基于感知掩蔽深度神经网络的单通道语音增强方法

韩 伟 [1] 张雄伟 [1] 闵 刚 [1, 2] 张启业 [3]

摘 要 本文将心理声学掩蔽特性应用于基于深度神经网络 (Deep neural network, DNN) 的单通道语音增强任务中, 提出了一种具有感知掩蔽特性的 DNN 结构. 首先, 提出的 DNN 对带噪语音幅度谱特征进行训练并分别得到纯净语音和噪声的幅度谱估计. 其次, 利用估计的纯净语音幅度谱计算噪声掩蔽阈值. 然后, 将噪声掩蔽阈值和估计的噪声幅度谱联合计算得到一个感知增益函数. 最后, 利用感知增益函数从带噪语音幅度谱中估计出增强语音幅度谱. 在 TIMIT 数据库上, 对不同信噪比下的 20 种噪声进行的仿真实验表明, 无论噪声类型是否在语音的训练集中出现, 所提出的感知掩蔽 DNN 都能够在有效去除噪声的同时保持较小的语音失真, 增强效果明显优于常见的 DNN 增强方法以及 NMF (Nonnegative matrix factorization) 增强方法.

关键词 语音增强, 深度神经网络, 感知增益函数, 掩蔽阈值

引用格式 韩伟, 张雄伟, 闵刚, 张启业. 基于感知掩蔽深度神经网络的单通道语音增强方法. 自动化学报, 2017, **43**(2): 248−258

DOI 10.16383/j.aas.2017.c150719

A Single-channel Speech Enhancement Approach Based on Perceptual Masking Deep Neural Network

HAN Wei[1] ZHANG Xiong-Wei[1] MIN Gang[1, 2] ZHANG Qi-Ye[3]

Abstract A new deep neural network (DNN) is proposed for single-channel speech enhancement, which incorporates the perceptual masking properties of psychoacoustic models. Firstly, the proposed DNN is trained to learn both the clean speech magnitude spectrum and the noise magnitude spectrum from the noisy magnitude spectrum. Secondly, the estimated clean speech magnitude spectrum is used to calculate the noise masking threshold. Then, the noise masking threshold and the estimated noise magnitude spectrum are combined to calculate a perceptual gain function. Finally, the enhanced speech magnitude spectrum are obtained by jointly training the perceptual gain function and the noisy speech magnitude spectrum. Experimental results on TIMIT with 20 noise types at various SNR (signal-noise ratio) levels demonstrate that the proposed perceptual masking DNN can effectively remove the noise while maintaining small speech distortion, so as to obtain better performance than the common DNN methods and the NMF (nonnegative matrix factorization) method, no matter noise conditions are included in the training set or not.

Key words Speech enhancement, deep neural network, perceptual gain function, masking threshold

Citation Han Wei, Zhang Xiong-Wei, Min Gang, Zhang Qi-Ye. A single-channel speech enhancement approach based on perceptual masking deep neural network. *Acta Automatica Sinica*, 2017, **43**(2): 248−258

语音增强的目标是抑制噪声影响, 提高人耳对带噪语音的感知质量和可懂度. 语音增强的主要任务是在保证语音可懂度的前提下, 从带噪语音信号中提取出尽可能纯净的语音信号. 单通道语音增强方法相比于多通道语音增强方法具有更易部署、成本更低、更易实现等优点, 在国内外引起了广泛的研究.

近几十年来, 各种不同的单通道语音增强算法出现在人们的视野中, 经典的增强方法主要有谱减法[1]、Wiener 滤波法[2]、基于最小均方误差的谱估计算法[3−4] 以及基于语音信号周期模型的增强算法[5] 等. 上述方法通常称为基于统计模型的无监督语音增强方法, 这类方法的优点是不需要先验信息, 即在语音增强时不需要知道具体的噪声类型或者特定说话人的语音特征, 就可以从带噪语音中估计出纯净语音. 基于统计模型的语音增强方法通常在高信噪比以及背景噪声结构较简单平稳的条件下效果较好, 但现实中的噪声往往具有复杂的非平稳特性,

收稿日期 2015-10-31 录用日期 2016-06-06
Manuscript received October 31, 2015; accepted June 6, 2016
国家自然科学基金 (61471394, 61402519), 江苏省自然科学基金 (BK20140071, BK20140074) 资助
Supported by National Natural Science Foundation of China (61471394, 61402519), Natural Science Foundation of Jiangsu Province (BK20140071, BK20140074)
本文责任编委 柯登峰
Recommended by Associate Editor KE Deng-Feng
1. 解放军理工大学 南京 210007 2. 西安通信学院 西安 710106 3. 中国人民解放军 96637 部队 北京 102101
1. PLA University of Science and Technology, Nanjing 210007 2. Xi′an Communications Institute, Xi′an 710106 3. Unit 96637 of PLA, Beijing 102101

基于统计模型的方法对噪声功率谱的估计难度较大, 语音增强的效果并不令人满意.

相对于基于统计理论的增强方法, 基于模型训练的语音增强方法在低信噪比、复杂背景噪声条件下表现出了更好的效果. 非负矩阵分解方法 (Nonnegative matrix factorization, NMF)[6−8] 就是近年来兴起的一种基于模型训练的方法, 该方法分别对语音和噪声信号进行建模, 或者只对语音和噪声二者之一进行建模, 并用语音和 (或) 噪声样本对所建模型进行训练, 估计出模型的具体参数, 然后利用所得参数从带噪语音中估计出纯净语音. 非负矩阵分解方法是建立了一个从带噪语音到纯净语音的线性映射, 但是线性映射并不能很好地表达出语音信号的复杂结构和特征.

近年来, 深度学习凭借其对复杂特征优秀的抽象和建模能力, 在语音信号处理领域引起了广泛的研究. 在深度学习应用于语音识别方面取得巨大成功的影响下, 一些学者们又将深度学习应用于语音增强任务中[9−14]. 基于深度学习的语音增强方法可以很好地学习到从带噪语音数据到纯净语音数据的复杂非线性映射函数, 相比 NMF 方法增强效果更加突出. Xu 等利用大数据的思想, 提出一种基于深度神经网络的语音增强方法[9], 通过 DNN (Deep neural network) 对带噪语音和纯净语音之间进行回归拟合建模. Huang 等提出一种将软掩蔽技术和深度递归神经网络 (Deep recurrent neural network, DRNN) 联合起来进行训练的语噪分离方法[10], 相比先通过 DRNN 训练得到纯净语音信号之后再利用软掩蔽技术进行处理的增强效果有很大提高. Wang 等分析了利用 DNN 对不同的目标进行训练的增强效果[11], 训练目标包括理想二值掩蔽 (Ideal binary mask, IBM)、理想浮值掩蔽 (Ideal ratio mask, IRM) 以及短时傅里叶变换掩蔽 (Short-time Fourier transform mask, SFFT-Mask) 等. 总的来说, 这些基于深度学习的方法及其掩蔽技术在语音增强任务取得了较好的效果, 但很少考虑人类的心理声学掩蔽特性对语音增强效果的影响.

利用深度神经网络在语音增强方面的优势, 本文提出一种将 DNN 和人类的心理声学掩蔽特性相结合的单通道语音增强方法. 该方法将心理声学掩蔽融合到目标语音的幅度谱估计中, 首先利用 DNN 对带噪语音特征进行训练得到一个具有心理声学掩蔽特性的感知增益函数, 然后将该感知增益函数与带噪语音幅度谱进行计算得到纯净语音的幅度谱估计. 实验结果表明, 本文方法消除噪声的性能显著优于 NMF 语音增强方法以及常见的 DNN 语音增强方法.

本文第 1 节介绍基本的 DNN 语音增强方法以及基于心理声学特性的掩蔽阈值计算方法; 第 2 节详述了本文提出的语音增强方法的网络结构和具体实现; 第 3 节针对提出的增强方法进行实验评价; 最后给出总结与讨论.

1　基于 DNN 的语音增强方法及噪声掩蔽阈值

使用 DNN 进行语音增强的基本方法是依据最小均方误差准则, 对带噪语音和纯净语音间的复杂关系进行回归拟合建模, 得到一个从带噪语音到纯净语音的高度非线性映射函数.

1.1　基于 DNN 的语音增强方法

1.1.1　DNN 网络结构

基本的 DNN 是由多层非线性层组合而成的前馈型神经网络, 经过逐层训练从带噪语音特征参数中提取出纯净语音的特征. 基于 DNN 的语音增强结构如图 1 所示.

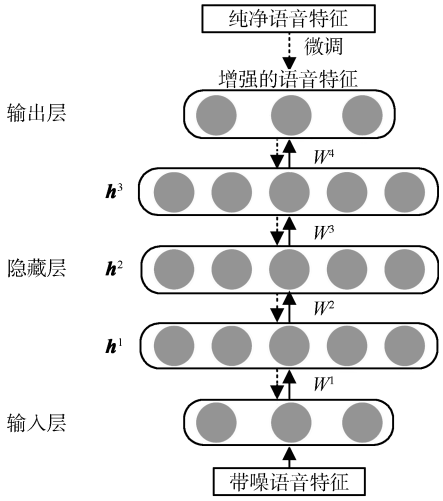


图 1　基于 DNN 的语音增强

Fig. 1　Speech enhancement based on DNN

DNN 的结构通常由 3 部分组成: 输入层、隐藏层和输出层. 输入层用来输入带噪语音的特征参数. 隐藏层一般由多层堆叠而成, 相邻层节点之间有连接, 同一层及跨层节点之间无连接. 输入层以及隐藏层的各个层之间利用激活函数传递数据, 上一层计算得到的输出作为下一层的输入变量, 如式 (1) 所示:

$$\boldsymbol{h}^l = \sigma(W^l \boldsymbol{h}^{l-1} + \boldsymbol{b}^l) \tag{1}$$

式中, $\sigma(\cdot)$ 是激活函数, l 是指网络的层数, W^l 是相邻层之间连接的权值矩阵, $\boldsymbol{b}^l$ 是偏置量, $\boldsymbol{h}^l$ 是第 l 层

的输出, 它由上一层的输出 $\boldsymbol{h}^{l-1}$ 计算得到, 首先由输入层的带噪语音特征参数 $\boldsymbol{h}^0=\boldsymbol{x}$ 开始计算.

输出层要根据需求选用合适的函数来得到输出结果, 函数既可以是线性的也可以是非线性的.

1.1.2 DNN 网络的训练

训练一个 DNN 通常包含两个阶段: 无监督的预训练和有监督的微调.

逐层贪婪无监督预训练是 2006 年 Hinton 等提出用来解决误差反向传播时, 网络易陷入局部最优的问题[15]. 预训练的思想主要是: 互相连接的每两层作为一个限制玻尔兹曼机 (Restricted Boltzmann machine, RBM), 网络由若干个 RBM 堆叠而成, 由底层到高层利用对比散度算法 (Contrastive divergence, CD) 依次对每个 RBM 进行训练来更新权值[16].

近年来, 文献 [17] 提出一种称为 ReLU (Rectified linear unit) 的激活函数 $f(x)=\max(0,x)$ 来代替 RBM 中常用的 sigmoid 激活函数 $\sigma(x)=1/(1+\mathrm{e}^{-x})$. 采用 ReLU 激活函数的 DNN 计算更简单, 网络的初始权值采用随机初始化时, 同样可以使网络得到很好的效果, 成为目前研究中的主流激活函数.

在有监督的微调阶段, 网络训练的目标是最小化网络输出的增强语音与纯净语音之间的误差, 网络训练的目标函数如下:

$$J_{\mathrm{MSE}}(W,\boldsymbol{b})=\frac{1}{N}\sum_{n=1}^{N}\frac{1}{2}\left\|\hat{\boldsymbol{S}}_n(W,\boldsymbol{b})-\boldsymbol{S}_n\right\|^2 \tag{2}$$

式中, n 是样本索引号, N 是所有的样本数, $\hat{\boldsymbol{S}}_n$ 是增强的语音特征而 $\boldsymbol{S}_n$ 是与其相对应的纯净语音特征.

计算得到增强语音与纯净语音的误差之后, 将该误差用反向传播算法来修正网络的权值. 对权值矩阵 W 和偏置量 $\boldsymbol{b}$ 的更新法则如下所示:

$$W^l=W^l-\varepsilon\frac{\partial J(W,\boldsymbol{b})}{\partial W^l},\quad 1\le l\le L+1 \tag{3}$$

$$\boldsymbol{b}^l=\boldsymbol{b}^l-\varepsilon\frac{\partial J(W,\boldsymbol{b})}{\partial \boldsymbol{b}^l},\quad 1\le l\le L+1 \tag{4}$$

式中, ε 是学习速率, L 指总的隐藏层数, $L+1$ 指网络的输出层.

网络参数经过一系列的迭代更新, 就可以得到一个训练好的 DNN 来实现语音增强.

1.2 心理声学模型及噪声掩蔽阈值的计算

心理声学模型是在研究人类听觉系统基础上抽象出来反映人类听觉感知特性的数学模型, 它描述了人类听觉系统对语音及噪声的感知和掩蔽能力.

计算听觉场景分析是典型的模拟人耳听觉感知过程的方法. 听觉场景分析认为人类的听觉感知系统在感知声音的过程中对听觉场景中的声学事件进行了组织, 将混合声音中属于同一声源的声音分量组织到一个听觉流中, 从而实现了语音分离, 而计算听觉场景分析就是利用计算机来模拟听觉场景分析的过程.

心理声学模型的掩蔽特性是: 当一些噪声在人类听觉掩蔽阈值之下的时候, 听觉系统就无法感知这些噪声的存在. 因此, 在实际中可以利用计算得到的掩蔽阈值来有效去除噪声. 语音分离的任务是将目标语音与干扰语音分开, 而心理声学模型拥有的感知掩蔽噪声特性, 在语音分离任务中得到了广泛应用.

依据心理声学模型, 输入信号频带需要按临界频带 (单位: Bark) 重新划分, 然后估计出每个临界频带的掩蔽阈值, 以此来对噪声进行整形, 使每个临界频段内的噪声功率小于该子带的掩蔽阈值, 从而能够被语音信号所掩蔽, 达到感知失真最小[18].

Johnston 提出了一种在各语音帧中, 估计背景噪声掩蔽阈值的一般方法[19], 该方法建立在临界带分析的基础上, 具体可以表述为以下 4 个步骤:

步骤 1. 掩蔽阈值是从纯净语音的各个分析帧得到的. 首先计算得到语音信号帧 $\boldsymbol{s}(t)$ 的功率谱 $\boldsymbol{P}(\omega)$ 为:

$$\boldsymbol{P}(\omega)=\mathrm{Re}^2(\boldsymbol{S}(\omega))+\mathrm{Im}^2(\boldsymbol{S}(\omega)) \tag{5}$$

式中, $\boldsymbol{S}(\omega)$ 是信号 $\boldsymbol{s}(t)$ FFT (Fast Fourier transform) 变换得到的频域信号. 然后计算各个临界带内的能量 B_i:

$$B_i=\sum_{\omega=bl_i}^{bh_i}\boldsymbol{P}(\omega) \tag{6}$$

式中, i 代表 24 个 Bark 临界带中的一个编号, bl_i 和 bh_i 分别表示第 i 段最低和最高的频率.

步骤 2. 为了计算相近的临界带存在的掩蔽效应, 从第一步得到的临界能量 B_i 需要与一个 "扩展矩阵" 做卷积. 令 S_{ij} 表示扩展矩阵 S 中元素, 并且 i 和 j 满足:

$$|\,j-i\,|\le 25 \tag{7}$$

式中, i 为掩蔽目标的 Bark 频率, j 是掩蔽源的 Bark 频率. 扩展 Bark 频域谱可由 S_{ij} 和 B_i 相互卷积得到:

$$C_i=S_{ij}*B_i \tag{8}$$

步骤 3. 根据 "掩蔽音" 具有的类似噪声或纯音的特性, 从上式计算的掩蔽门限中减去一个偏差[20].

纯音掩蔽噪声:

$$T_{\mathrm{N}} = C_i - 14.5 - i \tag{9}$$

噪声掩蔽纯音:

$$T_{\mathrm{T}} = C_i - 5.5 \tag{10}$$

为了判断信号是语音还是噪声, 需要用到谱平面测度 SFM (Spectral flatness measure), SFM 表示如下:

$$\mathrm{SFM}_{\mathrm{dB}} = 10\mathrm{lg}\left(\frac{G_{\mathrm{m}}}{A_{\mathrm{m}}}\right) \tag{11}$$

式中, G_{m} 和 A_{m} 分别表示该语音信号的几何平均和算术平均.

由 SFM 计算得到系数 α_{SFM}:

$$\alpha_{\mathrm{SFM}} = \min\left(\frac{\mathrm{SFM}_{\mathrm{dB}}}{\mathrm{SFM}_{\mathrm{dBmax}}}, 1\right) \tag{12}$$

$\mathrm{SFM}_{\mathrm{dBmax}}$ 影响语音信号是否都是纯音, $\mathrm{SFM}_{\mathrm{dBmax}}$ 的一个常用值是 $-60\,\mathrm{dB}$. α_{SFM} 取值范围是 0 到 1 之间. 当 $\mathrm{SFM}_{\mathrm{dB}} = 0\,\mathrm{dB}$ 时, $\alpha_{\mathrm{SFM}} = 0$, 表示语音信号全部都是噪声; 当 $\mathrm{SFM}_{\mathrm{dB}}$ 取值导致 $\alpha_{\mathrm{SFM}} = 1$ 时, 表示语音信号全部都是纯音, 实际语音信号通常是介于纯音和噪音之间的.

掩蔽能量的偏移可表示为:

$$o_i = \alpha_{\mathrm{SFM}}(14.5 + i) + (1 - \alpha_{\mathrm{SFM}}) \times 5.5 \tag{13}$$

则噪声的掩蔽阈值为:

$$T_i = 10^{\lg C_i - \left(\frac{o_i}{10}\right)} \tag{14}$$

步骤 4. 将步骤 3 得到的 Bark 刻度下的掩蔽门限 T_i 变回线性频率刻度, 从而得到 $\boldsymbol{T}(\omega)$, 其中 ω 是 DFT (Discrete Fourier transform) 被采样的频率.

2　基于感知掩蔽 DNN 的语音增强方法

心理声学的掩蔽特性可以将低于掩蔽阈值的残余噪声有效去除掉. 本节利用该优点, 提出一种将感知掩蔽技术融入到 DNN 的新型网络结构来实现语音增强.

2.1　感知增益函数的计算

假设带噪语音信号帧可以表示为纯净语音帧以及与语音不相关的加性噪声帧叠加得到的, 如下式所示:

$$\boldsymbol{y} = \boldsymbol{s} + \boldsymbol{n} \tag{15}$$

式中, $\boldsymbol{y}$, $\boldsymbol{s}$ 和 $\boldsymbol{n}$ 分别为带噪语音信号帧、干净语音信号帧和噪声信号帧. 通过离散傅里叶变换将上式转化到频域上, 表示为:

$$\boldsymbol{Y}(\omega) = F^{\mathrm{H}}\boldsymbol{y} = F^{\mathrm{H}}\boldsymbol{s} + F^{\mathrm{H}}\boldsymbol{n} = \boldsymbol{S}(\omega) + \boldsymbol{N}(\omega) \tag{16}$$

式中, F^{H} 表示 N 点离散傅里叶变换矩阵, H 表示 Hermite 算子. $\boldsymbol{S}(\omega)$ 和 $\boldsymbol{N}(\omega)$ 分别为干净语音和噪声的频谱分量.

假设增强后得到的语音频谱为 $\hat{\boldsymbol{S}}(\omega)$. $\hat{\boldsymbol{S}}(\omega)$ 可由一个增益函数 $\boldsymbol{G}$ 从带噪语音频谱 $\boldsymbol{Y}(\omega)$ 中估计得到, 如下式所示:

$$\hat{\boldsymbol{S}}(\omega) = \boldsymbol{G} \otimes \boldsymbol{Y}(\omega) \tag{17}$$

式中, $\otimes$ 表示对应元素相乘. 则语音信号频域估计误差可以表示为:

$$\begin{aligned}\boldsymbol{\varepsilon}(\omega) &= \hat{\boldsymbol{S}}(\omega) - \boldsymbol{S}(\omega) = \\ &(\boldsymbol{G} - \boldsymbol{I}) \otimes \boldsymbol{S}(\omega) + \boldsymbol{G} \otimes \boldsymbol{N}(\omega) = \\ &\boldsymbol{\varepsilon}_{\boldsymbol{S}}(\omega) + \boldsymbol{\varepsilon}_{\boldsymbol{N}}(\omega)\end{aligned} \tag{18}$$

式中, $\boldsymbol{\varepsilon}_{\boldsymbol{S}}(\omega)$ 表示语音失真频谱, $\boldsymbol{\varepsilon}_{\boldsymbol{N}}(\omega)$ 表示残余噪声频谱.

假设语音失真的能量为 $\boldsymbol{E}_{\boldsymbol{S}}(\omega) = \mathrm{E}(\boldsymbol{\varepsilon}_{\boldsymbol{S}}^{\mathrm{H}}(\omega) \times \boldsymbol{\varepsilon}_{\boldsymbol{S}}(\omega))$, 残余噪声能量为 $\boldsymbol{E}_{\boldsymbol{N}}(\omega) = \mathrm{E}(\boldsymbol{\varepsilon}_{\boldsymbol{N}}^{\mathrm{H}}(\omega) \times \boldsymbol{\varepsilon}_{\boldsymbol{N}}(\omega))$. 结合人耳的掩蔽效应, 最优的增益函数 $\boldsymbol{G}$ 应该使语音失真尽可能小的同时, 保证噪声处于人耳掩蔽阈值之下, 即满足如下最优化问题:

$$\begin{aligned}&\min_{G} \ \boldsymbol{E}_{\boldsymbol{S}}(\omega) \\ &\text{s.t.} \quad \boldsymbol{E}_{\boldsymbol{N}}(\omega) \le \boldsymbol{T}(\omega)\end{aligned} \tag{19}$$

式中, $\boldsymbol{T}(\omega)$ 为短时幅度谱分量的听觉掩蔽阈值估计值, 由上一节所述的心理声学模型计算得到. 为了求解上面的约束最优化问题, 构造 Lagrange 代价函数:

$$J = \boldsymbol{E}_{\boldsymbol{S}}(\omega) + \mu \cdot (\boldsymbol{E}_{\boldsymbol{N}}(\omega) - \boldsymbol{T}(\omega)) \tag{20}$$

其中 μ 是 Lagrange 乘子, 并且 $\mu > 0$. 文献 [21] 给出了该优化问题的详细求解过程, 在此不再详述. 感知增益函数 $\boldsymbol{G}$ 可以表示为:

$$\boldsymbol{G} = \frac{\boldsymbol{1}}{\boldsymbol{1} + \max\left(\sqrt{\dfrac{|\boldsymbol{N}(\omega)|^2}{\boldsymbol{T}(\omega)}} - \boldsymbol{1}, \boldsymbol{0}\right)} \tag{21}$$

2.2　网络结构

本节介绍提出的结合人耳感知听觉掩蔽特性的深度神经网络结构, 将这个网络命名为 PM-DNN (Perceptual masking deep neural network), PM-DNN 结构如图 2 所示.

如图 2 所示, 提出的 PM-DNN 结构是将感知增益函数与 DNN 网络结合为一个整体进行训练, 详细的训练过程可表述如下:

1) 利用 PM-DNN 同时得到增强的语音幅度谱 $\tilde{\boldsymbol{S}}$ 以及分离出的干扰噪声幅度谱 $\tilde{\boldsymbol{N}}$, 而常用的 DNN 训练输出只有增强的语音幅度谱特征.

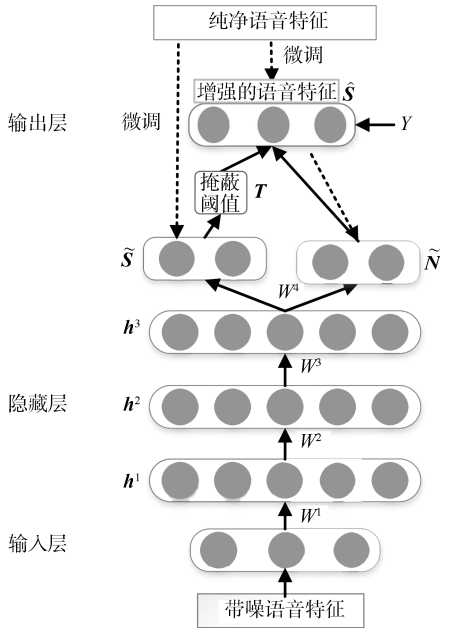


图 2 基于 PM-DNN 的语音增强

Fig. 2 Speech enhancement based on PM-DNN

2) 利用增强后的语音幅度谱 $\tilde{\boldsymbol{S}}$ 来计算得到掩蔽阈值 $\boldsymbol{T}$. 传统语音增强算法为了精确地估计掩蔽阈值, 通常是对带噪语音由噪声消减算法 (例如谱减法) 进行处理, 由增强后的频谱来计算掩蔽阈值.

3) 利用计算得到的掩蔽阈值 $\boldsymbol{T}$ 以及噪声的幅度谱 $\tilde{\boldsymbol{N}}$ 来计算式 (21) 所示的感知增益函数 $\boldsymbol{G}$.

4) 将感知增益函数 $\boldsymbol{G}$ 连同带噪语音幅度谱 $\boldsymbol{Y}$ 作为网络的额外一层叠加于原始网络的输出之上, 得到网络的最终输出 $\hat{\boldsymbol{S}}$, 如下所示:

$$\hat{\boldsymbol{S}} = \boldsymbol{G} \otimes \boldsymbol{Y} = \frac{\boldsymbol{1}}{\boldsymbol{1} + \max\left(\sqrt{\frac{\tilde{\boldsymbol{N}}^2}{\boldsymbol{T}}} - \boldsymbol{1},\ \boldsymbol{0}\right)} \otimes \boldsymbol{Y} \tag{22}$$

当得到网络的最终输出之后, 我们用纯净语音对网络权值参数进行有监督的微调, 网络的训练目标函数由两部分组成, 如式 (23) 所示:

$$J = \alpha||\hat{\boldsymbol{S}} - \boldsymbol{S}||_2^2 + \beta||\tilde{\boldsymbol{S}} - \boldsymbol{S}||_2^2 \tag{23}$$

式中, $||\hat{\boldsymbol{S}} - \boldsymbol{S}||_2^2$ 表示网络最终输出 $\hat{\boldsymbol{S}}$ 与纯净语音 $\boldsymbol{S}$ 之间的误差, $||\tilde{\boldsymbol{S}} - \boldsymbol{S}||_2^2$ 表示网络前一层得到的增强语音幅度谱 $\tilde{\boldsymbol{S}}$ 与纯净语音 $\boldsymbol{S}$ 之间的误差, α 和 β 分别是前后两项的权重值, 并且 $\alpha + \beta = 1$. 由于 $\tilde{\boldsymbol{S}}$ 与最后一层隐藏层 $\boldsymbol{h}_3$ 之间的权值, 即 W^4 的一部分值, 无法由最终输出 $\hat{\boldsymbol{S}}$ 和纯净语音 $\boldsymbol{S}$ 的误差通过反向传播算法来迭代更新 (误差 $||\hat{\boldsymbol{S}} - \boldsymbol{S}||_2^2$ 无法通过掩蔽阈值 $\boldsymbol{T}$ 反向传播至 $\tilde{\boldsymbol{S}}$), 因此需要通过在目标函数中增加一项 $||\tilde{\boldsymbol{S}} - \boldsymbol{S}||_2^2$, 利用 $\tilde{\boldsymbol{S}}$ 和 $\boldsymbol{S}$ 之间的误差对 $\tilde{\boldsymbol{S}}$ 与 $\boldsymbol{h}_3$ 之间的权值进行更新, 使估计的语音幅度谱 $\tilde{\boldsymbol{S}}$ 更准确, 从而计算出的掩蔽阈值也更准确.

网络的权值更新需要利用误差反向传播算法来计算. 依据链式规则, $\tilde{\boldsymbol{S}}$ 和 $\tilde{\boldsymbol{N}}$ 所在层的偏导数分别为:

$$\boldsymbol{\delta}_{\tilde{\boldsymbol{S}}} = \frac{\partial J}{\partial \tilde{\boldsymbol{S}}} \tag{24}$$

$$\boldsymbol{\delta}_{\tilde{\boldsymbol{N}}} = \frac{\partial J}{\partial \tilde{\boldsymbol{N}}} = \frac{\partial J}{\partial \hat{\boldsymbol{S}}} \otimes \frac{\partial \hat{\boldsymbol{S}}}{\partial \tilde{\boldsymbol{N}}} = - \frac{\partial J}{\partial \hat{\boldsymbol{S}}} \otimes \frac{\left(\dfrac{\tilde{\boldsymbol{N}}}{\left(\boldsymbol{T} \otimes \sqrt{\dfrac{\tilde{\boldsymbol{N}}^2}{\boldsymbol{T}}}\right)}\right) \otimes \boldsymbol{M}_{1,0}}{\left[\boldsymbol{1} + \max\left(\sqrt{\dfrac{\tilde{\boldsymbol{N}}^2}{\boldsymbol{T}}} - \boldsymbol{1}, \boldsymbol{0}\right)\right]^2} \tag{25}$$

式 (25) 中, $\boldsymbol{M}_{1,0}$ 表示对由 $\tilde{\boldsymbol{N}}$ 构成的复合函数 $\max(\boldsymbol{u},\ \boldsymbol{0})(\boldsymbol{u} = \sqrt{\tilde{\boldsymbol{N}}^2 / \boldsymbol{T}} - \boldsymbol{1})$ 求偏导的结果, $\max(\boldsymbol{u},\ \boldsymbol{0})$ 的偏导值是根据 $\boldsymbol{u}$ 是否大于 $\boldsymbol{0}$ 而得到的由 1 和 0 构成的向量.

得到偏导数 $\boldsymbol{\delta}_{\tilde{\boldsymbol{S}}}$ 和 $\boldsymbol{\delta}_{\tilde{\boldsymbol{N}}}$ 之后, 就按照常用的 DNN 训练方法对 PM-DNN 的权值参数 W 和偏置参数 $\boldsymbol{b}$ 进行更新.

网络权值参数经过若干次数的迭代更新, 就可得到一个训练好的具有听觉掩蔽效果的 PM-DNN.

2.3 PM-DNN 语音增强流程

利用 PM-DNN 进行语音增强任务主要分为训练和增强两个阶段, 具体流程框图如图 3 所示.

在训练阶段, 首先对带噪语音进行短时傅里叶变换 (Short-time Fourier transform, STFT) 求取带噪语音的幅度谱, 然后将带噪语音的幅度谱作为 PM-DNN 的输入特征, 对 PM-DNN 进行训练.

在增强阶段, 将测试的带噪语音幅度谱特征直接输入训练好的 PM-DNN 得到增强语音的幅度谱特征, 利用人耳对相位信息不敏感的特性, 使用带噪

语音的相位信息进行语音重构得到增强语音的频谱, 然后利用短时傅里叶逆变换 STFT^{-1} 得到时域的增强语音信号.

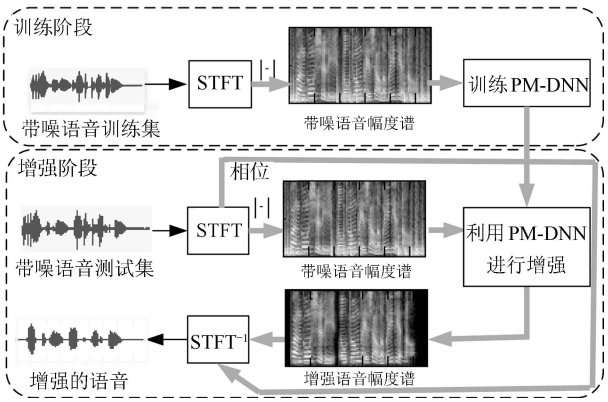


图 3　基于 PM-DNN 的语音增强框图

Fig. 3　The framework of speech enhancement based on PM-DNN

3　PM-DNN 语音增强方法性能评估

本节对所提出的 PM-DNN 语音增强方法进行实验仿真测试, 并就其性能进行评估.

3.1　实验数据及设置

实验中的纯净语音选自 TIMIT 标准语音库, 从 240 个不同的男性和女性说话人中随机选取 600 条语句作为训练语音. 实验所用的噪声来自 Noisex-92 标准噪声库以及实际生活中常见的一些非平稳噪声. Babble, Car, Casino, Cicadas, F16, Factory1, Frogs, HFchannel, Jungle, Restaurant, Street, White, Airport, Pink 和 Birds 这 15 种噪声作为训练噪声. 所有的纯净语音和噪声都采样到 8 kHz. 600 条训练语音中的每条语音在上述 15 种噪声中随机选取一种噪声, 并从 −6 dB, −3 dB, 0 dB, 3 dB, 6 dB, 9 dB, 12 dB 和 15 dB 这些信噪比中随机选取一种信噪比进行语音和噪声混合, 将混合好的 600 条多样性带噪语音作为网络的训练集. 带噪语音的幅度谱作为 PM-DNN 训练的输入特征, 进行 STFT 时所使用的语音帧长为 64 ms (512 点), 帧移为 16 ms (128 点), PM-DNN 隐藏层数设为 3 层, 每层 2 048 个节点, 激活函数选择 ReLU 函数, 网络连接权值采用随机初始化方法.

测试阶段, 另选 120 个与训练阶段不同的男性和女性说话人并从中随机抽取 60 条语句作为实验测试用的纯净语音. 我们除了选取带噪语音集中出现的 15 种噪声进行测试, 同时还另外选取输入带噪语音集中没有出现的 Exhibition, Subway, Train, Motorcycles 和 Ocean 这 5 种噪声来测试所提增强方法的泛化能力.

3.2　对比方法及评价指标

实验对比方法采用 NMF, DNN 以及隐式的傅里叶变换域的 IRM-DNN 作为基准. NMF 方法中纯净语音字典由 DNN 训练集中的 600 条纯净语音训练得到, 语音字典基设为 1 000, 噪声字典基设为 100. NMF 训练特征选择 257 维的幅度谱, 用 Hamming 窗计算幅度谱, 窗长为 64 ms, 帧移为 16 ms.

隐式的傅里叶变换域的 IRM-DNN 就是在基本的 DNN 之上叠加一个理想浮值掩蔽层, 联合带噪语音谱一起作为网络的训练目标输出, 傅里叶变换域的 IRM 定义如下:

$$\mathrm{IRM}=\frac{|\boldsymbol{S}(\omega)|^2}{|\boldsymbol{S}(\omega)|^2+|\boldsymbol{N}(\omega)|^2} \tag{26}$$

式中, $|\boldsymbol{S}(\omega)|^2$ 和 $|\boldsymbol{N}(\omega)|^2$ 分别表示纯净语音和噪声的能量. DNN 和 IRM-DNN 采用与 PM-DNN 相同的数据集进行训练, 隐藏层数同样设为 3 层, 每层 2 048 个节点. NMF 和 DNN 方法在最后利用软掩蔽技术来进一步提升语音信号的自然度, 获得更好的试听体验. 测试 PM-DNN 的增强性能时, 我们分别分析 PM-DNN 首次得到的增强语音幅度谱的效果, PM-DNN 不加掩蔽技术以及 PM-DNN 联合软掩蔽技术的增强效果.

在评价指标中, 采用感知语音质量评估方法 (Perceptual evaluation of speech quality, PESQ)[22]、对数谱距离 (Log-spectral distance, LSD)[23] 和频率加权分段信噪比 (Frequency weighted segmental SNR, fwSNRseg)[24] 来分别评估增强语音的质量和增强方法的实际性能. 其中 PESQ 是 ITU-T 推荐的评估方法, 是一种能够评价语音主观试听效果的客观计算方法, 可以很好地近似平均意见得分 (Mean opinion score, MOS), PESQ 的取值范围为 $-0.5\sim4.5$, 得分越高说明算法增强效果越好. LSD 衡量纯净语音和增强语音之间的对数谱距离, 其值与语音质量成反比, 即越小的值表示越好的增强效果. fwSNRseg 是衡量增强算法对噪声抑制能力的指标, 它是将分段信噪比 (Segmental SNR, SNRseg) 扩展到频域. 频域分段信噪比算法优于时域 SNRseg 的地方在于增加了对频谱上不同频带施加不同权重的灵活度. LSD 和 fwSNRseg 的计算公式分别为:

$$\mathrm{LSD}=\frac{1}{M}\sum_{m=0}^{M-1}\sqrt{\frac{1}{\frac{L}{2}+1}\sum_{l=0}^{\frac{L}{2}}10\lg\frac{|S(m,l)|^2}{|\hat{S}(m,l)|^2}} \tag{27}$$

$$\text{fwSNRseg} = \frac{10}{M}\sum_{m=0}^{M-1}\frac{\sum_{j=1}^{k} B_j \lg\left[\dfrac{F^2(m,j)}{(F(m,j)-\hat{F}(m,j))^2}\right]}{\sum_{j=1}^{k} B_j} \tag{28}$$

式中, M 是总的信号帧数, $S(m,l)$ 和 $\hat{S}(m,l)$ 分别为纯净语音和增强后语音经过短时傅里叶变换后的第 m 帧的第 l 个频谱分量, B_j 是第 j 个频带的权重, k 是频带个数, $F(m,j)$ 是第 m 帧纯净信号的第 j 个频带的滤波带幅度, $\hat{F}(m,j)$ 是增强信号在相同频带的滤波带幅度.

3.3 实验结果及分析

图 4 给出 PM-DNN 目标函数中的权重 α 和 β 对 20 种噪声的 PESQ 均值影响. 可以看出, PM-DNN 目标函数的前后两项的权重对增强性能的影响并不大, 这表明在误差反向传播时, 只要目标函数中前后两项存在, 网络经过一系列的迭代训练, 就可以得到增强效果较好的网络权值参数.

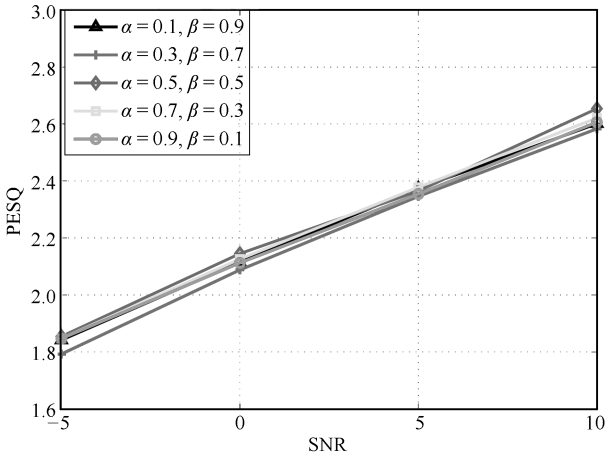


图 4 PM-DNN 目标函数中的权重 α 和 β 对 20 种噪声的 PESQ 均值影响

Fig. 4 The PESQ scores of PM-DNN objective function with different α and β (For each condition, the numbers are the mean values over all the 20 noise types.)

表 1 列举了在不同信噪比下, 4 种增强方法对于 20 种不同噪声环境的 PESQ 测量均值, 这些测量值表明 PM-DNN 方法在 4 种不同信噪比下, 都有好于 NMF, DNN 和 IRM-DNN 的增强效果. NMF 方法无论是否用软掩蔽做后处理, PESQ 值差距并不大, 但有研究发现 NMF 方法利用软掩蔽可以提高信源失真比 (Signal to distortion ratio, SDR)[25], 增强后的语音听起来也更自然. DNN 方法用软掩蔽技术效果更好, 而 IRM-DNN 和 PM-DNN 则不使用软掩蔽技术效果更好, 这是因为 DNN 直接以语音幅度谱作为训练目标, 而 IRM-DNN 和 PM-DNN 是将隐式的掩蔽函数作为训练目标. 对比 PM-DNN (First output) 和 DNN 可以看出, 随着信噪比的提高, PM-DNN (First output) 的增强效果与 DNN 的差距在逐渐减小. 将 PM-DNN 与 DNN 和 NMF 对比可以看出, 信噪比越高, PM-DNN 的增强性能提升更明显, 这是由于信噪比越高, 导致 PM-DNN 估计的第一次增强语音的幅度谱越来越准确, 导致根据它计算出的掩蔽阈值也更准确, 增强效果也越来越好.

图 5 给出了 4 种增强方法对于 20 种不同噪声, 在 4 种信噪比下的 PESQ 测量均值. 可以看出, 对于在训练集中出现的 15 种噪声, 除了 Cicadas 噪声, DNN 方法的增强效果都要好于 NMF 方法, 而 IRM-DNN 的增强效果相较 DNN 又有明显提高. PM-DNN 方法首次得到的增强语音效果稍弱于 DNN, 但大部分噪声下增强效果优于 NMF 方法, 当 PM-DNN 增加感知掩蔽层之后, PM-DNN 在首次得到的增强语音幅度谱基础之上效果有了显著提高, 无论是否利用软掩蔽技术, 效果都要远远好于 NMF, DNN 和 IRM-DNN 方法, 这表明 PM-DNN 直接训练感知增益函数得到的首次增强语音幅度谱并不是最优的, 这是因为将感知增益函数以及 DNN 作为一个整体进行训练, 训练得到的网络各个层的权值参数保证最终输出结果是最好的, 但并不能保证中间的某一输出也是最好的. 对于在训练集中没有出现的 5 种噪声, DNN 的增强效果有所下降, 这是因为 DNN 训练时没有学到这些噪声的结构特点, 泛化性能并不是很好, 但我们提出的 PM-DNN 对于训练集中没出现的噪声依然有很好的增强效果, 仅仅在 Motorcycles 下效果微弱于 NMF, 一方面因为 NMF 更擅长于去除 Motorcycles 这类具有低秩重复性结构的噪声, 另一方面就是该噪声没有出现在 PM-DNN 训练集中.

图 6 和图 7 分别给出了 4 种增强方法对于 20 种不同噪声, 在 4 种信噪比下的 LSD 测量均值和 fwSNRseg 测量均值. 从图 6 可以看出, 对于在训练集中出现的 15 种噪声, PM-DNN 相较于 NMF, DNN 和 IRM-DNN 都有较低的 LSD 值, 只在 Frogs, Jungle 和 Birds 噪声下与 DNN 效果持平, 表明采用 PM-DNN 增强方法导致的语音失真相对 NMF 和 DNN 较小. 对于训练集中没有出现的 5 种噪声, DNN 在 Exhibition 和 Subway 噪声下, LSD 指标较差于 NMF. IRM-DNN 在 LSD 指标的表现效果也并不好, PM-DNN 仅在 Motorcycles 噪声下稍弱于 NMF 和 DNN, 在 Subway 噪声下与 IRM-DNN 持平, 其他情况下, PM-DNN 的 LSD 值都要低于 3

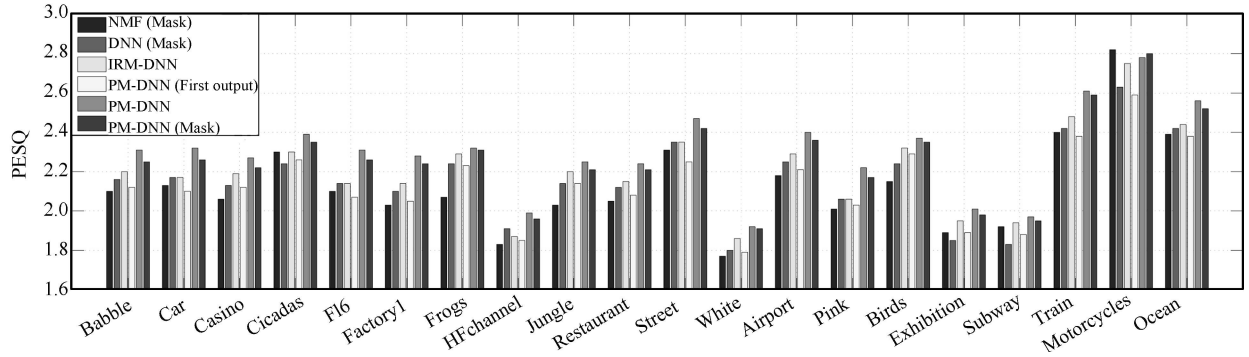


图 5　4 种增强方法在 20 种不同噪声情况下的 PESQ 值 (每种噪声的 PESQ 值是在 −5 dB, 0 dB, 5 dB 和 10 dB 4 种信噪比下的平均值.)

Fig. 5　The PESQ scores of the 4 enhancement methods for the 20 noise types (For each noise type, the numbers are the mean values over four input SNR conditions, i.e. from −5 dB to 10 dB spaced by 5 dB.)

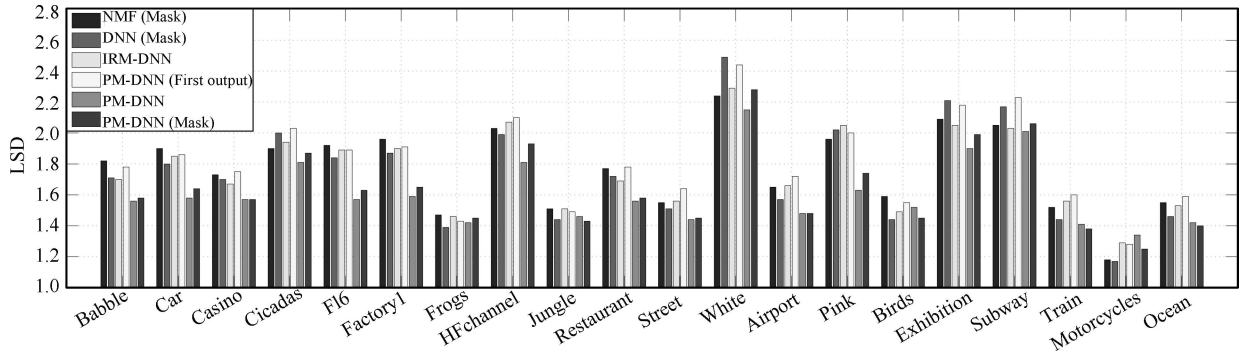


图 6　4 种增强方法在 20 种不同噪声情况下的 LSD 值 (每种噪声的 LSD 值是在 −5 dB, 0 dB, 5 dB 和 10 dB 4 种信噪比下的平均值.)

Fig. 6　The LSD values of the 4 enhancement methods for the 20 noise types (For each noise type, the numbers are the mean values over four input SNR conditions, i.e. from −5 dB to 10 dB spaced by 5 dB.)

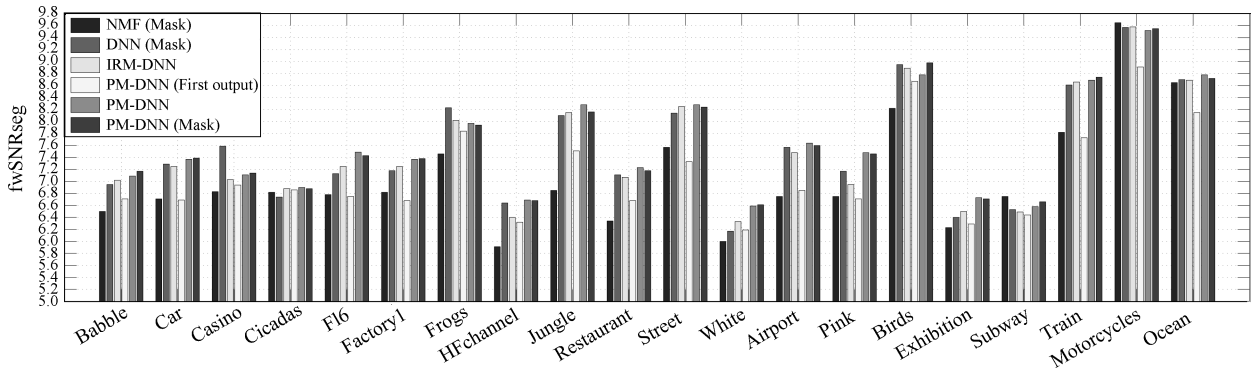


图 7　4 种增强方法在 20 种不同噪声情况下的 fwSNRseg 值 (每种噪声的 fwSNRseg 值是在 −5 dB, 0 dB, 5 dB 和 10 dB 4 种信噪比下的平均值.)

Fig. 7　The fwSNRseg values of the 4 enhancement methods for the 20 noise types (For each noise type, the numbers are the mean values over four input SNR conditions, i.e. from −5 dB to 10 dB spaced by 5 dB.)

表 1 4 种信噪比下, 不同方法对 20 种噪声的 PESQ 均值

Table 1 The PESQ scores of different methods at four different input SNR levels (For each condition, the numbers are the mean values over all the 20 noise types.)

SNR (dB)	NMF	DNN	IRM-DNN	PM-DNN (First output)	PM-DNN	NMF (Mask)	DNN (Mask)	IRM-DNN (Mask)	PM-DNN (Mask)
−5	1.705	1.740	1.787	1.732	1.875	1.701	1.775	1.740	1.834
0	2.002	1.995	2.061	1.996	2.165	1.995	2.034	2.015	2.122
5	2.261	2.194	2.350	2.256	2.445	2.262	2.284	2.308	2.411
10	2.524	2.350	2.631	2.518	2.714	2.520	2.535	2.596	2.691

种对比方法, 这表明 PM-DNN 既可保证语音失真较小又有良好的泛化性能. 从图 7 反映的对噪声抑制能力的 fwSNRseg 指标可以分析得出, NMF 方法似乎不善于提高频率加权分段信噪比的值, 而我们所提的 PM-DNN 方法在绝大部分噪声情况下, 都有好于 DNN 和 IRM-DNN 的 fwSNRseg 值, 这表明 PM-DNN 方法在噪声抑制方面的性能要优于 NMF, DNN 和 IRM-DNN. LSD 和 fwSNRseg 这些测量值都进一步验证了本文方法具有良好的增强性能.

虽然 PESQ 值、LSD 测度以及 fwSNRseg 测度都反映出本文方法的良好性能, 但这 3 个指标只是在宏观上反映了本文方法的性能, 为了更好地观察出增强语音信号的细节特征, 本文给出了 4 种增强方法对于输入信噪比为 5 dB, 被 F16 飞机噪声所污染的纯净语音进行增强前后的语谱图, 如图 8 所示.

由图 8 可见, 在 2 750 Hz 的频带附近, 使用 NMF 进行增强的结果依然存在少量噪声残留, 导致了类似音乐噪声的试听感受, 而 DNN 相比 NMF 去除噪声的效果要好很多, 但在一些低频段还有微弱的残余噪声, 而本文 PM-DNN 方法在有效去除高频噪声的同时, 对低频噪声也有很好的处理, 相比于 DNN 去噪效果更明显. PM-DNN 能够在有效去除噪声的前提下, 较好地保持语音信号的固有谐波特性, 增强效果明显优于 NMF 和 DNN 方法.

4 结论及展望

本文提出一种新颖的深度神经网络结构来实现单通道语音增强任务, 该网络结构结合了人类心理声学的特点, 将一个具有掩蔽阈值约束的感知增益函数和带噪语音幅度谱联合优化. 通过对不同信噪比下的 20 种不同噪声进行的仿真实验表明, 该语音增强方法的 PESQ, LSD 及 fwSNRseg 增强性能指标都显著优于非负矩阵分解方法和常见的深度神经网络方法, 无论在高信噪比还是低信噪比的噪声环境下, 都有很好的增强效果, 能够在保留语音信号固有谐波特性的同时很好地移除噪声.

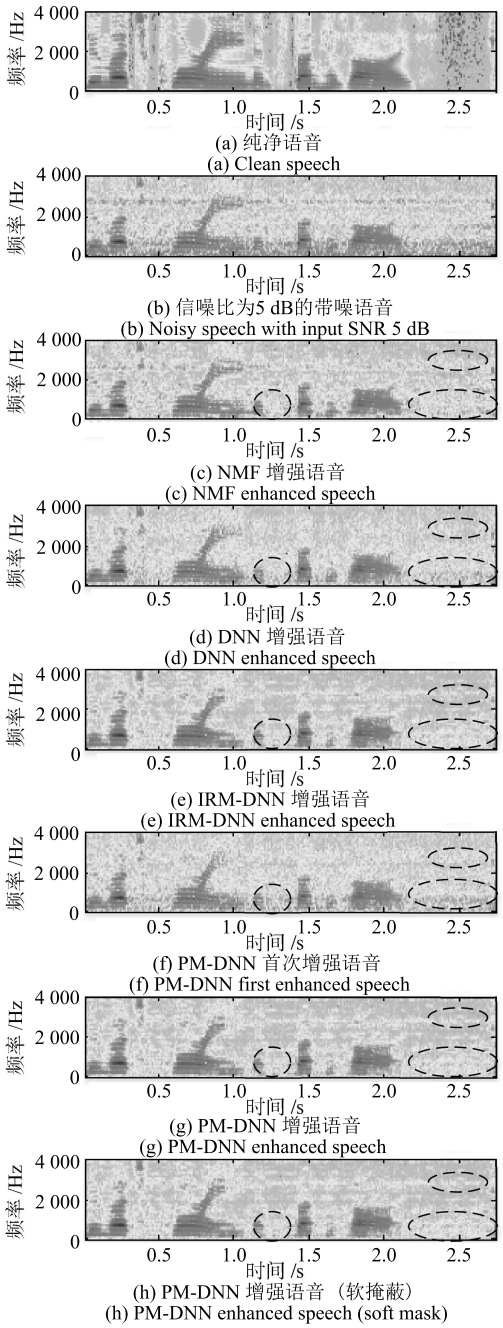


图 8 语谱图

Fig. 8 Spectrograms

在未来的工作中, 可针对如何提高网络训练过程中用来计算掩蔽阈值的纯净语音幅度谱的估计值来进行研究, 从而更进一步提高网络的增强效果.

References

1 Boll S F. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1979, **27**(2): 113−120

2 Chen J D, Benesty J, Huang Y T, Doclo S. New insights into the noise reduction Wiener filter. *IEEE Transactions on Audio, Speech and Language Processing*, 2006, **14**(4): 1218−1234

3 Ephraim Y, Malah D. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1984, **32**(6): 1109−1121

4 Gerkmann T, Hendriks R C. Unbiased MMSE-based noise power estimation with low complexity and low tracking delay. *IEEE Transactions on Audio, Speech, and Language Processing*, 2012, **20**(4): 1383−1393

5 Jensen J R, Benesty J, Christensen M G, Jensen S H. Enhancement of single-channel periodic signals in the time-domain. *IEEE Transactions on Audio, Speech, and Language Processing*, 2012, **20**(7): 1948−1963

6 Wilson K W, Raj B, Smaragdis P, Divakaran A. Speech denoising using nonnegative matrix factorization with priors. In: Proceedings of the 2008 IEEE International Conference on Acoustics, Speech and Signal Processing. Las Vegas, USA: IEEE, 2008. 4029−4032

7 Sun C L, Zhu Q, Wan M H. A novel speech enhancement method based on constrained low-rank and sparse matrix decomposition. *Speech Communication*, 2014, **60**: 44−55

8 Sun M, Li Y N, Gemmeke J, Zhang X W. Speech enhancement under low SNR conditions via noise estimation using sparse and low-rank NMF with Kullback-Leibler divergence. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2015, **23**(7): 1233−1242

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 Huang P S, Kim M, Hasegawa-Johnson M, Smaragdis P. Joint optimization of masks and deep recurrent neural networks for monaural source separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2015, **23**(12): 2136−2147

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

12 Sun M, Zhang X W, Van hamme H, Zheng T F. Unseen noise estimation using separable deep auto encoder for speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2016, **24**(1): 93−104

13 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

14 Narayanan A, Wang D L. Improving robustness of deep neural network acoustic models via speech separation and joint adaptive training. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2015, **23**(1): 92−101

15 Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets. *Neural Computation*, 2006, **18**(7): 1527−1554

16 Bengio Y. Learning deep architectures for AI. *Foundations and Trends® in Machine Learning*, 2009, **2**(1): 1−127

17 Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks. In: Proceedings of the 14th International Conference on Artificial Intelligence and Statistics. Fort Lauderdale, USA: JMLR, 2011. 315−323

18 Zhang Yong, Liu Yi, Liu Hong. A two-stage speech enhancement algorithm combined with human auditory perception. *Journal of Signal Processing*, 2014, **30**(4): 363−373
(张勇, 刘轶, 刘宏. 结合人耳听觉感知的两级语音增强算法. 信号处理, 2014, **30**(4): 363−373)

19 Johnston J D. Transform coding of audio signals using perceptual noise criteria. *IEEE Journal on Selected Areas in Communications*, 1988, **6**(2): 314−323

20 Udrea R M, Vizireanu N D, Ciochina S. An improved spectral subtraction method for speech enhancement using a perceptual weighting filter. *Digital Signal Processing*, 2008, **18**(4): 581−587

21 Hu Y, Loizou P C. Incorporating a psychoacoustical model in frequency domain speech enhancement. *IEEE Signal Processing Letters*, 2004, **11**(2): 270−273

22 Rix A W, Beerends J G, 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. Salt Lake City, USA: IEEE, 2001. 749−752

23 Zou Xia, Chen Liang, Zhang Xiong-Wei. Speech enhancement with Gamma speech modeling. *Journal on Communications*, 2006, **27**(10): 118−123
(邹霞, 陈亮, 张雄伟. 基于 Gamma 语音模型的语音增强算法. 通信学报, 2006, **27**(10): 118−123)

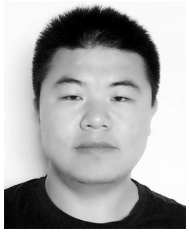
24 Hu Y, Loizou P C. Evaluation of objective quality measures for speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, 2008, **16**(1): 229−238

25 Huang P S, Kim M, Hasegawa-Johnson M, Smaragdis P. Deep learning for monaural speech separation. In: Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing. Florence, Italy: IEEE, 2014. 1562−1566

韩 伟 解放军理工大学指挥信息系统学院博士研究生. 2013 年获得解放军理工大学硕士学位. 主要研究方向为语音信号处理技术, 深度学习和语音分离.
E-mail: lan3533065@163.com
(**HAN Wei** Ph. D. candidate at the College of Command Information System, PLA University of Science and Technology. He received his master degree from PLA University of Science and Technology in 2013. His research interest covers acoustic and speech signal processing, deep learning and speech separation.)

张雄伟 解放军理工大学指挥信息系统学院教授. 1992 年获得南京通信工程学院博士学位. 主要研究方向为智能信息处理, 语音与图像信号处理, 数字通信. 本文通信作者.
E-mail: xwzhang9898@163.com
(**ZHANG Xiong-Wei** Professor at the College of Command Information System, PLA University of Science and Technology. He received his Ph. D. degree from Nanjing Institute of Communication Engineering in 1992. His research interest covers intelligence information processing, speech and image signal processing, and telecommunication systems. Corresponding author of this paper.)

闵 刚 解放军理工大学指挥信息系统学院博士研究生. 西安通信学院讲师. 2008 年获得解放军理工大学硕士学位. 主要研究方向为语音信号处理理论与技术, 语音编码, 语音增强.
E-mail: mgxaty@gmail.com
(**MIN Gang** Ph. D. candidate at the College of Command Information System, PLA University of Science and Technology and lecturer at Xi′an Communications Institute. He received his master degree from PLA University of Science and Technology in 2008. His research interest covers acoustic and speech signal processing theory and techniques, speech coding and speech enhancement.)



张启业 解放军 96637 部队助理工程师. 2013 年获得解放军理工大学硕士学位. 主要研究方向为光通信理论与技术.
E-mail: wangwangzhang555@163.com
(**ZHANG Qi-Ye** Assistant engineer at the Unit 96637 of PLA. He received his master degree from PLA University of Science and Technology in 2013. His research interest covers optical communication theory and techniques.)