利用深度卷积神经网络提高未知噪声下的语音增强性能

袁文浩 [1] 孙文珠 [1] 夏 斌 [1] 欧世峰 [2]

摘 要 为了进一步提高基于深度学习的语音增强方法在未知噪声下的性能, 本文从神经网络的结构出发展开研究. 基于在时间与频率两个维度上, 语音和噪声信号的局部特征都具有强相关性的特点, 采用深度卷积神经网络 (Deep convolutional neural network, DCNN) 建模来表示含噪语音和纯净语音之间的复杂非线性关系. 通过设计有效的训练特征和训练目标, 并建立合理的网络结构, 提出了基于深度卷积神经网络的语音增强方法. 实验结果表明, 在未知噪声条件下, 本文方法相比基于深度神经网络 (Deep neural network, DNN) 的方法在语音质量和可懂度两种指标上都有明显提高.

关键词 语音增强, 深度卷积神经网络, 深度神经网络, 噪声

引用格式 袁文浩, 孙文珠, 夏斌, 欧世峰. 利用深度卷积神经网络提高未知噪声下的语音增强性能. 自动化学报, 2018, **44**(4): 751−759

DOI 10.16383/j.aas.2018.c170001

Improving Speech Enhancement in Unseen Noise Using Deep Convolutional Neural Network

YUAN Wen-Hao[1] SUN Wen-Zhu[1] XIA Bin[1] OU Shi-Feng[2]

Abstract In order to further improve the performance of speech enhancement method based on deep learning in unseen noise, this paper focuses on the architecture of neural network. Based on the strong correlation between local characteristics of speech and noise signals in time and frequency domains, a deep convolutional neural network (DCNN) model is used to represent the complex nonlinear relationship between noisy speech and clean speech. By designing effective training features and training target, and establishing reasonable network architecture, a speech enhancement method based on DCNN is proposed. Experimental results show that under the condition of unseen noise, the proposed method significantly outperforms the methods based on deep neural network (DNN) in terms of both speech quality and intelligibility.

Key words Speech enhancement, deep convolutional neural network (DCNN), deep neural network (DNN), noise

Citation Yuan Wen-Hao, Sun Wen-Zhu, Xia Bin, Ou Shi-Feng. Improving speech enhancement in unseen noise using deep convolutional neural network. *Acta Automatica Sinica*, 2018, **44**(4): 751−759

语音增强是噪声环境下语音信号处理的必要环节[1]. 传统的基于统计的语音增强方法一般通过假设语音和噪声服从某种分布或者具有某些特性来从含噪语音中估计纯净语音, 这些方法对于平稳噪声具有较好的处理效果, 但在高度非平稳噪声和低信噪比情况下其处理性能将会急剧恶化[2−5].

近年来, 深度学习成为了机器学习领域的研究热点, 深度神经网络 (Deep neural network, DNN) 在图像分类和语音识别领域的成功应用为解决复杂多变噪声环境下的语音增强问题提供了思路. 与其他机器学习方法相比, 深度神经网络具有更加强大的学习能力, 通过使用大量纯净语音和含噪语音样本数据进行模型的训练, 能够有效提高语音增强方法对不同噪声的适应能力, 相比传统有监督方法具有更强的泛化能力, 对没有经过训练的未知噪声也有比较好的处理效果. 基于深度神经网络的语音增强方法的有效性已在很多文献中得到证明, 文献 [6] 训练 DNN 作为一个二值分类器来估计含噪语音的 IBM (Ideal binary mask), 克服了基于核函数的机器学习方法对大规模数据存在的计算复杂度难题, 提高了对未知噪声的适应能力, 取得了优于传统方法的语音增强性能. 文献 [7] 采用更加有效的 IRM (Ideal ratio mask) 代替 IBM 作为训练目标, 并通过实验证明了相比其他方法, 基于深度神经网络的语音增强方法明显提高了增强语音的质量和可懂度. 不同于上述方法中使用的基于掩蔽的训练目标, Xu

收稿日期 2017-01-03 录用日期 2017-07-18
Manuscript received January 3, 2017; accepted July 18, 2017
国家自然科学基金 (61701286, 61473179), 山东省自然科学基金 (ZR2015FL003, ZR2014FM007, ZR2017MF047) 资助
Supported by National Natural Science Foundation of China (61701286, 61473179), Shandong Provincial Natural Science Foundation of China (ZR2015FL003, ZR2014FM007, ZR2017MF047)
本文责任编委 党建武
Recommended by Associate Editor DANG Jian-Wu
1. 山东理工大学计算机科学与技术学院 淄博 255000 2. 烟台大学光电信息科学技术学院 烟台 264005
1. College of Computer Science and Technology, Shandong University of Technology, Zibo 255000 2. Institute of Science and Technology for Opto-electronic Information, Yantai University, Yantai 264005

等将纯净语音的对数功率谱 (Logarithmic power spectra, LPS) 作为训练目标, 以含噪语音的对数功率谱作为训练特征, 通过训练 DNN 得到一个高度非线性的回归函数, 来建立含噪语音对数功率谱与纯净语音对数功率谱之间的映射关系[8]; 并在文献 [9] 中采用 Global variance equalization、Dropout training 和 Noise-aware training 三种策略进一步改善该方法, 使其在低信噪比、非平稳噪声环境下的语音增强性能相比传统方法有了显著提升. 为了在语音增强时充分考虑相位信息, 文献 [10] 提出了复数域的掩蔽目标 cIRM (Complex IRM), 通过同时估计掩蔽目标的实部和虚部, 相比使用其他训练目标进一步提高了语音增强性能.

除了设计不同的训练特征和训练目标, 提高未知噪声下语音增强性能的另外一种重要思路是提高训练集中噪声的多样性. 文献 [9, 11] 分别采用包含 104 类和 115 类噪声的训练集, 提高了 DNN 对未知噪声的处理能力; 文献 [12 − 13] 更是通过训练包含 10 000 种不同噪声的 DNN 来提高对未知噪声的泛化能力, 主客观实验结果表明采用大数据量的训练集能显著提高未知噪声下的语音可懂度. 另外, 与直接增加训练集噪声类型数量的方法不同, 文献 [14] 采用对有限种类的噪声施加不同的扰动项的方式来提高噪声特性的多样性, 实验结果表明该方法同样能有效提高 DNN 的泛化能力.

上述基于深度神经网络的语音增强方法尽管在训练目标的设计、训练特征的选择以及训练集的规模上各有不同, 但是它们所采用的网络结构均是全连接的 DNN. 为了进一步提高未知噪声下的语音增强性能, 本文考虑使用深度学习的另外一种重要的网络结构 — 深度卷积神经网络 (Deep convolutional neural network, DCNN) 来进行语音增强. 深度卷积神经网络在图像识别等分类任务上已经取得了巨大成功[15], 其在二维图像信号处理上相比 DNN 表现出了更好的性能. 语音和噪声信号在时域的相邻帧和频域的相邻频带之间都具有很强的相关性, 因此在基于深度神经网络的语音增强方法中, 为了充分考虑时域和频域的上下文关系, 一般采用相邻多帧的特征作为网络的输入, 这种矩阵形式的输入在时间和频率两个维度上的局部相关性与图像中相邻像素之间的相关性非常类似. 如图 1 和图 2 所示, 假设使用连续 5 帧的对数功率谱作为网络的输入, 当网络结构为全连接的 DNN 时, 由于其输入层只有一个维度, 因此要将包含时频结构信息的矩阵转换为向量作为输入; 而当网络结构为 DCNN 时, 则可以直接使用矩阵作为输入, 不破坏时频结构. 可见, 得益于 DCNN 在二维平面上的局部连接特性, 使其相比 DNN 能够更好地表达网络输入在时间和频率两个维度的内在联系, 因而在语音增强时能够更充分地利用语音和噪声信号的时频相关性. 另外, DCNN 通过权值共享极大减少了神经网络需要训练的参数的个数, 具有更好的泛化能力, 对未训练噪声理论上应该有更好的处理性能.

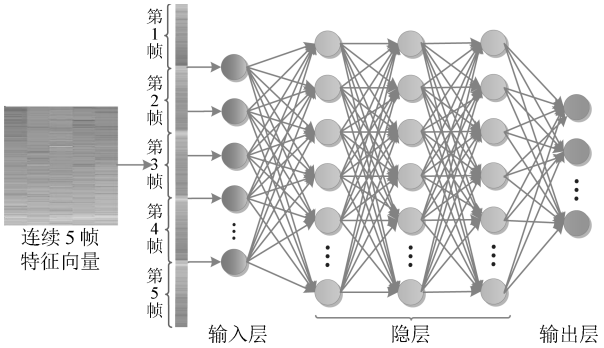


图 1 DNN 结构示意图

Fig. 1 Schematic diagram of DNN

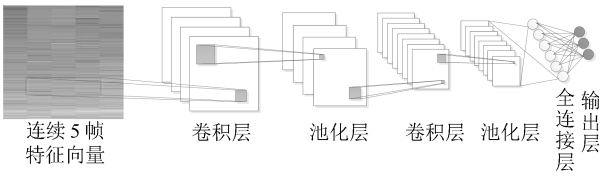


图 2 DCNN 结构示意图

Fig. 2 Schematic diagram of DCNN

实际上, 对于语音信号处理, CNN (Convolutional neural network) 以及 DCNN 已经在语音识别任务中得到成功应用, 取得了超越 DNN/HMM 系统的语音识别性能, 证明了其对于语音信号同样具有较好的特征提取能力[16−18], 文献 [19−23] 更是采用极深层的卷积神经网络显著提高了语音识别性能. 但是在语音识别任务中, DCNN 的最后一层一般采用 Softmax 来预测状态概率, 因此本质上也是一个分类问题; 而基于深度神经网络的语音增强方法一般将语音增强归结为回归问题进行解决, 因此传统的网络结构并不适合. 文献 [24] 以幅度谱向量作为训练特征和训练目标, 采用不包含全连接层的 FCNN (Fully convolutional neural network) 来进行语音增强, 虽然大幅度降低了训练参数的规模, 但是相比 DNN 并没有明显提高增强后语音的质量和可懂度; 文献 [25] 采用 CNN 对 LPS 特征进行建模, 通过同时学习纯净语音和信噪比, 研究了 SNR−aware 算法对语音增强性能的影响, 但是并没有对不同网络结构的语音增强性能进行深入分析. 为了提高语音增强性能, 特别是未知噪声下的语音增强性能, 本文通过对不同网络结构的语音增强性能进行对比与分析, 设计针对语音增强问题的合理 DCNN 网络结构, 提出基于深度卷积神经网络的语

音增强方法; 最后通过实验度量增强语音的质量和可懂度, 对方法在未知噪声下的语音增强性能进行客观评价.

1　训练特征与训练目标

假设含噪语音 y 由纯净语音 s 和加性噪声 d 组成,

$$y = s + d \tag{1}$$

语音增强的目的就是在已知 y 的条件下得到 s 的估计值 $\hat{s}$, 假设 y, s 和 $\hat{s}$ 在第 n 帧的短时傅里叶变换 (Short-time Fourier transform, STFT) 形式分别为 $Y_{n,k}\exp(\mathrm{j}\alpha_{n,k})$, $S_{n,k}\exp(\mathrm{j}\varphi_{n,k})$ 和 $\hat{S}_{n,k}\exp(\mathrm{j}\hat{\varphi}_{n,k})$, 其中 $k = 1, 2, \cdots, K$ 是频带序号, 忽略相位信息, 对第 n 帧的信号而言, STFT 域上的语音增强任务就是最小化如下的误差函数

$$Er = \sum_{k=1}^{K}\left(\hat{S}_{n,k} - S_{n,k}\right)^2 \tag{2}$$

令 S_n 和 $\hat{S}_n$ 分别表示纯净语音第 n 帧的幅度谱向量及其估计值, 该误差函数可以改写为

$$Er = \left\|\hat{S}_n - S_n\right\|_2^2 \tag{3}$$

基于深度学习的语音增强的基本思想可以描述为: 通过训练网络参数集合 θ 构造一个高度复杂的非线性函数 f_θ, 使得误差函数

$$Er = \left\|f_\theta\left(X_n\right) - S_n\right\|_2^2 \tag{4}$$

最小, 从而得到目标输出

$$\hat{S}_n = f_\theta\left(X_n\right) \tag{5}$$

其中

$$X_n = [Y_{n-N}, Y_{n-N+1}, \cdots, Y_n, \cdots, Y_{n+N-1}, Y_{n+N}] \tag{6}$$

表示第 n 帧的训练特征, 由以第 n 帧为中心的共 $(2N+1)$ 帧的含噪语音的幅度谱向量构成, $(2N+1)$ 即为输入窗长.

为了构造类似于图像处理 DCNN 的网络输入, 同时在保证时域语音信号重构简单的前提下提高网络性能, 我们采用对数运算对 X_n 和 S_n 的范围进行缩放, 设计如下的训练特征和训练目标

$$Z_n = \ln\left(X_n + 1\right) \tag{7}$$

$$T_n = \ln\left(S_n + 1\right) \tag{8}$$

其中, Z_n 和 T_n 是幅度谱的变换形式, 且其值不小于 0, 因此称其为非负对数幅度谱 (Nonnegative logarithmic amplitude spectra, NLAS).

DCNN 采用小批量梯度下降法进行训练, 本文使用的损失函数定义为

$$L(\theta) = \frac{1}{M}\sum_{n=1}^{M}\left\|f_\theta\left(Z_n\right) - T_n\right\|_2^2 \tag{9}$$

其中, M 代表网络训练所采用的 Mini-batch 的大小.

网络训练完成后, 在进行语音增强时, 对第 n 帧的纯净语音 s_n, 使用训练目标的估计值 $\hat{T}_n$ 与含噪语音第 n 帧的相位谱向量 α_n 进行时域信号的重构

$$\begin{aligned}\hat{s}_n =& \mathrm{ISTFT}(\hat{S}_n\exp(\mathrm{j}\hat{\varphi}_n)) = \\ & \mathrm{ISTFT}((\exp(\hat{T}_n) - 1)\exp(\mathrm{j}\alpha_n))\end{aligned} \tag{10}$$

$\hat{s}_n$ 即为增强后的语音信号.

2　网络结构

借鉴在图像识别中使用的典型 DCNN 的结构, 依据本文所采用的训练特征和训练目标, 构造如图 3 所示的 DCNN. 可见, 本文设计的网络结构与典型 DCNN 的最大不同在于最后几层全连接层的设计, 典型 DCNN 在全连接层后要经过一个 Softmax 层来计算分类结果, 而本文网络则是直接通过全连接层计算目标向量. 更深的网络结构、更多的节点数量或滤波器数量能够提高网络的性能, 但同时也增加了网络的复杂程度和训练难度, 对于本文实验, 依据训练集的数据规模, 通过权衡网络性能及训练难度之间的关系, 我们采用了包含 3 个卷积层和 2 个全连接层的网络结构, 其中全连接层的节点数量设为 1 024, 卷积层滤波器的个数除第一层为 64 外, 其余设为 128.

具体的网络结构设计如下:

1) 输入层

网络的输入是多帧非负对数幅度谱向量构成的特征矩阵.

2) 卷积层

本文网络包含 3 个卷积层, 第一层采用的卷积滤波器大小为 7×7, 其余两层的滤波器大小为 3×3, 步长均设为 1×1.

3) Batch normalization 层

在每个卷积层和激活函数层之间都有一个 Batch normalization 层.

4) 池化层

3 个激活函数层后是 3 个池化层, 均采用 Max-pooling, 滤波器大小 3×3, 步长为 2×2.

5) 全连接层

3 个卷积层之后是 2 个全连接层 (Fully connected) 和 2 个激活函数层.

6) 输出层

网络的最后一层是 129 个节点的全连接层, 对应 129 维的目标输出.

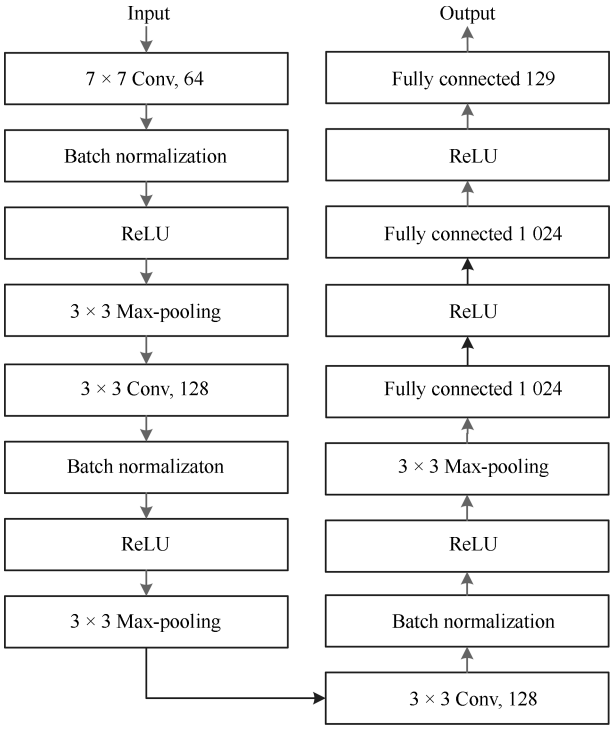


图 3 本文 DCNN 的结构框图

Fig. 3 Structure diagram of the proposed DCNN

3 实验与结果分析

3.1 实验配置

实验所用的纯净语音全部来自 TIMIT 语音数据库[26], 所用的噪声包含俄亥俄州立大学 Perception and Neurodynamics 实验室的 100 类噪声[27], 以及文献 [11] 中的 15 类噪声. 语音和噪声信号的采样频率均转换为 8 kHz, 短时傅里叶变换的帧长为 32 ms (256 点), 帧移为 16 ms (128 点), 相应的非负对数幅度谱特征向量和训练目标的维度为 129. 训练集由 100 000 段含噪语音 (约 80 小时) 构成, 使用 TIMIT 语音库的 Training 集的 4 620 段纯净语音和 115 类噪声按照 −5 dB、0 dB、5 dB、10 dB 和 15 dB 五种不同的信噪比合成得到. 每段含噪语音的具体合成方法如下: 每次从 4 620 段纯净语音中随机选取 1 段, 并从 115 类噪声中随机选取 1 类, 然后将该类噪声的随机截取片段按照从 5 种信噪比中随机选取的 1 种混入语音中. 测试集采用 TIMIT 语音库的 Core test 集的 192 段语音合成, 噪声数据采用来自 Noisex92 噪声库的与训练集噪声完全不同的 4 类未知噪声[28], 分别是 Factory2、Buccaneer1、Destroyer engine、HF channel 噪声. 对于每一类噪声, 将 192 段语音分别按照 −5 dB、0 dB 和 5 dB 的全局信噪比与该类噪声的随机截取片段进行混合, 4 类噪声合成的测试集总共包含 2 304 ($192 \times 3 \times 4$) 段含噪语音.

本文通过对增强语音进行客观评价来比较不同方法的语音增强性能, 主要采用 PESQ (Perceptual evaluation of speech quality) 作为指标来评价增强语音的质量[29], 并采用 STOI (Short time objective intelligibility) 作为指标来评价增强语音的可懂度[30]. PESQ 即语音质量感知评估是 ITU-T (国际电信联盟电信标准化部) 推荐的语音质量评估指标, 其得分范围为 $-0.5 \sim 4.5$, 越高的得分表示越高的语音质量. STOI 即短时客观可懂度, 则主要衡量语音的可懂度, 其得分范围为 $0 \sim 1$, 越高的得分表示语音具有越好的可懂度.

下面通过一系列实验对本文提出的 DCNN 的语音增强性能以及可能影响网络性能的关键因素进行分析.

3.2 DNN 与 DCNN 的比较

为了验证本文所提出的 DCNN 在语音增强中的有效性, 我们将其与 DNN 进行比较. 作为对比的 DNN 具有 5 个隐层, 每个隐层有 1 024 个节点, 激活函数为 ReLU; 为了防止过拟合, 提高泛化能力, 每个隐层后面均伴有一个 Dropout 层, Dropout 的比例为 0.2. DNN 和 DCNN 均采用式 (7) 定义的非负对数幅度谱作为训练目标, 并采用式 (8) 定义的训练特征作为网络的输入; 其中, 对于 DCNN, 为了适应其网络结构, 输入窗长设为 15 帧; 对于 DNN, 为了更好地进行对比, 其输入窗长分别设为与文献 [9] 相同的 11 帧 (DNN_11F), 以及与 DCNN 相同的 15 帧 (DNN_15F). mini-batch 的大小均为 128, 冲量因子均设为 0.9, 迭代次数均为 20. 本文的所有网络均使用微软的 Cognitive Toolkit 进行训练 [31].

首先通过比较 DNN 和 DCNN 的训练误差和测试误差来分析两种网络的性能, 图 4 给出了不同训练阶段所对应的训练集和测试集的均方误差, 可见, 两种 DNN 在训练集和测试集上的均方误差 (MSE) 都十分接近, 这表明两种 DNN 具有相似的语音增强性能; 而 DCNN 在训练集和测试集上的均方误差都要明显小于两种 DNN, 表明 DCNN 具有更好的语音增强性能.

为了进一步比较 DNN 和 DCNN 的语音增强性能, 我们对测试集含噪语音通过三种方法进行增强后得到的增强语音的平均语音质量和可懂度进行比较, 表 1 和表 2 分别给出了在 4 类不同噪声和 3 种不同信噪比下增强语音的平均 PESQ 和 STOI 得分, 并给出了未处理的含噪语音的平均 PESQ 和

STOI 得分作为对比. 可见, 通过采用多类噪声进行训练, 对于 4 种未经训练的噪声类型, 两种方法均能有效提升语音质量和可懂度, 并且在两种不同的指标中, DCNN 在不同噪声类型和不同信噪比条件下均取得了优于两种 DNN 的结果.

表 1　三种方法的平均 PESQ 得分
Table 1　The average PESQ score for three methods

噪声类型	信噪比 (dB)	含噪语音	DNN_11F	DNN_15F	DCNN
Factory2	−5	1.73	2.25	2.27	**2.33**
	0	2.07	2.57	2.58	**2.65**
	5	2.40	2.83	2.82	**2.89**
Buccaneer1	−5	1.36	1.88	1.92	**1.93**
	0	1.63	2.24	2.26	**2.27**
	5	1.95	2.54	2.54	**2.56**
Destroyer engine	−5	1.59	2.01	1.99	**2.15**
	0	1.81	2.27	2.26	**2.46**
	5	2.10	2.53	2.55	**2.76**
HF channel	−5	1.36	1.7	1.71	**2.03**
	0	1.58	2.04	2.06	**2.37**
	5	1.85	2.38	2.39	**2.65**

表 2　三种方法的平均 STOI 得分
Table 2　The average STOI score for three methods

噪声类型	信噪比 (dB)	含噪语音	DNN_11F	DNN_15F	DCNN
Factory2	−5	0.65	0.76	0.76	**0.78**
	0	0.76	0.85	0.84	**0.86**
	5	0.85	0.89	0.89	**0.91**
Buccaneer1	−5	0.51	0.66	0.66	**0.68**
	0	0.63	0.77	0.77	**0.78**
	5	0.75	0.85	0.85	**0.86**
Destroyer engine	−5	0.57	0.62	0.63	**0.70**
	0	0.69	0.75	0.75	**0.82**
	5	0.81	0.85	0.85	**0.90**
HF channel	−5	0.57	0.69	0.69	**0.73**
	0	0.69	0.78	0.79	**0.82**
	5	0.80	0.86	0.86	**0.88**

另外, 我们还在表 3 给出了含噪语音和增强语音的分段信噪比 (Segmental SNR, SegSNR), 分段信噪比同样是衡量语音质量的重要指标, 它比全局信噪比更接近实际的语音质量; 分段信噪比越大, 代表主观的语音质量越好. 与 PESQ 和 STOI 指标下的结果一致, 采用 DCNN 增强后的语音取得了最佳的分段信噪比. 值得注意的是, 两种 DNN 在三种指标下都取得了非常相近的结果, 这与文献 [9] 的描述是一致的.

表 3　三种方法的平均 SegSNR
Table 3　The average SegSNR for three methods

噪声类型	信噪比 (dB)	含噪语音 (dB)	DNN_11F (dB)	DNN_15F (dB)	DCNN (dB)
Factory2	−5	−6.90	−0.69	−0.59	**−0.05**
	0	−4.50	0.34	0.42	**0.95**
	5	−1.57	1.24	1.29	**1.80**
Buccaneer1	−5	−7.21	−1.52	−1.40	**−0.96**
	0	−4.90	−0.50	−0.39	**0.11**
	5	−2.03	0.46	0.53	**1.03**
Destroyer engine	−5	−7.15	−2.86	−2.81	**−2.16**
	0	−4.90	−1.37	−1.24	**−0.54**
	5	−1.91	0.04	0.21	**0.89**
HF channel	−5	−7.24	−1.13	−1.21	**0.35**
	0	−4.91	0.05	−0.02	**1.34**
	5	−2.09	1.04	1.02	**2.03**

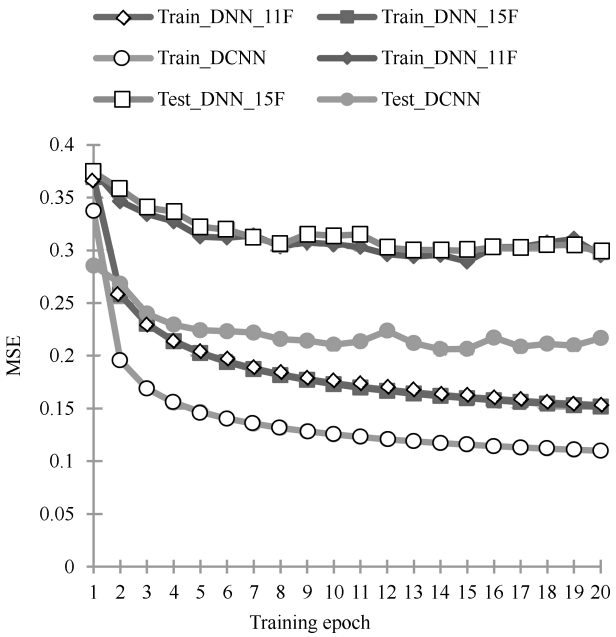


图 4　两种网络的训练误差和测试误差
Fig. 4　Training error and testing error of two networks

为了更加直观地比较两种方法的语音增强性能, 我们分别采用三种方法对一段含有 Factory2 噪声信噪比为 −5 dB 的含噪语音进行语音增强, 然后比较其增强语音的语谱图. 图 5 (a) 和 (b) 分别给出了含噪语音与其相应的纯净语音的语谱图, 图 5 (c) ∼ (e) 则分别给出了采用 DNN_11F、DNN_15F 以及 DCNN 增强后语音的语谱图. 可以看到, DCNN 增强后语音的残留噪声成分更少, 语音的纯净度更高, 其语谱图与纯净语音的语谱图更加接

近.

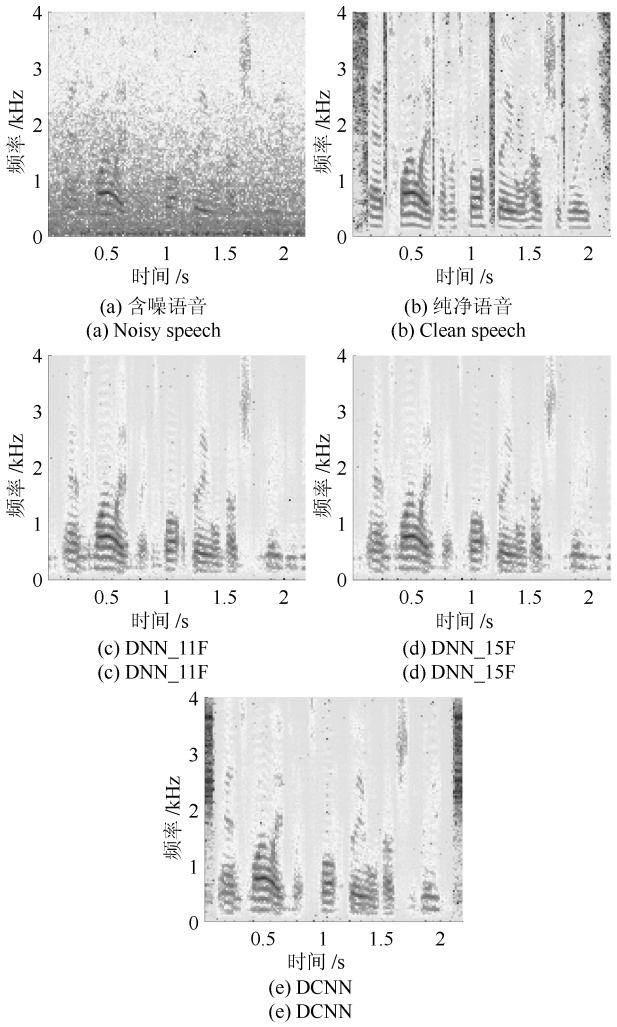


图 5 −5 dB 的 Factory2 噪声下的增强语音语谱图示例

Fig. 5 An example of spectrogram of enhanced speech under Factory2 noise at −5 dB SNR

3.3 卷积层数量的影响

对于图像和语音这种具有局部强相关性的信号, 卷积层具有很好的特征提取能力, 但是由于语音增强是一个回归问题, 网络的最后输出对应的是纯净语音的功率谱, 所以还需要通过全连接层来进行数据的拟合. 在本文使用的网络结构中, 不同的卷积层和全连接层的数量会带来网络性能的差别, 图 6 给出了不同网络配置下增强后语音的平均 PESQ 得分提升和平均 STOI 得分提升. 可见, 当网络包含 3 个卷积层和 2 个全连接层时, 在 3 种不同的信噪比下两种指标都得到了最高的提升值, 表明该网络结构具有最好的语音增强性能.

3.4 池化层的影响

Max-pooling 的直接作用是通过选取特征的局部最大值达到降低特征维度的目的. 在含噪语音功率谱的相邻时频单元中, 局部最大值一般含有语音成分, 而局部最小值一般为噪声成分, 传统的基于最小统计的噪声估计方法正是基于此原则. 因此, 池化层的存在将对时频单元起到一定的筛选作用, 能够通过筛掉局部较小值达到抑制噪声成分的目的.

为了检验池化层对于网络性能的影响, 我们将卷积层的步长设为 2, 并去掉池化层, 训练得到不含池化层的网络模型. 图 7 给出了不同信噪比下包含池化层 (Max-pooling) 和不含池化层 (No pooling) 的网络增强后语音的平均 PESQ 得分提升和平均 STOI 得分提升, 综合分析两种指标可知, 在较低信噪比的 −5 dB 和 0 dB 两种情况下, 包含池化层的网络的语音增强性能略好于不含池化层的网络.

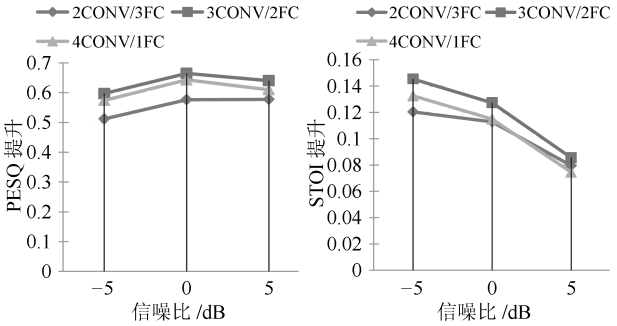


图 6 卷积层数量对网络性能的影响

Fig. 6 The influence of the number of convolutional layers on the network performance

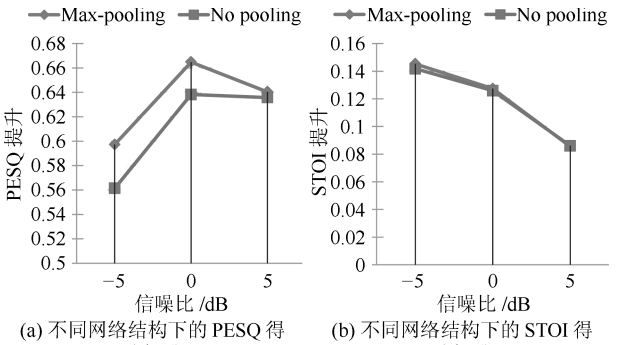


图 7 池化层对网络性能的影响

Fig. 7 The influence of the pooling layers on the network performance

通过对比两种网络增强后语音的语谱图来进一步观察池化层的影响, 图 8(a) 和图 8(b) 分别给出了一段含有 −5 dB 的 HF channel 噪声的含噪语音与其相应的纯净语音的语谱图, 图 8(c) 和图 8(d) 则分别给出了采用包含池化层和不含池化

层的网络增强后语音的语谱图. 由图 8 可见, 与上述分析一致, 包含池化层的网络增强后语音的残留噪声明显少于不含池化层网络增强后语音, 表明 Max-pooling 的存在确实能带来更好的噪声抑制效果.

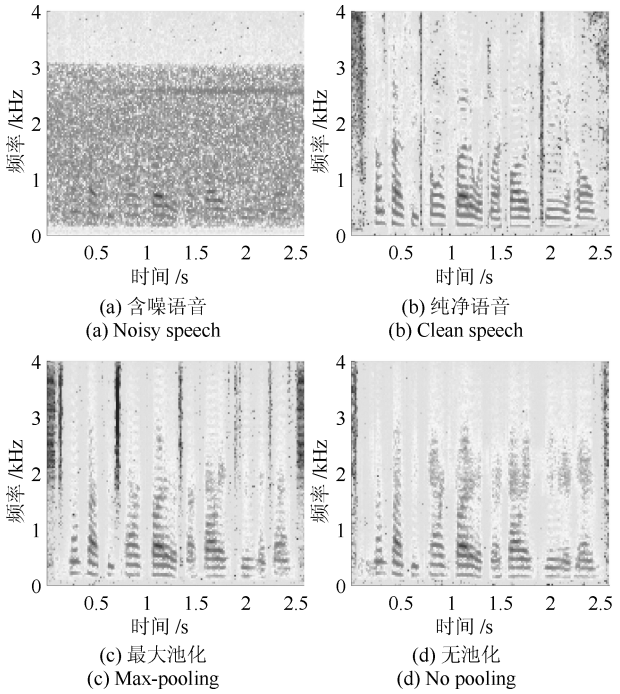


图 8　−5 dB 的 HF channel 噪声下的增强语音语谱图示例

Fig. 8　An example of spectrogram of enhanced speech under HF channel noise at −5 dB SNR

3.5　Batch normalization 层的影响

Batch normalization 是深度卷积神经网络中的常用技术, Batch normalization 层的引入往往可以加快收敛过程, 提升训练速度, 并能防止过拟合. 为了检验 Batch normalization 层对本文网络结构的影响, 我们去掉网络中的 Batch normalization 层, 训练得到不含 Batch normalization 层的网络模型. 图 9 给出了不同信噪比下包含 Batch normalization 层 (BN) 和不含 Batch normalization 层 (No BN) 的网络增强后语音的平均 PESQ 得分提升和平均 STOI 得分提升, 在两种指标下, 不包含 Batch normalization 层的网络模型都略好于包含 Batch normalization 层的网络模型, 表明 Batch normalization 层的引入并没有提升本文网络结构的语音增强性能. 可见, 对于本文相对简单的网络结构, Batch normalization 并没有明显的作用, 可以去掉.

3.6　LPS 与 NLAS 的比较

下面通过实验对文献 [9] 采用的 LPS 与本文采用的 NLAS 两种特征进行比较, 分别采用 DNN 和 DCNN 对两种特征进行训练. 其中, 训练 LPS 的 DNN (LPS-DNN) 与训练 NLAS 的 DNN (NLAS-DNN) 均为与前文相似的包含 5 个隐层的 DNN, 需要注意的是两种 DNN 采用的激活函数是 Sigmoid 函数, 因为在我们的实验中, 当训练特征为 LPS 时, 如果采用 ReLU 作为激活函数, 会造成训练过程不收敛; 训练 LPS 的 DCNN (LPS-DCNN) 与前文的 NLAS-DCNN 结构一致. 图 10 分别给出了 4 种测试集噪声在不同信噪比下采用 4 种方法增强后语音的平均 PESQ 和 STOI 得分. 可见, 在相同特征下, DCNN 的语音增强性能明显好于 DNN; 在相同的网络结构下, 采用 NLAS 特征训练得到的网络模型在 3 种不同信噪比下都取得了较好的语音可懂度, 并且在低信噪比 (−5 dB) 下取得了较好的语音质量, 表明 NLAS 特征能够更好地保留含噪语音中的语音成分, 更加适用于低信噪比下的语音增强.

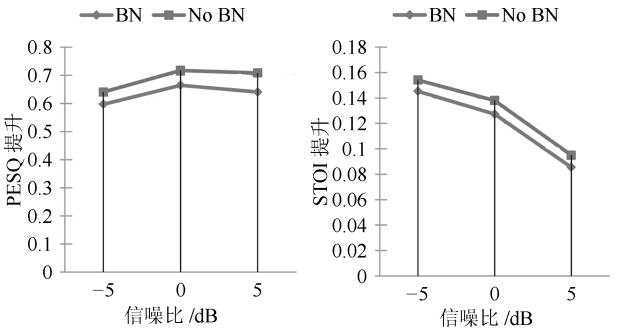


图 9　Batch normalization 层对网络性能的影响

Fig. 9　The influence of the batch normalization layers on the network performance

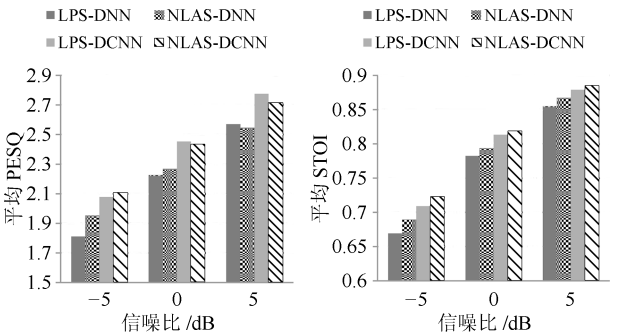
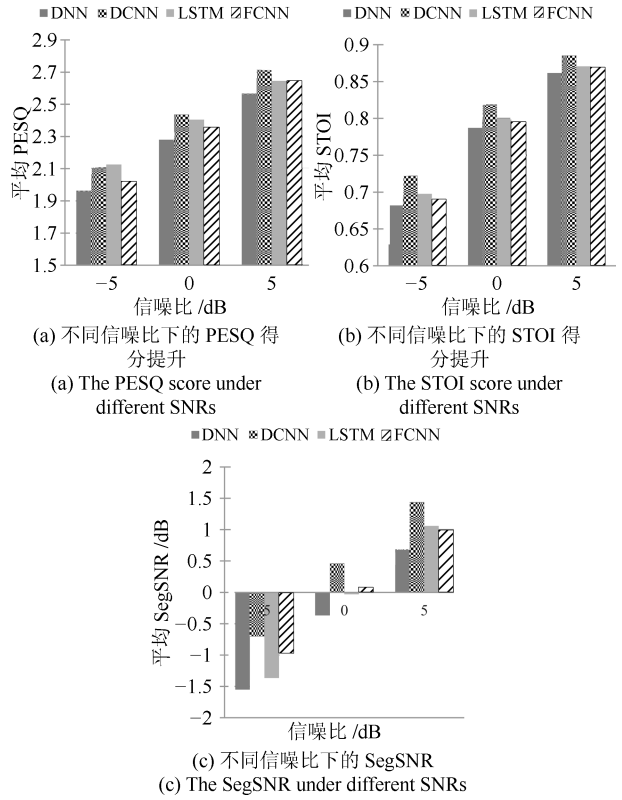


图 10　两种特征训练得到的 DNN 和 DCNN 的性能比较

Fig. 10　The performance comparisons for DNN and DCNN trained using two kinds of feature

3.7 与其他方法的比较

为了进一步验证本文 DCNN 的语音增强性能, 将其与 LSTM (Long-short term memory) 以及文献 [24] 中的 FCNN 进行比较. 其中 LSTM 包含 5 个隐层, Cell 维度为 256; FCNN 包含 16 个卷积层, 每层滤波器的个数分别为: 10, 12, 14, 15, 19, 21, 23, 25, 23, 21, 19, 15, 14, 12, 10, 1. 图 11 分别给出了各种方法增强后语音的平均 PESQ、平均 STOI 和平均 SegSNR, 同时给出 DNN 对应的结果作为对比. 通过综合分析 3 种指标可知, DCNN 取得了最佳的语音增强性能, LSTM 次之, FCNN 略好于 DNN.



(a) 不同信噪比下的 PESQ 得分提升
(a) The PESQ score under different SNRs
(b) 不同信噪比下的 STOI 得分提升
(b) The STOI score under different SNRs
(c) 不同信噪比下的 SegSNR
(c) The SegSNR under different SNRs

图 11 两种特征训练得到的 DNN 和 DCNN 的性能比较

Fig. 11 The performance comparisons for DNN and DCNN trained using two kinds of feature

4 结论

为了进一步提高未知噪声下的语音增强性能, 考虑 DCNN 相比 DNN 具有更好的局部特征表达能力, 能够更好地利用语音和噪声信号的时频相关性, 本文采用深度卷积神经网络建立回归模型来表达含噪语音和纯净语音之间的复杂非线性关系. 通过使用非负对数幅度谱作为训练特征和训练目标, 设计与训练了不同结构的 DCNN 并对其语音增强性能进行了比较, 得到了适合于语音增强问题的合理网络结构, 提出了基于深度卷积神经网络的语音增强方法. 实验结果表明, 在与 DNN 及其他方法的对比中, 本文提出的 DCNN 在测试集上取得了更小的误差, 表现出了更好的噪声抑制能力, 在各类噪声和各种信噪比条件下都显著提升了增强后语音的语音质量和可懂度, 进一步提高了未知噪声下的语音增强性能.

References

1 Loizou P C. *Speech Enhancement: Theory and Practice*. Florida: CRC Press, 2013.

2 Ephraim Y, Malah D. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1985, **33**(2): 443−445

3 Cohen I. Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging. *IEEE Transactions on speech and audio processing*, 2003, **11**(5): 466−475

4 Mohammadiha N, Smaragdis P, Leijon A. Supervised and unsupervised speech enhancement using nonnegative matrix factorization. *IEEE Transactions on Audio, Speech, and Language Processing*, 2013, **21**(10): 2140−2151

5 Liu Wen-Ju, Nie Shuai, Liang Shan, Zhang Xue-Liang. Deep learning based speech separation technology and its developments. *Acta Automatica Sinica*, 2016, **42**(6): 819−833 (刘文举, 聂帅, 梁山, 张学良. 基于深度学习语音分离技术的研究现状与进展. 自动化学报, 2016, **42**(6): 819−833)

6 Wang Y X, Wang D L. Towards scaling up classification-based speech separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 2013, **21**(7): 1381−1390

7 Wang Y X, Narayanan A, Wang D L. On training targets for supervised speech separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 2014, **22**(12): 1849−1858

8 Xu Y, Du J, Dai L R, Lee C H. An experimental study on speech enhancement based on deep neural networks. *IEEE Signal Processing Letters*, 2014, **21**(1): 65−68

9 Xu Y, Du J, Dai L R, Lee C H. A regression approach to speech enhancement based on deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2015, **23**(1): 7−19

10 Williamson D S, Wang Y X, Wang D L. Complex ratio masking for monaural speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2016, **24**(3): 483−492

11 Xu Y, Du J, Huang Z, Dai L R, Lee C H. Multi-objective learning and mask-based post-processing for deep neural network based speech enhancement. In: Proceedings of the 16th Annual Conference of the International Speech Communication Association. Dresden, Germany: ISCA, 2015. 1508−1512

12 Wang Y X, Chen J T, Wang D L. Deep Neural Network Based Supervised Speech Segregation Generalizes to Novel Noises Through Large-scale Training, Technical Report OSU-CISRC-3/15-TR02, Department of Computer Science and Engineering, The Ohio State University, Columbus, Ohio, USA, 2015

13 Chen J T, Wang Y X, Yoho S E, Wang D L, Healy E W. Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises. *The Journal of the Acoustical Society of America*, 2016, **139**(5): 2604−2612

14 Chen J T, Wang Y X, Wang D L. Noise perturbation for supervised speech separation. *Speech Communication*, 2016, **78**: 1−10

15 Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. In: Proceedings of the International Conference on Neural Information Processing Systems. Nevada, USA: Curran Associates Inc. 2012. 1097−1105

16 Abdel-Hamid O, Mohamed A, Jiang H, Penn G. Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition. In: Proceedings of the 2012 IEEE International Conference on Acoustics, Speech and Signal Processing. Kyoto, Japan: IEEE, 2012. 4277−4280

17 Abdel-Hamid O, Deng L, Yu D. Exploring convolutional neural network structures and optimization techniques for speech recognition. In: Proceedings of the 14th Annual Conference of the International Speech Communication Association. Lyon, France: ISCA, 2013. 3366−3370

18 Sainath T N, Kingsbury B, Saon G, Soltau H, Mohamed A R, Dahl G, Ramabhadran B. Deep convolutional neural networks for large-scale speech tasks. *Neural Networks*, 2015, **64**: 39−48

19 Qian Y M, Bi M X, Tan T, Yu K. Very deep convolutional neural networks for noise robust speech recognition. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 2016, **24**(12): 2263−2276

20 Bi M X, Qian Y M, Yu K. Very deep convolutional neural networks for LVCSR. In: Proceedings of the 16th Annual Conference of the International Speech Communication Association. Dresden, Germany: ISCA, 2015. 3259−3263

21 Qian Y, Woodland P C. Very deep convolutional neural networks for robust speech recognition. In: Proceedings of the 2016 IEEE Spoken Language Technology Workshop. San Juan, Puerto Rico: IEEE, 2016. 481−488

22 Sercu T, Puhrsch C, Kingsbury B, LeCun Y. Very deep multilingual convolutional neural networks for LVCSR. In: Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing. Shanghai, China: IEEE, 2016. 4955−4959

23 Sercu T, Goel V. Advances in very deep convolutional neural networks for LVCSR. In: Proceedings of the 16th Annual Conference of the International Speech Communication Association. California, USA: ISCA, 2016. 3429−3433

24 Park S R, Lee J. A fully convolutional neural network for speech enhancement. arXiv:1609.07132, 2016.

25 Fu S W, Tsao Y, Lu X. SNR-Aware convolutional neural network modeling for speech enhancement. In: Proceedings of the 17th Annual Conference of the International Speech Communication Association. San Francisco, USA: ISCA, 2016. 8−12

26 Garofolo J S, Lamel L F, Fisher W M, Fiscus J G, Pallett D S, Dahlgren N L, Zue V. TIMIT acoustic-phonetic continuous speech corpus. Linguistic Data Consortium, Philadelphia, 1993.

27 Hu G N. 100 nonspeech sounds [online], available: http://web.cse.ohio-state.edu/pnl/corpus/HuNonspeech/HuCorpus.html, April 20, 2004

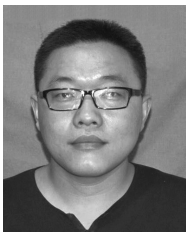
28 Varga A, Steeneken Herman J M. Assessment for automatic speech recognition: II. NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Communication*, 1993, **12**(3): 247−251

29 Beerends J G, Rix A W, Hollier M P, Hekstra A P. Perceptual evaluation of speech quality (PESQ) — a new method for speech quality assessment of telephone networks and codecs. In: Proceedings of the 2001 IEEE International Conference on Acoustics, Speech and Signal Processing. Utah, USA: IEEE, 2001. 749−752

30 Taal C H, Hendriks R C, Heusdens R, Jensen J. An algorithm for intelligibility prediction of time-frequency weighted noisy speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 2011, **19**(7): 2125−2136

31 Yu D, Eversole A, Seltzer M L, Yao K S, Huang Z H, Guenter B, Kuchaiev O, Zhang Y, Seide F, Wang H M, Droppo J, Zweig G, Rossbach C, Currey J, Gao J, May A, Peng B L, Stolcke A, Slaney M. An Introduction to Computational Networks and the Computational Network Toolkit, Technical Report, Tech. Rep. MSR, Microsoft Research, 2014.

袁文浩 博士, 山东理工大学计算机科学与技术学院讲师. 主要研究方向为语音信号处理, 语音增强. 本文通信作者.
E-mail: why_sdut@126.com
(**YUAN Wen-Hao** Ph. D., lecturer at the College of Computer Science and Technology, Shandong University of Technology. His research interest covers speech signal processing and speech enhancement. Corresponding author of this paper.)



孙文珠 博士, 山东理工大学计算机科学与技术学院讲师. 主要研究方向为多媒体信号传输, 视频编码.
E-mail: swz_lw@sina.com
(**SUN Wen-Zhu** Ph. D., lecturer at the College of Computer Science and Technology, Shandong University of Technology. His research interest covers multimedia signal processing and video coding.)

夏 斌 博士, 山东理工大学计算机科学与技术学院副教授. 主要研究方向为信号处理.
E-mail: xiabin@sdut.edu.cn
(**XIA Bin** Ph. D., associate professor at the College of Computer Science and Technology, Shandong University of Technology. His main research interest is signal processing.)

欧世峰 博士, 烟台大学光电信息科学技术学院副教授. 主要研究方向为语音信号处理, 盲信号处理.
E-mail: ousfeng@126.com
(**OU Shi-Feng** Ph. D., associate professor at the Institute of Science and Technology for Opto-electronic Information, Yantai University. His research interest covers speech signal processing and blind source separation.)