音增强提供了新思路。Stoller 等[4]将 U-Net 网络应用于声源分离(Wave-U-Net)，Macartney 等[5]利用 Wave-U-Net 从带噪语音信号中分离出噪声信号，实现了语音增强，其效果优于维纳算法和 SEGAN，并吸引许多学者对 Wave-U-Net 模型结构进行改进且取得了一定成果[6-7]。与传统语音增强算法相比，基于深度学

收稿日期： 2021-03-25; **修回日期：** 2021-06-30
基金项目： 新疆维吾尔自治区自然科学基金项目(2017D01C044)，国家科技部重点研发项目子课题(2018YFC0823402)。
作者简介： 康峥(1996—)，男，山西忻州人，硕士研究生，研究方向为语音增强。
通信作者： 黄志华，E-mail: zhhuang@xju.edu.cn

习的算法凭借强大的学习能力可以适应各种噪声类型，克服传统算法所要求的前提条件和低信噪比时去噪性能差的问题，且通过模型训练可有效保留语音特征信息，取得较好的语音增强结果。但基于深度学习的语音增强模型对数据集内噪声的去噪效果较好，而对集外噪声的去噪效果较差。

2006 年 Donoho 等[8]提出压缩感知(Compressed Sensing, CS)理论，在信号处理领域引起了研究热潮，并在图像处理、无线传感领域得到了迅速应用。近几年，CS 在语音增强领域的应用也备受关注。Sreenivas 等[9]对语音信号的稀疏表示进行了探索，证明了 CS 在语音重构上的可行性。随后许多学者对语音信号的稀疏基进行研究并取得了一定进展[10-12]。基于 CS 的语音增强算法可以解决传统算法中非平稳噪声去除较差的问题[10-12]，但是 CS 要求语音信号必须是稀疏的，即使用少量数据表示目标语音信号，该过程可能会造成原始语音有效信息丢失，从而降低重构语音质量。

最近，CS 与深度学习结合受到了广泛关注，相关学者们利用深度神经网络实现传统 CS 中的信号稀疏、观测降维和信号重构过程[13-15]。Bora 等提出利用生成模型代替信号稀疏过程(Compressed Sensing using Generative Models, CSGM)[13]，实现了图像重构，但该模型较复杂且训练缓慢。谷歌 DeepMind 提出了深度压缩感知模型(Deep Compressed Sensing, DeepCS)[15]，相比 CSGM 重构质量更好且速度更快。本文提出了语音增强的 DeepCS 模型框架，构建基于深度压缩感知的语音增强模型(Speech Enhancement based on Deep Compressed Sensing, SEDCS)，在增强效果和测试阶段的时间效率上取得了一定的效果。

1　结合深度学习的压缩感知模型简介

1.1　压缩感知原理

CS 在采样的同时压缩信号，可用较少的数据重构原始信号[8]。设信号为 $\boldsymbol{x}\in R^N$，可通过正交稀疏基矩阵 $\boldsymbol{\Psi}$ 获得信号的稀疏表示，即：

$$\boldsymbol{x}=\sum_{i=1}^{N}\boldsymbol{\psi}_i\boldsymbol{\alpha}_i=\boldsymbol{\Psi A} \tag{1}$$

其中：$\boldsymbol{A}$ 是只有 k 个非零元素的稀疏表示矩阵。信号稀疏处理之后，信号降维获得观测向量 $\boldsymbol{m}$ 为

$$\boldsymbol{m}=\boldsymbol{\Phi x}=\boldsymbol{\Phi\Psi A} \tag{2}$$

其中：$\boldsymbol{\Phi}$ 为 $M\times N$ 维测量矩阵（$M<<N$），可以选择随机高斯矩阵等作为测量矩阵，$\boldsymbol{\Phi}$ 与 $\boldsymbol{\Psi}$ 不相关，且 $\boldsymbol{\Phi}$ 与 $\boldsymbol{\Psi}$ 相乘后获得的矩阵满足有限等距性质(Restricted Isometry Property, RIP)，即：

$$\left(1-\delta_k\right)\|\boldsymbol{A}\|_2^2\leqslant\|\boldsymbol{\Phi\Psi A}\|_2^2\leqslant\left(1+\delta_k\right)\|\boldsymbol{A}\|_2^2 \tag{3}$$

其中，$\delta_k\in(0,1)$ 是等距常数。信号重构是一个最优化问题，即：

$$\hat{\boldsymbol{x}}=\underset{x}{\operatorname{argmin}}\|\boldsymbol{x}\|_0,\quad s.t.\boldsymbol{m}=\boldsymbol{\Phi x} \tag{4}$$

其中，$\|\cdot\|_0$ 表示 l_0 范数，$\hat{\boldsymbol{x}}$ 表示获得的重构信号。由于式(4)的求解为 N-P 难问题，因此将其简化为求解 l_1 范数问题，即：

$$\hat{\boldsymbol{x}}=\underset{x}{\operatorname{argmin}}\|\boldsymbol{x}\|_1,\quad s.t.\boldsymbol{m}=\boldsymbol{\Phi x} \tag{5}$$

其中，$\|\cdot\|_1$ 表示 l_1 范数。重构信号常用优化算法为正交匹配追踪(Orthogonal Matching Pursuit, OMP)算法[16]，而 CS 只需重构稀疏表示矩阵，再通过逆变换即可得到重构信号 $\hat{\boldsymbol{x}}$，因此该算法主要通过迭代优化获得信号的稀疏表示矩阵：

$$\hat{\boldsymbol{A}}=\underset{A}{\operatorname{argmin}}\left\|(\boldsymbol{\Phi\Psi})\boldsymbol{A}-\boldsymbol{m}_t\right\|_2\quad s.t.\|\boldsymbol{A}\|_0\leqslant\mathrm{k} \tag{6}$$

其中，$\boldsymbol{A}$、$\hat{\boldsymbol{A}}$ 分别表示初始稀疏表示矩阵以及更新后的稀疏表示矩阵，$\boldsymbol{m}_t$ 表示初始观测向量。通过迭代可获得 $\hat{\boldsymbol{A}}$，迭代次数由稀疏度 k 决定。最后通过逆变换即可得到重构信号。

1.2　结合深度神经网络的压缩感知模型

信号在稀疏化过程中可能造成信息丢失，为解决此问题，Bora 等提出了 CSGM 模型[13]，并定义了损失函数：

$$L(\boldsymbol{z})=\left\|\boldsymbol{\Phi}G_\theta(\boldsymbol{z})-\boldsymbol{m}\right\|_2^2 \tag{7}$$

其中，$\boldsymbol{z}$ 表示潜在随机输入，G_θ 表示参数为 θ 的生成模型，$\boldsymbol{\Phi}$ 表示测量矩阵，$\boldsymbol{m}$ 表示原始信号通过测量矩阵获得的观测信号。对 $\boldsymbol{z}$ 进行优化，如果优化后的随机输入 $\hat{\boldsymbol{z}}$ 使损失函数最小，那么 $\hat{\boldsymbol{x}}=G_\theta(\hat{\boldsymbol{z}})$ 即为重构信号。但该模型结构复杂，重构速率仍有待提升。

为进一步提升 CSGM 性能，谷歌 DeepMind 将 $\boldsymbol{\Phi}$ 替换为参数为 β 的神经网络 F_β，并定义了与 RIP 相关的损失函数以保证 F_β 实现信号观测降维过程[15]，以下分别是文献[15]中定义的 F_β 与 G_θ 的损失函数：

$$L_F=\mathrm{E}_{x_1,x_2}\left[\left(\left\|F_\beta(\boldsymbol{x}_1)-F_\beta(\boldsymbol{x}_2)\right\|_2-\left\|\boldsymbol{x}_1-\boldsymbol{x}_2\right\|_2\right)^2\right] \tag{8}$$

$$L_G=\mathrm{E}_{x\sim p_{\mathrm{data}}(x)}\left[\left\|F_\beta(G_\theta(\boldsymbol{z}))-F_\beta(\boldsymbol{x})\right\|_2^2\right] \tag{9}$$

其中，$\boldsymbol{x}_1$、$\boldsymbol{x}_2$ 分别表示从真实数据分布 $p_{\mathrm{data}}(\boldsymbol{x})$ 与生成数据分布 $G_\theta(\boldsymbol{z})$ 采样的数据。DeepCS 在重构速度

与重构质量上均优于传统CS以及CSGM。

2 基于深度压缩感知的语音增强模型

2.1 模型框架

受DeepCS模型[15]启发，本文将基于CS的语音增强算法中语音信号稀疏过程用生成模型代替，即不再需要考虑对语音信号进行稀疏，测量矩阵用卷积神经网络代替，称为测量模型。通过模型训练，直接在时域去除噪声信号恢复出干净的语音信号。

设语音信号表示为$\boldsymbol{x}$，噪声信号表示为$\boldsymbol{n}$，因此带噪语音信号$\boldsymbol{y}$可以表示为

$$\boldsymbol{y}=\boldsymbol{x}+\boldsymbol{n} \tag{10}$$

$\boldsymbol{y}$经过预处理后获得$\hat{\boldsymbol{y}}$，输入生成模型G_θ，可以获得生成语音信号$\hat{\boldsymbol{x}}=G_\theta(\hat{\boldsymbol{y}})$，然后$\boldsymbol{x}$与$\hat{\boldsymbol{x}}$通过测量模型获得各自观测信号。观测信号作为优化对象，当两条语音观测信号之间误差最小时，$\hat{\boldsymbol{x}}$即为重构的增强语音。

生成模型采用如图1所示的U-Net结构，由编码网络、跳跃连接(Skip)、注意力机制(Attention mechanism)以及解码网络四部分组成。编码网络由11个下采样模块组成，其中语音信号通过卷积操作获得语音特征信息，激活函数选择PReLU。解码网络由11个上采样模块组成，是编码网络的逆过程，采用解卷积恢复出与输入信号相同时间长度的语音信号。

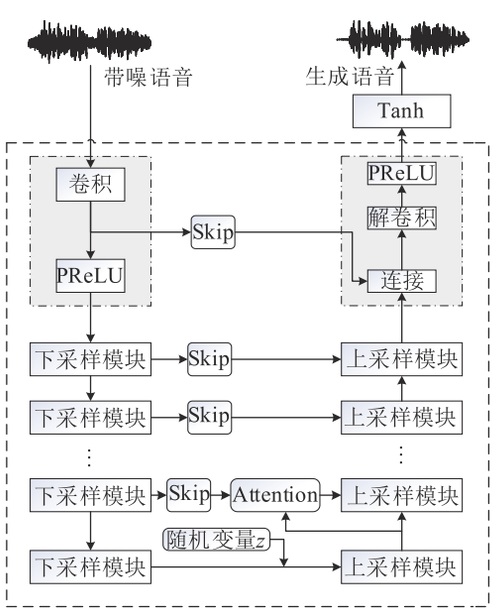


图1 增强语音的生成模型

Fig.1 Generative model of enhanced speech

为防止语音信号细节特征丢失，在编码网络和解码网络之间添加跳跃连接，在最后一层跳跃连接中添加注意力机制[6-7]，对语音信号特征进行修剪，去除无关语音特征。注意力机制结构图如图2所示。

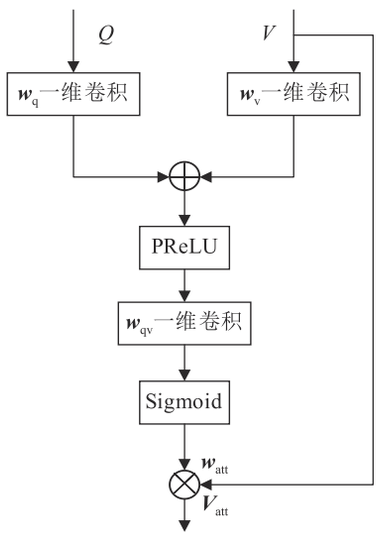


图2 注意力机制结构图

Fig.2 Structural diagram of attentional mechanism

注意力机制的输入为生成模型中下采样模块输出(V)和上采样模块输出(Q)，$\boldsymbol{w}_q$、$\boldsymbol{w}_v$以及$\boldsymbol{w}_{qv}$表示一维卷积的权重，$\boldsymbol{b}_q$、$\boldsymbol{b}_v$与$\boldsymbol{b}_{qv}$表示偏置，$\boldsymbol{w}_{att}$表示获得的注意力系数，与V相乘后获得通过注意力后的输出特征图$\boldsymbol{V}_{att}$。信号通过注意力机制的过程可以表示成求解式(11)的过程，其中P[·]和S{·}分别表示激活函数PReLU和Sigmoid：

$$\boldsymbol{V}_{att}=V\mathrm{S}\left\{\boldsymbol{w}_{qv}\mathrm{P}\left[\left(\boldsymbol{w}_q\boldsymbol{Q}+\boldsymbol{b}_q\right)+\left(\boldsymbol{w}_v\boldsymbol{V}+\boldsymbol{b}_v\right)\right]+\boldsymbol{b}_{qv}\right\} \tag{11}$$

测量模型为卷积神经网络(Convolutional Neural Network, CNN)，用于对语音信号进行观测降维，其输入是干净语音信号或生成语音信号，通过11个卷积层和1个全连接层获得指定观测维数的语音特征向量。模型框图如图3所示，其中S表示卷积步幅，每个卷积层的S均为2；C_i表示语音通过每一个卷积层后的通道数，角标i表示卷积层的索引号(即1-11)；F表示输入语音信号长度，通过卷积后，每层的特征向量维度为F/S^i。最后一层卷积结束后获得8×1 024维的语音特征图，将其通过线性平滑后输入全连接层，全连接层输出维数由观测维数决定。测量模型中激活函数选择LeakyReLU，为防止梯度爆炸和消失在激活函数之前添加批量归一化模块(Batch Normalization, BN)。

2.2 元学习与损失函数

2.2.1 元学习

元学习(Meta-Learning)是一种模型的学习策略，其目的是让模型学会学习[17]。面对不同的深度

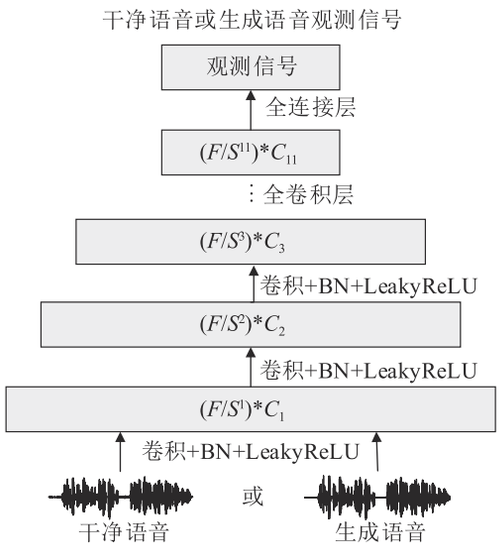


图3　卷积神经网络构成的测量模型

Fig.3　Measurement model composed of convolutional neural network

学习任务，元学习构建不同的损失函数预训练模型，再由各任务的损失函数共同构建一个总损失函数，通过优化总损失函数获得针对具体学习任务的模型。其最终目的是让模型在应对不同深度学习任务时收敛速度更快，模型收敛更好。

模型无关元学习(Model-Agnostic Meta-Learning, MAML)为元学习的一种应用拓展[18]，旨在通过有限的梯度更新获得更好的模型初始参数。如设具体学习任务的模型参数为$\boldsymbol{\Phi}$，不同深度学习任务的模型参数为μ_n，n表示所有参与元学习的任务中的第n个任务；因此，各任务的模型参数更新过程为

$$\hat{\mu}_n \leftarrow \mu_n - \eta \frac{\partial L(\mu_n)}{\partial \mu_n} \tag{12}$$

具体任务的损失函数设为$L(\boldsymbol{\Phi})$，与各任务损失函数的关系为

$$L(\boldsymbol{\Phi}) = \sum_n L(\hat{\mu}_n) \tag{13}$$

其中，η表示学习率。通过有限次数的梯度下降更新之后，将更新后的参数应用于具体学习任务的模型训练，梯度更新公式为

$$\hat{\boldsymbol{\Phi}} \leftarrow \boldsymbol{\Phi} - \eta \frac{\partial L(\boldsymbol{\Phi})}{\partial \boldsymbol{\Phi}} \tag{14}$$

其中，$\hat{\boldsymbol{\Phi}}$表示针对具体学习进行任务梯度下降更新之后的模型参数。

为减少训练周期、使模型收敛更快，本文所提模型采用与MAML类似的优化方式对生成模型进行优化训练。

将带噪语音信号看作输入模型之前的待优化参数，对其进行预处理优化，此时生成模型和测量模型不更新，优化次数可指定，优化过程为

$$\hat{\boldsymbol{y}} \leftarrow \hat{\boldsymbol{y}} - \alpha \frac{\partial \left[\left\| F_\beta(G_\theta(\boldsymbol{y})) - F_\beta(\boldsymbol{x}) \right\|_2^2 \right]}{\partial \boldsymbol{y}} \Bigg|_{\boldsymbol{y}=\hat{\boldsymbol{y}}} \tag{15}$$

其中，α表示优化学习率，$\boldsymbol{y}$与$\hat{\boldsymbol{y}}$分别表示原始带噪语音信号和优化后的带噪语音信号。然后将优化之后的带噪语音信号输入生成模型，再通过最小化损失函数更新模型。

2.2.2　模型损失函数

生成模型损失函数采用均方误差函数(Mean Square Error, MSE)，为了稳定模型训练，同时最小化生成语音与干净语音之间的差距，添加了L_1正则化项，因此生成模型损失函数为

$$L_{\mathrm{G}} = \mathrm{E}_{x\sim p_{\mathrm{data}}(x)} \left[\left\| F_\beta(G_\theta(\hat{\boldsymbol{y}})) - F_\beta(\boldsymbol{x}) \right\|_2^2 \right] + \lambda \left\| (G_\theta(\hat{\boldsymbol{y}})) - \boldsymbol{x} \right\|_1 \tag{16}$$

其中，λ表示控制L_1正则化项对整个损失函数影响的超参数。测量模型损失函数选择DeepCS中基于RIP性质的损失函数即：

$$L_{\mathrm{F}} = \mathrm{E}_{\substack{\boldsymbol{x}\sim p_{\mathrm{data}}(\boldsymbol{x}) \\ \hat{\boldsymbol{x}}\sim G_\theta(\hat{\boldsymbol{y}})}} \left[\left(\left\| F_\beta(\boldsymbol{x}) - F_\beta(\hat{\boldsymbol{x}}) \right\|_2 - \left\| \boldsymbol{x} - \hat{\boldsymbol{x}} \right\|_2 \right)^2 \right] \tag{17}$$

本文所提方法采用两个模型联合训练的方式，即最优化总损失函数：$\min\left(L_{\mathrm{total}}\right) = L_{\mathrm{G}} + L_{\mathrm{F}}$。

2.3　模型算法

在训练模型时由于对带噪语音信号进行预处理优化，因此训练模型时算法时间复杂度为$\mathrm{O}(n^3)$。为更详细表述本文所提模型中压缩感知与深度学习如何结合，通过下列训练算法的伪代码总结概述：

输入：干净语音信号$\boldsymbol{x}$、带噪语音信号$\boldsymbol{y}$、带噪语音的优化次数T、每个epoch包括的batch数N、其他超参数、测量模型F_β、生成模型G_θ。

输出：增强语音信号$\hat{\boldsymbol{x}}$，步骤如下：

(1) 初始化模型参数；

(2) for i=1，…，N

(3) $\boldsymbol{y}$由G_θ获得生成语音信号$G_\theta(\boldsymbol{y})$；

(4) $\boldsymbol{x}$与$G_\theta(\boldsymbol{y})$分别由F_β获得观测信号；

(5) for i=1，…，T

(6) 根据式(15)优化$\boldsymbol{y}$，获得优化后带噪语音$\hat{\boldsymbol{y}}$；

(7) end for

(8) $\hat{\boldsymbol{y}}$由G_θ获得优化后生成语音信号$G_\theta(\hat{\boldsymbol{y}})$；

(9) $G_\theta(\hat{\boldsymbol{y}})$由$F_\beta$获得优化后观测信号；

(10) 根据式(16)计算生成模型损失：L_{G}；

(11) 根据式(17)计算测量模型损失：L_{F}；

(12) 最优化总损失：$\min(L_{\mathrm{total}}) = L_{\mathrm{G}} + L_{\mathrm{F}}$；

(13) 更新G_θ与F_β模型参数；

(14) end for

(15) 直到全部epoch结束，获得最优$\hat{\boldsymbol{x}}$。

3 实验过程与结果分析

3.1 数据集

本文选择由Valentini等[19]构建的公开数据集，该数据集中包含从Voice Bank语料库中选取的30个说话人[20]，其中28个说话人用于构建训练集，2个说话人用于构建测试集。带噪训练集中噪声类型一共有10种(其中2种人为环境噪声，8种从Demand噪声集中选取的环境噪声[21])，信噪比设为0、5、10、15 dB，因此共40种噪声条件，11 572条带噪语音。带噪测试集中噪声选择Demand噪声集中5种不同于训练集噪声的环境噪声，信噪比设为2.5、7.5、12.5、17.5 dB，因此共20种噪声条件，824条带噪语音。以上语音文件采样率均为16 kHz。

考虑到传统CS语音增强算法的局限性，本文构建一个新测试集用于评测本文所提算法与传统CS算法的性能。该测试集从Valentini等构建的包含2个说话人的干净语音测试集中随机选择60条语音，其中每个说话人有30条语音。噪声选择Noise92数据集的babble、white、volvo[22]。信噪比设为−5、0、5、10 dB，采样率为16 kHz，因此可获得12种与训练集完全不同的噪声条件，构建了720条带噪语音。

3.2 实验设置

3.2.1 基线模型

为评测本文所提模型的可行性及有效性，分别从深度学习算法、传统算法和传统CS算法等角度进行对比评估。因此本文基线模型选择由Pascual等提出的SEGAN[3]、维纳语音增强算法[1]以及基于字典学习的传统CS语音增强算法。本文对这三个基线模型进行复现，SEGAN模型的参数全部遵照原文参数配置[3]。基于字典学习的传统CS算法基于文献[10]的思想选择K-SVD(K-means Singular Value Decomposition)算法学习字典，选择离散余弦变换基(Discrete Cosine Transform, DCT)对语音信号进行离散余弦变换，选择OMP算法重构稀疏信号，在稀疏域完成语音去噪，通过逆离散余弦变换获得去噪后语音。

3.2.2 评价指标

本文用于评测语音质量的客观指标包括：感知语音质量评估(Perceptual Evaluation of Speech Quality, PESQ)[23]、平均意见得分(Mean Opinion Score, MOS)[24]以及分段信噪比(Segmental SNR, SSNR)。MOS包括：针对语音信号失真的平均意见得分CSIG、背景噪声干扰性平均意见得分CBAK、总体增强效果平均意见得分COVL。PESQ取值在−0.5~4.5之间，CSIG、CBAK、COVL取值均在1~5之间，这5个指标得分越大语音质量越好。评测语音可懂度采用短时客观可懂度(Short-Time Objective Intelligibility, STOI)，取值在0~1之间，得分越高表示语音可懂度越高。

3.2.3 模型参数

本文提出的模型要对带噪语音进行预处理优化，优化学习率α设为0.01，优化次数T设为3，这使得模型可以更快收敛，因此模型训练epoch设为20，优化函数选择RMSprop，模型学习率设为0.000 2，batch_size设为16，两个模型的卷积核均为31，卷积步幅S均为2，生成模型损失函数中L_1正则化项的参数因子设为100。训练数据读取方式与SEGAN读取方式相同[3]，输入语音段长度F取16 384个采样点，帧叠设为50%，在输入模型之前对语音数据采用系数为0.95的预加重处理。

3.3 结果分析

3.3.1 模型可行性分析

表1为本文所提模型(SEDCS)与维纳滤波算法(Wiener)和SEGAN模型的对比结果，其中Noise表示公开原始带噪语音测试集各评价指标的得分结果，由于该公开测试集信噪比相对较高，因此其STOI得分偏高，m表示SEDCS采用的观测维数，测量模型采用CNN或多层感知机(Multilayer Perceptron, MLP)。SEDCS(CNN, m=50)和SEDCS(MLP, m=50)分别记录了当m取50时，采用两种不同测量模型的评测结果，发现当测量模型选择MLP时各评价指标得分均低于采用CNN的得分，去噪效果较差，原因是通过MLP获得的语音观测信号丢失了某些语音关键信息，因此后续实验测量模型均选择CNN。为比较采用不同观测维数时模型的性能，将m取值为25、50、75、100分别进行实验，并记录各自指标得分于表1中SEDCS(CNN, m=25)、SEDCS(CNN, m=50)、SEDCS(CNN, m=75)和SEDCS(CNN, m=100)中，发现m=100时各指标均低于其他模型，去噪效果最差。m=25时与取其他观测维数的模型结果相比，SSNR与CSIG得分较高，分别为6.01和3.35，语音失真不严重。m=50时评价指标PESQ、CBAK、COVL及STOI均高

于取其他观测维数时的模型结果，语音整体增强效果优于其他模型，因此后续实验观测维数均取 m=50。

本文在生成模型中添加注意力机制，评价指标得分如表 1 所示。根据各评价指标得分可以发现，SEDCS 模型(Atten,CNN, m=50)，相比没有添加注意力机制的模型其去噪性能有所提升，但是去噪效果仍不及 SEGAN 模型，原因是 SEGAN 模型采用对抗训练方式训练两个模型，从而使生成信号更接近原始数据分布，而 SEDCS 采用联合训练方式同时训练两个模型，因此拟合效果不及 SEGAN。与传统维纳算法相比，虽然 PESQ 得分低 0.03，但是其他各指标均高于维纳算法，这也证明将 CS 与深度学习结合应用于语音增强领域的可行性及有效性。

3.3.2　模型泛化性与压缩感知算法对比分析

由于基于字典学习的传统 CS 语音增强算法需要学习噪声字典，因此要求预先确定噪声类型。为与传统 CS 语音增强算法进行比较，同时测试模型泛化性能，本文构建了一个新测试集，噪声条件同 3.1 节。

表 2 记录了分别利用三种方法对新测试集增强之后的 PESQ 与 STOI 的得分情况。其中 Noise 所在行表示自行构建的带噪语音测试集的 PESQ 与 STOI 的得分结果，SEDCS 表示添加注意力机制、测量模型选用 CNN、m=50、且通过公开训练集训练完成的本文所提模型，SEGAN 表示通过公开训练集训练完成的基于 GAN 的语音增强模型[3]。

实验结果表明，在不同信噪比条件下 SEDCS 的 PESQ 得分均高于传统 CS 算法的 PESQ 得分，SEDCS 的 PESQ 平均得分为 1.62，传统 CS 算法的 PESQ 平均得分为 1.42，证明其总体增强效果优于传统 CS 算法。造成此现象的原因是传统 CS 算法在信号稀疏过程中丢失了语音有效信息，同时重构语音含有较多的残余噪声，从而造成重构语音质量较差。SEDCS 的 STOI 得分在信噪比为 10 dB 时为 0.882，高于传统 CS 算法，其他信噪比时，传统 CS 算法的 STOI 得分高于 SEDCS，造成此现象的原因是构建的测试集相对于 SEDCS 模型是未知的噪声条件，因此 STOI 得分略低于传统 CS。在不同信噪比条件下 SEDCS 的 PESQ 得分均高于 SEGAN。在信噪比为 10 dB 噪声条件下两个模型的 STOI 得分相等为 0.882，其他噪声条件下 SEDCS 的 STOI 得分均高于 SEGAN，由两个模型的平均得分可知，SEDCS 的泛化性能优于 SEGAN。

本文提出的 SEDCS 模型主要为解决传统 CS 语音增强算法中通过稀疏语音信号造成的有效信息丢失和重构语音信号残余噪声较多的问题。为进一步说明传统 CS 算法中信号稀疏重构对最终增强语音信号的影响，图 4 和图 5 分别给出了 SEDCS 和传统 CS 算法在 0 dB 的 volvo 噪声条件下、时域和频域中的噪声抑制情况对比图。为便于观察，在图中用红色标记出增强语音中丢失信息的位置，用黑色标记出增强语音中存在残余噪声的位置。

对比图 4(c)和图 5(c)可以发现，传统 CS 算法由于对语音信号进行稀疏，造成了增强语音信号细节信息的丢失，且存在较多残余噪声，因此造成增强语音质量的下降；通过 SEDCS 算法获得的增强语音的去噪效果较好，残余噪声较少，可有效解决传统 CS 算法所造成的信息丢失和存在较多残余噪声的问题。

3.3.3　模型语音增强时效对比

为评测 SEDCS 模型在测试阶段语音增强任务上的时效，进一步验证 SEDCS 模型的优势，本文从 Valentini 等构建的公开带噪语音测试集中选取不同时间长度的语音信号共 260 条，分别利用同一训练集下训练的 SEGAN 和添加注意力机制，测量模型选用 CNN，m=50 的 SEDCS 模型进行增强处理。

表 1　不同模型评价指标得分对比
Table 1　Comparison of evaluation index scores of different models

算法	PESQ	SSNR	CSIG	CBAK	COVL	STOI
Noise	1.97	1.68	3.34	2.44	2.63	0.916
Wiener	2.22	5.07	3.23	2.68	2.67	0.914
SEGAN	2.24	7.15	3.47	2.93	2.84	0.931
SEDCS(CNN, m=25)	2.02	6.01	3.35	2.73	2.65	0.915
SEDCS(CNN, m=50)	2.15	5.50	3.31	2.78	2.71	0.918
SEDCS(CNN, m=75)	2.06	5.95	3.32	2.72	2.68	0.913
SEDCS(CNN, m=100)	1.95	5.51	3.21	2.69	2.56	0.916
SEDCS(MLP, m=50)	1.89	3.01	3.11	2.45	2.46	0.912
SEDCS(Atten, CNN, m=50)	2.19	6.71	3.41	2.88	2.78	0.921

表2 SEDCS算法与传统CS算法的PESQ与STOI指标的得分对比

Table 2 Comparison of PESQ and STOI scores between SEDCS and traditional CS algorithm

指标	信噪比/dB	Noise	SEGAN	SEDCS	传统CS
PESQ	−5	1.14	1.20	1.29	1.24
	0	1.25	1.36	1.50	1.36
	5	1.43	1.60	1.74	1.50
	10	1.69	1.84	1.95	1.59
	平均值	1.38	1.50	1.62	1.42
STOI	−5	0.657	0.625	0.676	0.795
	0	0.767	0.730	0.776	0.842
	5	0.823	0.819	0.841	0.858
	10	0.876	0.882	0.882	0.867
	平均值	0.781	0.764	0.794	0.841

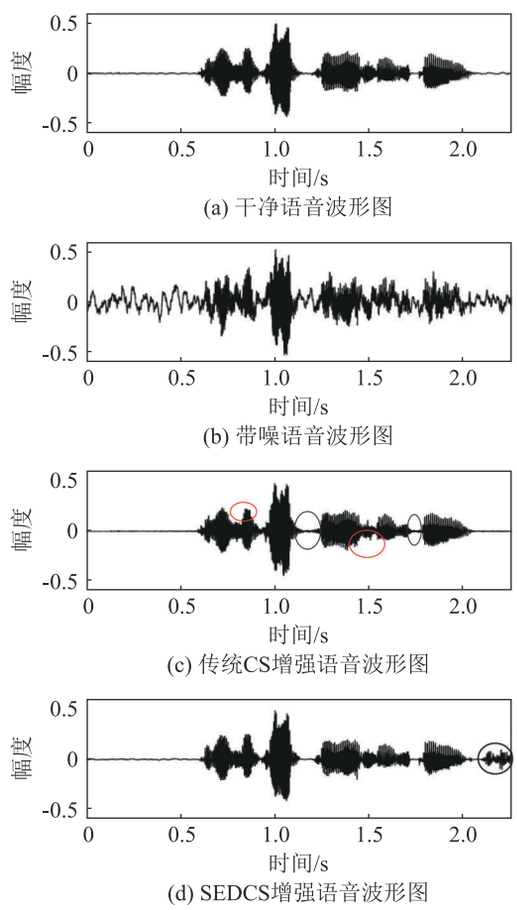


图4 SEDCS算法与传统CS算法抑制噪声后的语音波形对比

Fig.4 Comparison between speech waveforms after noise suppressed by SEDCS and traditional CS algorithms

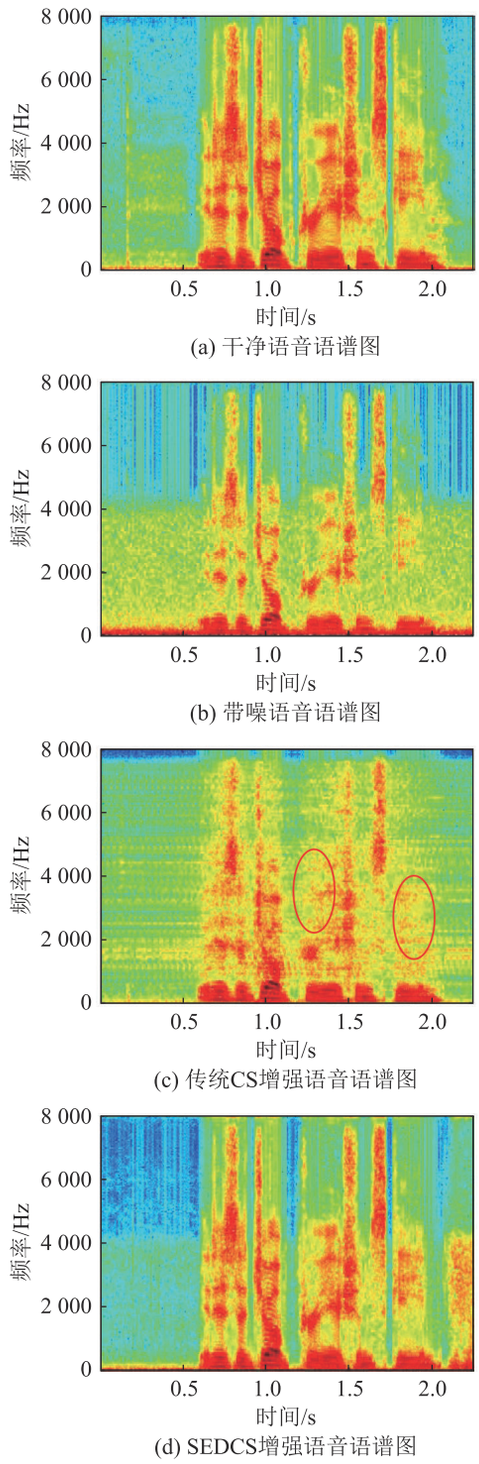


图5 SEDCS算法与传统CS算法抑制噪声后的语谱图对比

Fig.5 Comparison of between speech spectrograms after noise suppressed by SEDCS and traditional CS algorithms

增强所采用的硬件与软件配置分别是系统为Ubuntu 16.04、CPU为Inter(R) Xeon(R) Gold 5218 (2.30 GHz)、GPU为Nvidia Tesla P100(16 GB)的服务器以及在该服务器上通过Anaconda 3搭建的以CUDA-9.0加速的Tensorflow 1.12.0环境平台。

表3记录了测试阶段的语音增强算法耗时，表4记录了模型的参数量和测试阶段算法的浮点运算量。由表3和表4可以发现，在参数量级基本一致的情况下，SEDCS在公开测试集上增强效果虽然略低于SEGAN，但在测试阶段调用训练完成的模型增强语音时，增强速率优于SEGAN，耗时较短，能更快完成语音增强任务。原因是SEDCS在测试阶段的算法浮点运算量约为SEGAN的17.44%，因此测试阶段算法加载并运行保存的模型参数时，

SEDCS 算法的运行速率比 SEGAN 快，能更快调用模型完成增强任务。此外，表 4 中测试阶段算法的浮点运算量存在差距的原因是 SEGAN 在加载并运行模型参数时不仅计算了模型结构的浮点运算量，还计算了模型以外其他操作的浮点运算量(包括损失计算、梯度计算等操作，梯度计算大幅增加了浮点运算量)。而 SEDCS 在测试阶段只需对模型的相关操作进行计算，因此浮点运算量较低。

表 3　SEDCS 算法和 SEGAN 算法增强运算的耗时对比

Table 3　Comparison of times spent on enhancement operation between SEDCS and SEGAN algorithms

待处理语音时长/s	SEGAN 算法耗时/s	SEDCS 算法耗时/s
1	5.1	2.18
3	5.44	2.41
5	5.8	2.57
7	6.28	2.96
平均耗时/s	5.66	2.53

表 4　SEDCS 与 SEGAN 的模型参数量和测试阶段算法运算量对比

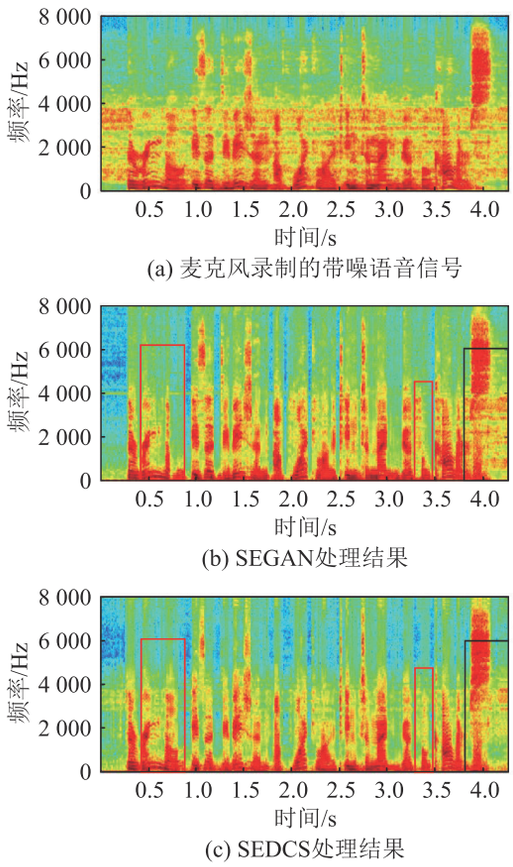
Table 4　Comparsion of model parameters and floating point operations of test code between SEDCS and SEGAN algorithms

模型算法	模型参数量	测试阶段算法浮点运算量
SEGAN	97.47×10^6	33.26×10^9
SEDCS	98.26×10^6	5.80×10^9

3.3.4　麦克风录制语音的去噪效果分析

本节对比分析了 SEDCS 与 SEGAN 处理由麦克风录制的带噪语音时的增强结果。录制环境为相对安静的实验室，录制设备采用笔记本电脑内置麦克风，并通过移动设备播放噪声来模拟现实环境噪声。录制的语音信息内容是：“The rainbow is a division of white light into many beautiful colors.”录制的带噪语音信号及 SEGAN、SEDCS 语音增强的语谱图如图 6 所示。

通过对比处理结果发现，SEDCS 可以对麦克风录制的带噪语音进行去噪处理。对比 SEGAN 的处理结果发现，由 SEDCS 增强的语音中残留的噪声更少(如图中黑色方框所标注位置)。此外，虽然经过 SEGAN 与 SEDCS 增强的结果中都存在一定的语音信息丢失，但是 SEDCS 的处理结果略优于 SEGAN(如图中红色方框所标注位置)。本实验也进一步说明了 SEDCS 具备一定的泛化性。



(a) 麦克风录制的带噪语音信号

(b) SEGAN处理结果

(c) SEDCS处理结果

图 6　SEDCS 与 SEGAN 语音增强处理前后的麦克风录制的带噪语音的语谱图对比

Fig.6　Comparison of noisy speech spectrogram recorded by microphone before and after speech enhancement processed by SEDCS and SEGAN algorithms

4　结 论

本文提出了一种基于 CS 与深度学习相结合的语音增强模型 SEDCS。使用编解码生成模型代替 CS 中的信号稀疏过程，使用深度卷积神经网络代替 CS 中观测降维过程，采用两个模型联合训练的方式，获得增强之后的语音信号。本文将 CS 与深度学习相结合的方法用于语音增强，实验结果表明，SEDCS 虽然增强效果一般，但是通过深度学习解决了 CS 中语音信号稀疏和重构问题，且具有较好的泛化性能。此外，通过计算测试阶段的算法浮点运算量，发现本文算法在测试阶段的浮点运算量较小，有助于提高语音增强的时间效率。通过对麦克风录制语音的去噪效果分析发现，本文，取得较好结果，SEDCS 的处理结果略优于 SEGAN，进一步说明本文模型具备一定泛化性。

在今后的工作中，将针对模型降噪性能、运行代价以及算法的实时性，进一步对生成模型和测量模型进行研究。在有效提取语音特征的前提下，研究通过更优的编解码方式(如扩张卷积、线性插值等)或训练方式(如对抗训练、强化学习等)提升去噪效果。在保证去噪性能的前提下，将尝试通过权值剪枝、减少网络层数或采用更高效的预处理方法降

低模型运行代价，同时尽可能提升模型测试阶段的实时性能。以上改进思路将是后续研究基于深度压缩感知的语音增强模型的主要研究方向。

参 考 文 献

[1] LOIZOU P C. Speech Enhancement[M]. BoCa Raton: CRC Press, 2013.

[2] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[EB/OL]. 2014: arXiv: 1406.2661[stat.ML]. https://arxiv.org/abs/1406.2661

[3] PASCUAL S, BONAFONTE A, SERRÀ J. SEGAN: speech enhancement generative adversarial network[C]//Interspeech 2017. ISCA: ISCA, 2017: 3642-3646..

[4] STOLLER D, EWERT S, DIXON S. Wave-U-net: a multi-scale neural network for end-to-end audio source separation [EB/OL]. 2018: arXiv: 1806.03185[cs. SD]. https://arxiv. org/abs/1806.03185

[5] MACARTNEY C, WEYDE T. Improved speech enhancement with the wave-U-net[EB/OL]. 2018: arXiv: 1811.11307 [cs.SD]. https://arxiv.org/abs/1811.11307

[6] GIRI R, ISIK U, KRISHNASWAMY A. Attention wave-U-net for speech enhancement[C]//2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. New Paltz, NY, USA. IEEE, 2019: 249-253.

[7] DENG F, JIANG T, WANG X R, et al. NAAGN: noise-aware attention-gated network for speech enhancement[C]//Interspeech 2020. ISCA: ISCA, 2020: 2457-2461.

[8] DONOHO D L. Compressed sensing[J]. IEEE Transactions on Information Theory, 2006, **52**(4): 1289-1306.

[9] SREENIVAS T V, KLEIJN W B. Compressive sensing for sparsely excited speech signals[C]//2009 IEEE International Conference on Acoustics, Speech and Signal Processing. Taipei, Taiwan, China. IEEE, 2009: 4125-4128.

[10] SIGG C D, DIKK T, BUHMANN J M. Speech enhancement using generative dictionary learning[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2012, **20**(6): 1698-1712.

[11] WANG J C, LEE Y S, LIN C H, et al. Compressive sensing-based speech enhancement[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2016, **24**(11): 2122-2131.

[12] SRIDHAR K V, KISHORE KUMAR T. Performance evaluation of CS based speech enhancement using adaptive and sparse dictionaries[C]//2019 4th International Conference and Workshops on Recent Advances and Innovations in Engineering (ICRAIE). Kedah, Malaysia. IEEE, 2019: 1-7.

[13] BORA A, JALAL A, PRICE E, et al. Compressed sensing using generative models[C]//International Conference on Machine Learning. PMLR, 2017: 537-546.

[14] MOUSAVI A, BARANIUK R G. Learning to invert: signal recovery via deep convolutional networks[C]//2017 IEEE International Conference on Acoustics, Speech and Signal Processing. New Orleans, LA, USA. IEEE, 2017: 2272-2276.

[15] WU Y, ROSCA M, LILLICRAP T. Deep compressed sensing [EB/OL]. 2019: arXiv: 1905.06723[cs. LG]. https://arxiv. org/abs/1905.06723

[16] TROPP J A, GILBERT A C. Signal recovery from random measurements via orthogonal matching pursuit[J]. IEEE Transactions on Information Theory, 2007, **53**(12): 4655-4666.

[17] SCHMIDHUBER J. Evolutionary principles in self-referential learning, or on learning how to learn: the m-eta-meta-… hook [D]. Technische Universität München,1987.

[18] FINN C, ABBEEL P, LEVINE S. Model-agnostic meta-learning for fast adaptation of deep networks[C]//ICML'17: Proceedings of the 34th International Conference on Machine Learning - Volume 70.2017: 1126-1135.

[19] VALENTINI-BOTINHAO C, WANG X, TAKAKI S, et al. Investigating RNN-based speech enhancement methods for noise-robust Text-to-Speech[C]//9th ISCA Workshop on Speech Synthesis Workshop (SSW 9). ISCA: ISCA, 2016: 146-152.

[20] VEAUX C, YAMAGISHI J, KING S. The voice bank corpus: Design, collection and data analysis of a large regional accent speech database[C]//2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE). Gurgaon, India. IEEE, 2013: 1-4.

[21] THIEMANN J, ITO N, VINCENT E. The diverse environments multi-channel acoustic noise database: a database of multichannel environmental noise recordings[J]. The Journal of the Acoustical Society of America, 2013, **133**(5): 3591.

[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] Wideband extension to recommendation P.862 for the assessment of wideband telephone networks and speech codecs[J]. 2007.

[24] HU Y, LOIZOU P C. Evaluation of objective quality measures for speech enhancement[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2008, **16**(1): 229-238.