文章编号：1003-0530(2024)07-1197-11

神经网络辅助估计先验语音存在概率的多通道降噪方法

雷 菁[1,2] 王劲夫[1,2] 杨飞然[1] 杨 军*[1,2]
（1. 中国科学院声学研究所，北京 100190；2. 中国科学院大学，北京 100049）

摘 要：噪声功率谱密度矩阵的估计在波束形成中非常关键。基于多通道语音存在概率(Multichannel Speech Presence Probability，MCSPP)估计噪声功率谱密度矩阵的方法，利用语音存在概率逐帧更新噪声功率谱密度矩阵。因此，语音存在概率的精度直接影响到噪声功率谱密度矩阵的估计精度。传统方法估计语音存在概率时依赖于噪声平稳假设。在变化较快的非平稳噪声上，估计的语音存在概率存在拖尾现象，这会导致降噪效果变差。本文从理论上解释了传统方法估计语音存在概率的拖尾现象成因。传统方法中语音存在概率由长期信噪比(Signal to Noise Ratio，SNR)线性映射得到，而本文证明当语音存在时当前时刻的长期信噪比仅为上一时刻长期信噪比的小幅衰减。当噪声快速变化时，长期信噪比变化缓慢，这导致语音存在概率出现拖尾现象。为解决该问题，本文提出了一种神经网络辅助估计先验语音存在概率的多通道降噪方法。所提方法利用时域卷积网络(Temporal Convolutional Network，TCN)来估计单通道观测信号的先验语音存在概率，而后利用多通道观测信号的空间信息来改善先验语音存在概率的估计。时域卷积网络估计先验语音存在概率不依赖于噪声的平稳假设，提升了噪声功率谱密度矩阵估计的精度。本文在CHiME-3数据集上进行测试，当SNR为5 dB时，所提方法取得的PESQ相比传统方法提升了0.09，fwSegSNR提升了0.78，COVL提升了0.08。结果表明，所提方法在非平稳噪声情况下能取得更好的降噪效果。

关键词：多通道降噪；神经网络；语音存在概率
中图分类号：TN912 **文献标识码：**A **DOI：**10.16798/j.issn.1003-0530.2024.07.002

引用格式：雷菁，王劲夫，杨飞然，等. 神经网络辅助估计先验语音存在概率的多通道降噪方法[J]. 信号处理，2024，40(7)：1197-1207. DOI：10.16798/j.issn.1003-0530.2024.07.002.
Reference format：LEI Jing，WANG Jinfu，YANG Feiran，et al. NN-supported a priori speech presence probability estimation for multichannel noise reduction[J]. Journal of Signal Processing，2024，40(7)：1197-1207. DOI: 10.16798/j.issn.1003-0530.2024.07.002.

NN-Supported a Priori Speech Presence Probability Estimation for Multichannel Noise Reduction

LEI Jing[1,2] WANG Jinfu[1,2] YANG Feiran[1] YANG Jun*[1,2]
（1. Institute of Acoustics，Chinese Academy of Sciences，Beijing 100190，China；
2. University of Chinese Academy of Sciences，Beijing 100049，China）

Abstract: The estimation of the noise power spectral density matrix is crucial in beamforming-based multichannel noise reduction methods. The multichannel speech presence probability（MCSPP）can be used to continuously control the ad-

收稿日期：2023-07-31；修回日期：2023-09-21
*通信作者：杨军 jyang@mail.ioa.ac.cn *Corresponding Author：YANG Jun，jyang@mail.ioa.ac.cn
基金项目：国家自然科学基金(62171438)；北京市自然科学基金-小米创新联合基金(L223032)；中国科学院声学研究所自主部署“前沿探索”类项目(QYTS202111)
Foundation Items：The National Natural Science Foundation of China（62171438）；Beijing Natural Science Foundation（L223032）；IACAS Frontier Exploration Project（QYTS202111）

aptation of the noise power spectral density matrix. Accordingly, the estimation accuracy of the noise power spectral density matrix is directly related to the accuracy of the speech presence probability estimation. Traditional techniques for estimating speech presence probability are based on the assumption of stationary noise. However, they frequently encounter a parameter trailing issue when dealing with non-stationary noise, leading to diminished noise suppression in practical applications. In this study, we first theoretically explain the rationale for the trailing problem in traditional methods for speech presence probability estimation. Speech presence probability is linearly related to the long-term signal-to-noise ratio (SNR) in traditional methods. Furthermore, we found that the long-term SNR of the current frame is only a small attenuation of the long-term SNR of the last frame when speech exists. When noise changes rapidly, the long-term SNR changes slowly, resulting in estimation trailing problem in the estimated speech presence probability. To address this problem, we proposed using the temporal convolutional network (TCN) to estimate the a priori speech presence probability. Furthermore, by integrating the estimated a priori speech presence probability into the MCSPP framework, we achieve a more accurate estimation of the posterior speech presence probability. TCN can directly estimate speech presence probability without relying on the noise stationary assumption, and the trailing problem can be effectively avoided. Therefore, a priori speech presence probability estimated by TCN can improve the accuracy of the noise power spectral density matrix estimation with non-stationary noise. The performance of the different methods was assessed using the CHiME-3 dataset. Simulation results demonstrate that the proposed method outperforms other methods in terms of noise reduction and speech quality in non-stationary noise environments. Specifically, the proposed method achieved a PESQ improvement of 0.09, a fwSegSNR improvement of 0.78, and a COVL improvement of 0.08 over the traditional method on the test dataset with an SNR of 5 dB.

Key words: multichannel noise reduction; neural network; speech presence probability

1 引言

在语音通信及语音交互场景中,期望语音常常被各种噪声污染。语音增强技术可以抑制带噪语音中的背景噪声,提升语音的质量。

多通道语音增强可额外利用空间信息,因此可获得比单通道语音增强[1-2]更好的降噪效果。多通道维纳滤波、带参数的维纳滤波、最小方差无失真响应滤波(Minimum Variance Distortionless Response, MVDR)等都属于典型的多通道语音增强方法。在这些方法中,估计多通道噪声的功率谱密度矩阵非常关键。在两通道的情况下,文献[3-5]提出进行语音活动性检测,在语音不存在时更新多通道的噪声功率谱密度矩阵的非对角项,对角项则以同样的方式更新或者由最小统计量方法估计。由于语音存在时噪声功率谱密度矩阵没有更新,可能会导致噪声功率谱密度矩阵估计不够准确。文献[6]使用交叉功率谱减法分别对两路信号进行增强,不再需要语音活动性检测,而只需要利用最小追踪方法估计噪声功率谱密度矩阵的非对角项。文献[7-9]利用噪声场扩散的假设,推导了噪声或语音互功率谱密度与扩散噪声场相干函数和观测信号之间的关系。这类方法利用扩散噪声场相干函数和观测信号计算噪声或语音功率谱密度矩阵,可用于抑制非平稳扩散噪声。此外,也有非盲方法利用目标声源方向信息来估计噪声参考信号,进而利用噪声参考信号估计噪声功率谱密度矩阵中的非对角项[10]。多通道语音存在概率(Multichannel Speech Presence Probability, MCSPP)方法利用语音存在概率来确定更新噪声功率谱密度矩阵所需的遗忘因子[11]。该遗忘因子与语音存在概率成正比,使得噪声功率谱密度矩阵在语音存在时更新速率较小,从而降低语音对噪声功率谱密度矩阵估计的干扰。为了得到准确的语音存在概率,MCSPP方法利用贝叶斯准则对先验语音存在概率进行优化得到后验语音存在概率。但现有的先验语音存在概率的计算方法依赖于噪声平稳假设,在噪声非平稳的情况下会出现拖尾现象。近年来,深度学习方法取得了较大的进步[12]。在这类方法中,神经网络被用于估计目标信号的频谱[13-17]或者一个作用于观测信号频谱的掩膜[18-21]。当利用神经网络估计作用于观测信号频谱的掩膜时,通常使用在0~1之间的理想比例掩膜(Ideal Ratio Mask, IRM)。神经网络估计的掩膜实际上可以表征观测信号频谱中每个时频点处存在说话人的概率,即我们所需要的语音存在概率。由于深度学习方法利用数据驱动模型,并不依赖于特定的先验假设,所以该类方法在非平稳噪声情况下性能表现相对稳定。

本文首先对估计先验语音存在概率的传统方法[22]进行了分析。语音存在概率由长期信噪比线

性映射得到。我们证明当语音存在时，当前时刻长期信噪比仅为上一时刻长期信噪比的小幅衰减，而这会导致语音存在概率不能追踪噪声的快速变化。为解决该问题，我们提出利用时域卷积网络（Temporal Convolutional Network，TCN）来估计先验语音存在概率。所提方法在噪声非平稳情况下仍能得到精确的语音存在概率估计，从而可以更准确地估计噪声功率谱密度矩阵。实验表明，对于非平稳噪声场景，所提 TCN-MCSPP 相比单独的 TCN 网络和单独的 MCSPP 能够取得更好的效果。

2　MCSPP

2.1　信号模型

考虑含 N 个传声器的阵列及单目标声源场景，则观测信号可表示为

$$y_n(t)=g_n(t)*s(t)+v_n(t)=x_n(t)+v_n(t) \tag{1}$$

其中 * 为卷积算子，$g_n(t)$ 为声源到第 n 个传声器的信道冲击响应，$s(t)$ 为目标语音信号，$v_n(t)$ 为第 n 个传声器处的噪声信号，$x_n(t)$ 为第 n 个传声器处的语音信号。在短时傅里叶变换（Short Time Fourier Transform，STFT）域有

$$Y_n(k,l)=X_n(k,l)+V_n(k,l) \tag{2}$$

其中 k,l 分别为频率和时间帧序号，$Y_n(k,l)$，$X_n(k,l)$，$V_n(k,l)$ 分别表示第 n 个传声器处观测信号、语音信号和噪声信号的短时傅里叶变换。

2.2　MVDR

波束形成方法将一组滤波器作用于观测信号来提取目标语音信号。定义传声器信号向量 $\boldsymbol{y}(k,l)=[Y_1(k,l),Y_2(k,l),\cdots,Y_N(k,l)]^{\mathrm{T}}$，则波束的输出为

$$\hat{S}(k,l)=\boldsymbol{w}^{\mathrm{H}}(k,l)\boldsymbol{y}(k,l) \tag{3}$$

其中 $\hat{S}(k,l)$ 为估计的目标信号的短时傅里叶变换，$\boldsymbol{w}(k,l)$ 为长度为 N 的滤波器系数向量。

一种广泛使用的波束形成器是 MVDR。MVDR 在目标方向信号无失真的条件下最小化输出功率。我们定义噪声向量 $\boldsymbol{v}(k,l)=[V_1(k,l),V_2(k,l),\cdots,V_N(k,l)]^{\mathrm{T}}$，则噪声功率谱密度矩阵和信号功率谱密度矩阵分别为 $\boldsymbol{\Phi}_{vv}(k,l)=\mathrm{E}\{\boldsymbol{v}(k,l)\boldsymbol{v}^{\mathrm{H}}(k,l)\}$ 和 $\boldsymbol{\Phi}_{yy}(k,l)=\mathrm{E}\{\boldsymbol{y}(k,l)\boldsymbol{y}^{\mathrm{H}}(k,l)\}$。MVDR 的滤波器系数向量为[23]

$$\boldsymbol{w}(k,l)=\frac{(\hat{\boldsymbol{\Phi}}_{vv}^{-1}(k,l)\hat{\boldsymbol{\Phi}}_{yy}(k,l)-\boldsymbol{I}_N)\boldsymbol{u}_1}{\mathrm{tr}(\hat{\boldsymbol{\Phi}}_{vv}^{-1}(k,l)\hat{\boldsymbol{\Phi}}_{yy}(k,l))-N} \tag{4}$$

其中 $\boldsymbol{I}_N$ 是维数为 $N\times N$ 的单位矩阵，$\boldsymbol{u}_1=[1,0,\cdots,0]^{\mathrm{T}}$，$\mathrm{tr}(\cdot)$ 表示求矩阵的迹。

2.3　MCSPP 估计

为了计算 $\boldsymbol{w}(k,l)$，我们需要知道观测信号功率谱密度矩阵 $\boldsymbol{\Phi}_{yy}(k,l)$ 和噪声功率谱密度矩阵 $\boldsymbol{\Phi}_{vv}(k,l)$。$\boldsymbol{\Phi}_{yy}(k,l)$ 可以直接计算为

$$\boldsymbol{\Phi}_{yy}(k,l)=\alpha_y(k,l)\boldsymbol{\Phi}_{yy}(k,l-1)+\left[1-\alpha_y(k,l)\right]\boldsymbol{y}(k,l)\boldsymbol{y}^{\mathrm{H}}(k,l) \tag{5}$$

其中 $\alpha_y(k,l)$ 为带噪信号估计的遗忘因子，一般选择为时频点无关的常数 α_y。噪声功率谱密度矩阵 $\boldsymbol{\Phi}_{vv}(k,l)$ 的估计非常关键，估计方法为[22]

$$\boldsymbol{\Phi}_{vv}(k,l)=\tilde{\alpha}_v(k,l)\boldsymbol{\Phi}_{vv}(k,l-1)+\left[1-\tilde{\alpha}_v(k,l)\right]\boldsymbol{y}(k,l)\boldsymbol{y}^{\mathrm{H}}(k,l) \tag{6}$$

其中 $\tilde{\alpha}_v(k,l)$ 为噪声估计的遗忘因子。当语音存在时，遗忘因子 $\tilde{\alpha}_v(k,l)$ 的值应该较大，此时当前帧信号对噪声功率谱密度矩阵更新影响小，可以减轻语音失真问题。遗忘因子 $\tilde{\alpha}_v(k,l)$ 与语音存在概率 $p(k,l)$ 的关系为

$$\tilde{\alpha}_v(k,l)=\alpha_v+(1-\alpha_v)p(k,l) \tag{7}$$

其中 α_v 为平滑因子。$p(k,l)$ 可以视作观测信号为 $\boldsymbol{y}(k,l)$ 的条件下语音存在的后验概率。MCSPP 方法假设语音和噪声都符合多元高斯分布[11]，则根据贝叶斯准则得到 $p(k,l)$

$$p(k,l)=\left\{1+\frac{q(k,l)}{1-q(k,l)}\left[1+\xi(k,l)\right]\exp\left[-\frac{\beta(k,l)}{1+\xi(k,l)}\right]\right\}^{-1} \tag{8}$$

其中 $\xi(k,l)=\mathrm{tr}\left[\boldsymbol{\Phi}_{vv}^{-1}(k,l)\boldsymbol{\Phi}_{xx}(k,l)\right]$，$\beta(k,l)=\boldsymbol{y}^{\mathrm{H}}(k,l)\boldsymbol{\Phi}_{vv}^{-1}(k,l)\boldsymbol{\Phi}_{xx}(k,l)\boldsymbol{\Phi}_{vv}^{-1}(k,l)\boldsymbol{y}(k,l)$，$q(k,l)$ 为先验语音不存在概率。

2.4　$q(k,l)$ 估计及其存在问题

先验语音不存在概率 $q(k,l)$ 的计算方式有两种[22]。一种方法是设置 $q(k,l)$ 为一个固定值，但是这类方法不能适应不同的噪声环境。另一种方法则是利用观测信号自适应估计 $q(k,l)$。在自适应计算 $q(k,l)$ 时，需要得到当前时刻噪声功率谱密度矩阵 $\boldsymbol{\Phi}_{vv}(k,l)$ 的估计，因而需要对 $q(k,l)$ 进行两次估计。第一次估计利用上一时刻的噪声功率谱密度矩阵 $\boldsymbol{\Phi}_{vv}^{-1}(k,l-1)$ 计算先验概率的第一次估计 $q^{(1)}(k,l)$ 和当前时刻的噪声功率谱密度矩阵 $\boldsymbol{\Phi}_{vv}^{(1)}(k,l)$。第二次估计利用当前时刻的噪声功率谱密度矩阵 $\boldsymbol{\Phi}_{vv}^{(1)}(k,l)$ 来计算最终的先验概率 $q(k,l)$。传统方法首先估计瞬时信噪比 $\psi^{(1)}(k,l)$ 和长期信噪

比 $\tilde{\psi}^{(\mathrm{l})}(k,l)$[22]

$$\psi^{(\mathrm{l})}(k,l)=\boldsymbol{y}^{\mathrm{H}}(k,l)\boldsymbol{\Phi}_{vv}^{-1}(k,l-1)\boldsymbol{y}(k,l) \tag{9}$$

$$\tilde{\psi}^{(\mathrm{l})}(k,l)=\mathrm{tr}(\boldsymbol{\Phi}_{vv}^{-1}(k,l-1)\boldsymbol{\Phi}_{yy}(k,l)) \tag{10}$$

利用 $\psi^{(\mathrm{l})}(k,l)$ 和 $\tilde{\psi}^{(\mathrm{l})}(k,l)$，可以计算先验概率 $q^{(\mathrm{l})}(k,l)$

$$\hat{q}(k,l)=\begin{cases}1 & \text{if } \tilde{\psi}(k,l)<N \text{ and } \psi(k,l)<\psi_0\\ \dfrac{\tilde{\psi}_0-\tilde{\psi}(k,l)}{\tilde{\psi}_0-N} & \text{if } N\leqslant\tilde{\psi}(k,l)<\tilde{\psi}_0 \text{ and } \psi(k,l)<\psi_0\\ 0 & \text{else}\end{cases} \tag{11}$$

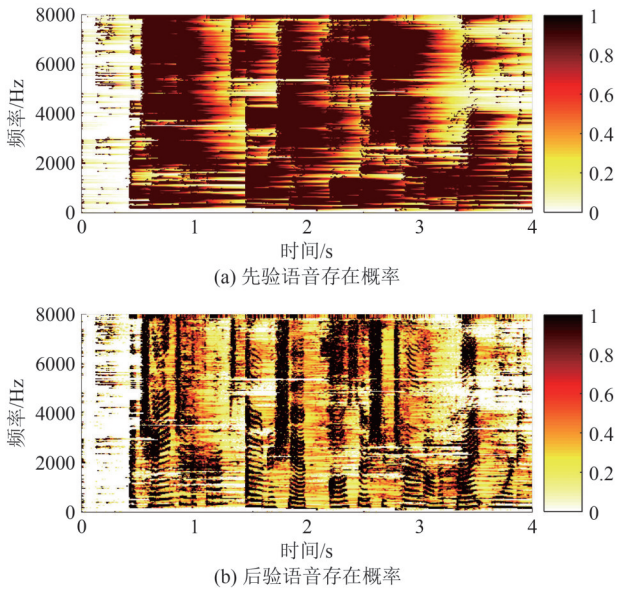
其中 ψ_0 和 $\tilde{\psi}_0$ 为控制虚警率的阈值。

传统方法[22]首先利用式(9)~(11)估计瞬时信噪比 $\psi^{(\mathrm{l})}(k,l)$、长期信噪比 $\tilde{\psi}^{(\mathrm{l})}(k,l)$ 和先验概率 $q^{(\mathrm{l})}(k,l)$，然后利用式(9)~(11)估计瞬时信噪比 $\psi^{(2)}(k,l)$、长期信噪比 $\tilde{\psi}^{(2)}(k,l)$ 和先验概率 $q(k,l)$。其中第一次估计使用上一时刻的噪声功率谱密度矩阵 $\boldsymbol{\Phi}_{vv}^{-1}(k,l-1)$，而第二次估计使用当前时刻的噪声功率谱密度矩阵 $\boldsymbol{\Phi}_{vv}^{(\mathrm{l})}(k,l)$。为了得到当前时刻的噪声功率谱密度矩阵 $\boldsymbol{\Phi}_{vv}^{(\mathrm{l})}(k,l)$，将先验概率 $q^{(\mathrm{l})}(k,l)$ 和上一时刻的噪声功率谱密度矩阵 $\boldsymbol{\Phi}_{vv}^{-1}(k,l-1)$ 代入公式(8)，得到语音存在概率的第一次估计 $p^{(\mathrm{l})}(k,l)$。平滑后的语音存在概率 $\hat{p}(k,l)=\alpha_p p(k,l-1)+(1-\alpha_p)p^{(\mathrm{l})}(k,l)$ 代入式(7)和(6)可得到遗忘因子 $\tilde{\alpha}_v^{(\mathrm{l})}(k,l)$ 和噪声功率谱密度矩阵 $\boldsymbol{\Phi}_{vv}^{(\mathrm{l})}(k,l)$，其中 α_p 为平滑因子。先验概率 $q(k,l)$ 和噪声功率谱密度矩阵 $\boldsymbol{\Phi}_{vv}^{(\mathrm{l})}(k,l)$ 代入式(8)可得到后验语音存在概率 $p(k,l)$。

传统方法[22]基于平稳噪声假设，所以在非平稳噪声环境下性能大幅降低。图1展示了传统方法估计的先验语音存在概率 $1-q(k,l)$ 和后验语音存在概率 $p(k,l)$。

图1对应的观测信号为一段采样率为16 kHz，长度为4 s的4通道带噪语音，噪声为采集自咖啡馆中的非平稳噪声。从图1可以看出，先验概率 $q(k,l)$ 在非平稳噪声存在的情况下出现了严重的拖尾现象。由于先验概率 $q(k,l)$ 的估计不准确，直接导致后验概率 $p(k,l)$ 的估计也出现拖尾现象。这会导致后续噪声功率谱密度矩阵的估计不精确，可能产生较大的残留。

如式(11)所示，先验概率 $q(k,l)$ 是长期信噪比 $\tilde{\psi}^{(2)}(k,l)$ 的线性映射，因此我们可以从长期信噪比 $\tilde{\psi}^{(2)}(k,l)$ 入手分析拖尾现象的成因。我们经过推导证明(推导过程见附录)，当 $l-1$ 时刻存在语音时，l



(a) 先验语音存在概率

(b) 后验语音存在概率

图1 非平稳噪声下的传统方法估计的先验概率 $1-q(k,l)$ (a)和后验概率 $p(k,l)$ (b)

Fig. 1 Priori speech presence probability $1-q(k,l)$ (a) and posteriori speech presence probability $p(k,l)$ (b) estimated by traditional method in the presence of nonstationary noise

时刻的长期信噪比 $\tilde{\psi}^{(2)}(k,l)$ 与上一时刻长期信噪比 $\tilde{\psi}(k,l-1)$ 关系为

$$\tilde{\psi}^{(2)}(k,l)\approx\alpha_y\tilde{\psi}(k,l-1) \tag{12}$$

其中 $\alpha_y=0.95$。可以看出当前时刻的长期信噪比 $\tilde{\psi}^{(2)}(k,l)$ 仅为上一时刻长期信噪比 $\tilde{\psi}(k,l-1)$ 的小幅衰减。如果噪声快速变化，当前时刻的长期信噪比 $\tilde{\psi}^{(2)}(k,l)$ 将不能及时追踪噪声的变化。而根据式(11)，先验概率 $q(k,l)$ 由当前时刻的长期信噪比 $\tilde{\psi}^{(2)}(k,l)$ 线性映射得到。长期信噪比 $\tilde{\psi}^{(2)}(k,l)$ 的缓慢变化，限制了先验概率 $q(k,l)$ 的变化速度，这会导致先验概率 $q(k,l)$ 的估计不准确，出现拖尾现象。

我们认为，这是传统方法[22]使用式(6)来更新 $\boldsymbol{\Phi}_{vv}^{(\mathrm{l})}(k,l)$ 导致的。在式(6)中，当语音存在时，遗忘因子 $\tilde{\alpha}_v(k,l)$ 较大，当前包含语音的信号对噪声功率谱密度矩阵更新影响小，这避免了语音对噪声功率谱密度矩阵估计的干扰，但同时也导致在语音存在的时段，噪声功率谱密度矩阵变化缓慢，难以追踪快速变化的噪声。

3 TCN-MCSPP

为了克服非平稳噪声的影响，本文提出利用DNN来估计先验概率[24]，替代原来MCSPP中的估

计结果。基于DNN的语音增强方法在非平稳噪声存在的情况下表现良好，但是我们前期实验发现其估计结果可能在部分情况下相对不够准确。我们提出的TCN-MCSPP方法可以结合DNN的非平稳先验估计以及MCSPP的多通道信息，进一步增强算法的性能。

时域卷积网络（Temporal Convolutional Network，TCN）最早被提出应用于计算机视觉中的行为分割任务，后被Luo Yi成功地应用于语音分离任务中[25]。考虑到其参数量小、能更好利用信号长时相关性的优点，我们采用TCN来估计先验语音存在概率。

如图2所示，TCN中的每一层为一个1-D Conv卷积块，卷积块可以分为若干个stack（即图2中的一列）。每个stack中的卷积块的空洞率逐渐增大，从而增大模型的感受野，更好地利用信号的长时相关特性。实验中TCN网络采用3个stack，每个stack 8个1-D Conv卷积块，每个stack中卷积块空洞率d为1、2、4……、2^7。

1-D Conv模块结构如图3所示。1-D Conv有output和skip connection两路输出，其中output作为下一个卷积块的输入，skip connection则被加起来作为TCN的输出。实验中，TCN的输入为第一个通道的观测信号频谱$Y_1(k,l)$，输出为观测信号对应的掩膜估计$\hat{M}$。

训练的过程中，观测信号对应的掩膜估计$\hat{M}$与理想比值掩膜M的均方误差作为损失，用以指导网络的训练，计算公式如下

$$L(M,\hat{M})=\sum_{l=1}^{L}\sum_{k=1}^{K}(M(k,l)-\hat{M}(k,l))^2 \tag{13}$$

理想比值掩膜M可以由参考语音信号$X_1(k,l)$和噪声信号$V_1(k,l)$计算得到

$$M(k,l)=\frac{|X_1(k,l)|^2}{|X_1(k,l)|^2+|V_1(k,l)|^2} \tag{14}$$

我们将TCN估计的掩膜$\hat{M}$转换为先验语音不存在概率

$$q_{NN}(k,l)=1-\hat{M}(k,l) \tag{15}$$

表1展示了利用$q_{NN}(k,l)$计算后验语音存在概率$p(k,l)$算法流程。

4　实验设置及结果

数据集采用CHiME3[26]。CHiME3中的噪声采集自公交车、咖啡馆、路口、步行区四种环境，基本涵盖了实际中各类场景的非平稳噪声。其中，公交车噪声1.37 h，咖啡馆噪声1.45 h，路口噪声1.43 h，步行区噪声1.39 h。干净语音与独立采集的噪声混合得到混合信号。采样率为16 kHz，每条语音长度为4 s。其中训练集中使用独立采集的噪声，验证集与测试集中的噪声由带噪信号减去对应的目标语音。

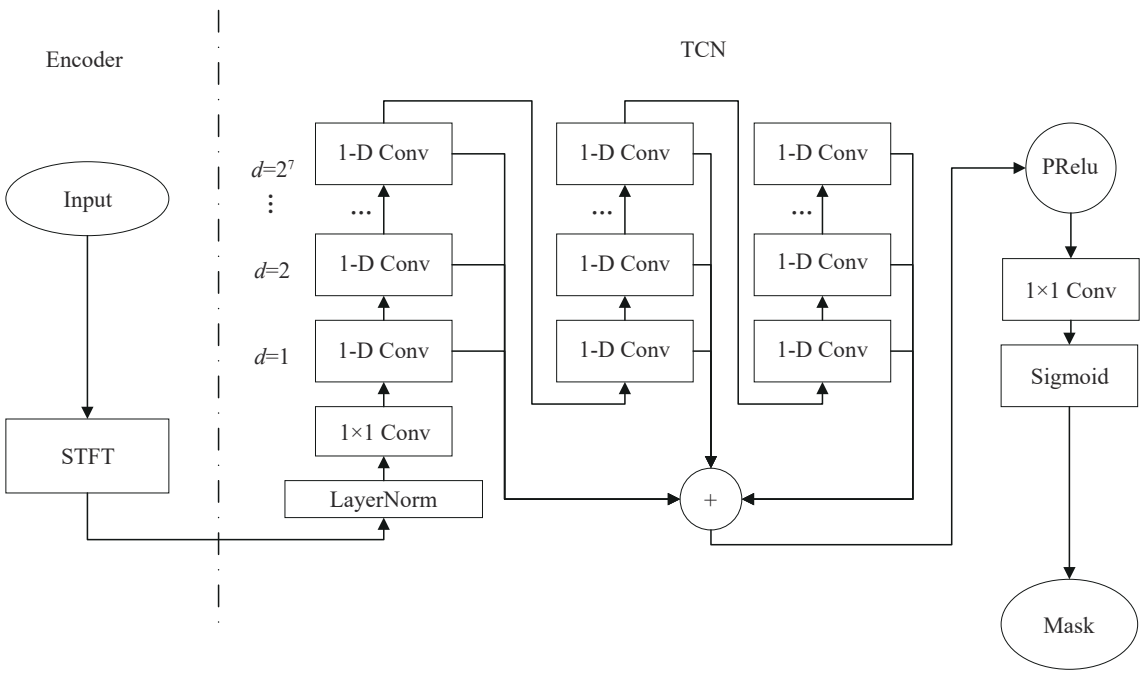


图2　TCN网络结构

Fig. 2　Architecture of the TCN network

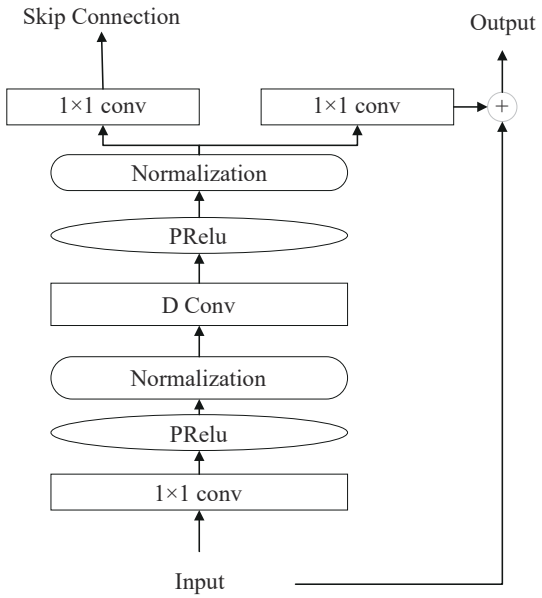


图3 1-D Conv模块结构

Fig. 3 Architecture of the 1-D Conv module

表1 利用$q_{NN}(k,l)$计算后验语音存在概率$p(k,l)$算法流程

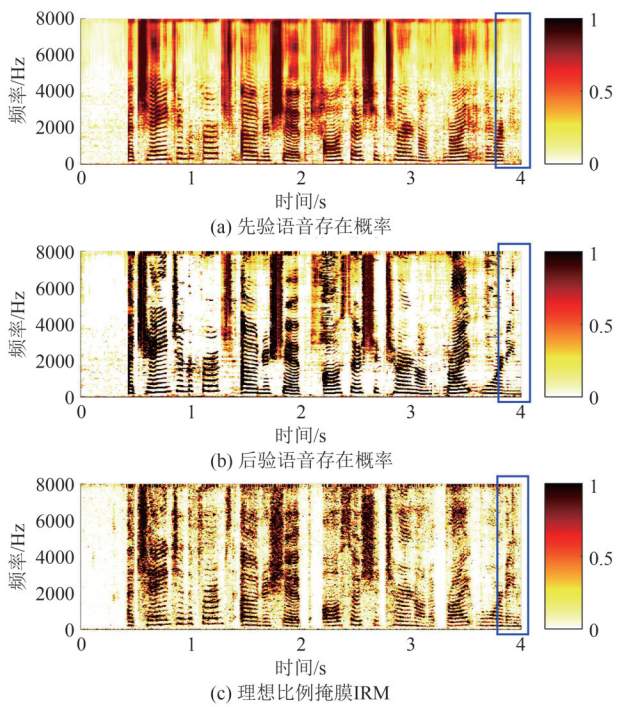
Tab. 1 Computation of posteriori speech presence probability $p(k,l)$ using $q_{NN}(k,l)$

所提算法
1. 将$q_{NN}(k,l)$和$\boldsymbol{\Phi}_{vv}(k,l-1)$代入式(8)中即可得到语音存在概率$p^{(l)}(k,l)$。
2. 根据$p^{(l)}(k,l)$计算遗忘因子$\tilde{\alpha}_v=\alpha_v+(1-\alpha_v)p^{(l)}(k,l)$。
3. 将$\tilde{\alpha}_v$代入式(6)中即可更新噪声功率谱密度矩阵$\boldsymbol{\Phi}_{vv}^{(l)}(k,l)$。
4. 将$q_{NN}(k,l)$和$\boldsymbol{\Phi}_{vv}^{(l)}(k,l)$代入式(8)中即可得到后验语音存在概率$p(k,l)$。

在训练阶段,我们仅使用训练集中第一个通道的数据来训练TCN。在测试阶段,我们使用前4个通道的数据,并设置语音与噪声混合时的信噪比分别为0 dB、5 dB、10 dB。训练集、验证集和测试集的总时长分别为15.15 h、2.88 h和2.01 h。训练时batch size为16,采用Adam优化器,初始学习率为0.001。

我们采用了四种测试指标,包括客观语音质量评估(Perceptual Evaluation of the Speech Quality, PESQ)[27],降噪系数(Noise Reduction factor, NR)[28],对数频谱距离(Log Spectral Distance, LSD)[29],频带加权分段信噪比(Frequency-Weighted Segmental Signal-to-Noise-Ratio, fwSegSNR)[30],信号失真测度(CSIG)[30],噪声失真测度(CBAK)[30]和综合质量测度(COVL)[30]。其中LSD指标越小代表语音质量越高,其余指标数值越大代表语音质量越高。

图4展示了所提方法估计的先验语音存在概率$1-q_{NN}(k,l)$和后验语音存在概率$p(k,l)$以及对应的IRM。图4对应的观测信号与图1中的观测信号相同,为一段采样率为16 kHz,长度为4 s的4通道带噪语音。可以发现,和图1所示传统方法[22]的结果相比,利用NN估计先验概率时将不再受到拖尾问题的影响。此外,如图4中蓝框部分所示,在4 s处的中低频部分,贝叶斯公式得到的后验概率能够补足NN估计中缺少的频谱结构,进一步提升语音存在概率的估计精度。这是由于所提算法估计后验概率$p(k,l)$的过程利用了多个通道的观测信号,融合了信号中的空间信息,因而进一步改善了先验概率$1-q_{NN}(k,l)$。



(a) 先验语音存在概率

(b) 后验语音存在概率

(c) 理想比例掩膜IRM

图4 非平稳噪声下所提方法估计先验概率(a),后验概率(b),IRM(c)

Fig. 4 Priori speech presence probability (a), posteriori speech presence probability (b) and IRM (c) estimated by proposal in the presence of nonstationary noise

我们在测试集上对所提算法和另外两种方法进行了测试,其中NN方法直接利用TCN估计的掩膜$\hat{M}$作用于$Y_1(k,l)$计算目标信号的频谱,MCSPP为传统方法[22]。需要说明的是本文的目的并不是开发一种基于深度学习的多通道语音增强,而是对传统的先验语音存在概率估计方法[22]存在问题进行分析并指出利用深度学习估计该参数的优势,因而本文侧重

比较传统算法[22]和本文所提方法的性能。图5展示了信噪比在0 dB、5 dB和10 dB的测试集上三种方法取得的PESQ、NR、LSD、fwSegSNR、CSIG、CBAK和COVL。当信噪比在5 dB时，所提算法的PESQ、fwSegSNR、CSIG、CBAK和COVL相比MCSPP分别提升了0.09，0.78，0.1，0.14和0.08，LSD相比MCSPP降低了0.21。这说明NN得到的mask作为先验概率比传统方法得到的先验概率更加准确。同时，所提算法的PESQ、fwSegSNR、CSIG、CBAK和COVL相比NN分别提升了0.62，2.55，0.73，0.37和0.65，LSD相比NN降低了0.37。这说明利用空间信息能够进一步提升NN处理的音频质量。此外，在信噪比取不同值时，所提算法在LSD、fwSegSNR、CSIG和CBAK上的表现均优于NN和MCSPP。

需要指出，尽管所提算法的PESQ、LSD和fwSegSNR均有了明显的提升，但在NR上的表现却落后于MCSPP。这提示我们，MCSPP可能存在对噪声的过估计，而所提算法改善了这一情况。值得注意的是，当信噪比在0 dB时，虽然所提算法在LSD和fwSegSNR上的表现优于MCSPP，但所提算法的PESQ比MCSPP略低。这可能是信噪比较低时NN估计的mask不够准确导致的。

为了验证所提方法在不同类型噪声存在情况下的性能，我们按照噪声类型划分测试集，并在SNR=5 dB的情况下分别对三种方法进行测试。其中，bus、caf、ped和str分别代表采集自公交车、咖啡馆、步行区和路口环境的噪声。测试结果如表2所示。可以发现，所提方法除NR以外的各项指标均

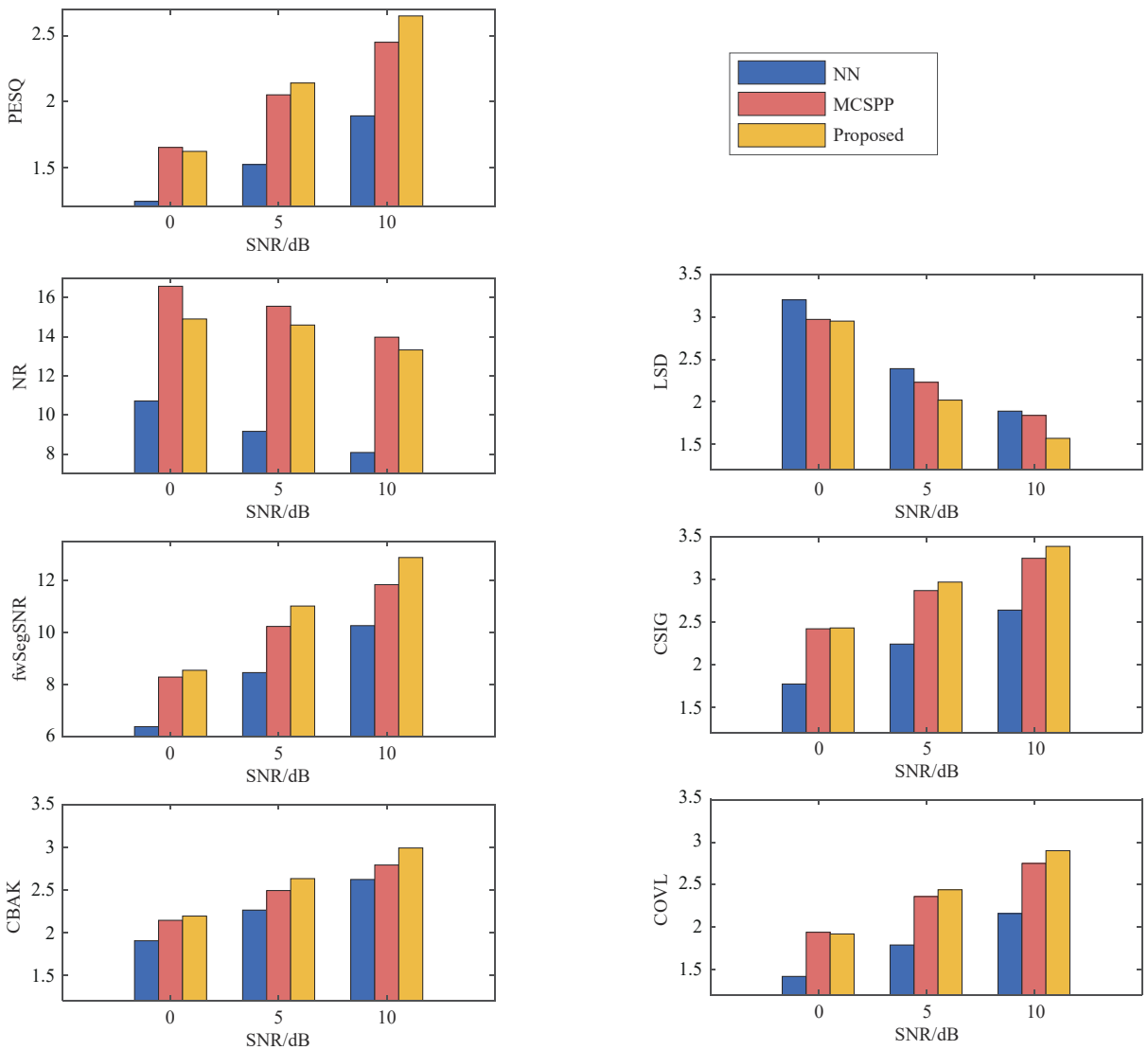


图5　不同信噪比的测试集上三种方法取得的PESQ、NR、LSD、fwSegSNR、CSIG、CBAK和COVL

Fig. 5　PESQ, NR, LSD, fwSegSNR, CSIG, CBAK and COVL of three methods for different SNRs

表 2 不同类型噪声下三种方法的测试结果

Tab. 2 Results of three methods on test sets of different noise types

Noise Type	Method	PESQ	NR	LSD	fwSegSNR	CSIG	CBAK	COVL
bus	Proposed	**2.34**	15.23	**1.87**	**11.97**	**3.21**	**2.74**	**2.66**
	MCSPP	2.20	**16.52**	1.99	11.10	3.08	2.59	2.54
	NN	1.67	9.54	2.17	9.36	2.45	2.36	1.97
caf	Proposed	**1.97**	13.75	**2.07**	**10.52**	**2.80**	**2.54**	**2.28**
	MCSPP	1.89	**14.29**	2.40	9.68	2.72	2.39	2.21
	NN	1.42	8.78	2.45	8.06	2.16	2.20	1.71
ped	Proposed	**2.07**	14.42	**2.23**	**10.14**	**2.79**	**2.56**	**2.32**
	MCSPP	2.01	**15.29**	2.44	9.50	2.73	2.43	2.27
	NN	1.44	9.27	2.57	7.78	2.08	2.19	1.68
str	Proposed	**2.18**	14.94	**1.93**	**11.41**	**3.06**	**2.66**	**2.51**
	MCSPP	2.08	**16.01**	2.11	10.63	2.95	2.53	2.42
	NN	1.53	9.15	2.36	8.66	2.28	2.29	1.82

优于 MCSPP 和 NN。此外，针对不同类型的噪声，所提方法展现的性能也不同。在 bus，caf，ped 和 str 噪声存在的情况下，所提方法取得的 PESQ 相比 MCSPP 分别提升了 0.14，0.08，0.06 和 0.1，CSIG 相比 MCSPP 分别提升了 0.13，0.08，0.06 和 0.11，COVL 相比 MCSPP 分别提升了 0.12，0.07，0.05 和 0.09。我们认为两方面的原因导致了这种差异。首先，在不同类型噪声存在情况下，NN 方法的性能存在差异，而所提方法的性能会直接受到 NN 估计的 mask 的精度的影响。表 2 中，bus 噪声下所提方法 PESQ 提升幅度最大，相应地其 NN 取得的 PESQ 也最高。其次，不同类型非平稳噪声的变化速度存在差异。理论上，噪声变化速度越快，所提方法相对 MCSPP 的提升越大。

此外，我们还设计实验验证了 MCSPP 方法中贝叶斯迭代的效果。在所提算法中，我们对 NN 估计的先验概率 $q_{NN}(k,l)$ 进行贝叶斯迭代得到后验概率 $p(k,l)$。为了验证其中贝叶斯迭代步骤的作用，我们进行了没有贝叶斯迭代的实验（Proposed without Bayesian），即直接令后验概率 $p(k,l)=1-q_{NN}(k,l)$。表 3 展示了四种方法在测试集 SNR=5 dB 时的测试结果。如表 3 中数据所示，所提算法的 PESQ 相比 MCSPP 方法提升了 0.09，相比 NN 方法提升了 0.62。但是，没有进行贝叶斯迭代的方法（Proposed without Bayesian）的 PESQ 低于所提算法和 MCSPP 方法。这表明，直接使用 NN 估计的 mask 作为语音存在概率并不利于提升增强效果。NN 估计的 mask 只有通过贝叶斯迭代过程与多通道信息进一步融合，才能更好地辅助后续的噪声功率谱密度矩阵估计。

表 3 测试集上贝叶斯迭代的增强结果对比

Tab. 3 Comparison of enhancement result for Bayesian iteration on the test set

Method	PESQ	NR	LSD	fwSegSNR	CSIG	CBAK	COVL
Proposed	**2.14**	14.60	**2.02**	**11.02**	**2.97**	**2.63**	**2.44**
MCSPP	2.05	**15.56**	2.23	10.24	2.87	2.49	2.36
NN	1.52	9.17	2.39	8.47	2.24	2.26	1.79
Proposed without Bayesian	2.01	13.04	2.24	10.29	2.84	2.50	2.31

5　结论

传统MCSPP在进行先验概率估计时往往会存在参数估计拖尾问题，而这会直接影响多通道语音增强算法的性能。我们首先从理论上解释了拖尾现象的成因，然后给出了一种新的结合神经网络的参数估计方法。所提方法将神经网络估计的掩膜作为先验概率的估计值，然后融合多通道信息与概率模型进行后验语音存在概率的估计。相较传统MCSPP估计方法，本文所提方法可以有效避免参数估计的拖尾问题；而相较纯神经网络的方法，本文所提的方法可以更为准确地估计语音信号的存在概率。实验结果表明，该方法在非平稳噪声下相比于传统的方法有明显的优势。

附录

为了表达简洁，我们接下来的推导过程将省略频率序号 k。用 $(\boldsymbol{\Phi}_{vv}^{(\mathrm{l})}(l))^{-1}$ 替换式(9)和(10)中的 $\boldsymbol{\Phi}_{vv}^{-1}(l-1)$，得到长期信噪比的第二次估计 $\tilde{\psi}^{(2)}(l)$

$$\tilde{\psi}^{(2)}(l)=\mathrm{tr}((\boldsymbol{\Phi}_{vv}^{(\mathrm{l})}(l))^{-1}\boldsymbol{\Phi}_{yy}(l)) \tag{16}$$

由式(6)可得

$$(\boldsymbol{\Phi}_{vv}^{(\mathrm{l})}(l))^{-1}=(\tilde{\alpha}_v^{(\mathrm{l})}(l)\boldsymbol{\Phi}_{vv}(l-1)+\left[1-\tilde{\alpha}_v^{(\mathrm{l})}(l)\right]\boldsymbol{y}(l)\boldsymbol{y}^{\mathrm{H}}(l))^{-1}=$$

$$\frac{1}{\tilde{\alpha}_v^{(\mathrm{l})}(l)}\left(\boldsymbol{\Phi}_{vv}^{-1}(l-1)-\frac{\boldsymbol{\Phi}_{vv}^{-1}(l-1)y(l)y^{\mathrm{H}}(l)\boldsymbol{\Phi}_{vv}^{-1}(l-1)}{\frac{\tilde{\alpha}_v^{(\mathrm{l})}(l)}{1-\tilde{\alpha}_v^{(\mathrm{l})}(l)}+\psi^{(\mathrm{l})}(l)}\right) \tag{17}$$

将式(17)代入式(16)中，可得

$$\tilde{\psi}^{(2)}(l)=\frac{1}{\tilde{\alpha}_v^{(\mathrm{l})}(l)}\tilde{\psi}^{(\mathrm{l})}(l)-\frac{1}{\tilde{\alpha}_v^{(\mathrm{l})}(l)}\frac{\mathrm{tr}(\boldsymbol{\Phi}_{vv}^{-1}(l-1)\boldsymbol{y}(l)\boldsymbol{y}^{\mathrm{H}}(l)\boldsymbol{\Phi}_{vv}^{-1}(l-1)\boldsymbol{\Phi}_{yy}(l))}{\frac{\tilde{\alpha}_v^{(\mathrm{l})}(l)}{1-\tilde{\alpha}_v^{(\mathrm{l})}(l)}+\psi^{(\mathrm{l})}(l)} \tag{18}$$

根据 $\mathrm{tr}(\boldsymbol{AB})=\mathrm{tr}(\boldsymbol{BA})$，并将式(5)代入可得

$$\mathrm{tr}(\boldsymbol{\Phi}_{vv}^{-1}(l-1)\boldsymbol{y}(l)\boldsymbol{y}^{\mathrm{H}}(l)\boldsymbol{\Phi}_{vv}^{-1}(l-1)\boldsymbol{\Phi}_{yy}(l))=$$
$$\mathrm{tr}(\boldsymbol{y}^{\mathrm{H}}(l)\boldsymbol{\Phi}_{vv}^{-1}(l-1)\boldsymbol{\Phi}_{yy}(l)\boldsymbol{\Phi}_{vv}^{-1}(l-1)\boldsymbol{y}(l))=$$
$$\alpha_y\psi'+(1-\alpha_y)(\psi^{(\mathrm{l})}(l))^2 \tag{19}$$

其中 $\psi'=\boldsymbol{y}^{\mathrm{H}}(l)\boldsymbol{\Phi}_{vv}^{-1}(l-1)\boldsymbol{\Phi}_{yy}(l-1)\boldsymbol{\Phi}_{vv}^{-1}(l-1)\boldsymbol{y}(l)$。将式(5)代入式(10)中，可得

$$\tilde{\psi}^{(\mathrm{l})}(l)=\alpha_y\mathrm{tr}(\boldsymbol{\Phi}_{vv}^{-1}(l-1)\boldsymbol{\Phi}_{yy}(l-1))+$$
$$(1-\alpha_y)\mathrm{tr}(\boldsymbol{\Phi}_{vv}^{-1}(l-1)\boldsymbol{y}(l)\boldsymbol{y}^{\mathrm{H}}(l))=$$
$$\alpha_y\tilde{\psi}(l-1)+(1-\alpha_y)\psi^{(\mathrm{l})}(l) \tag{20}$$

将式(19)和(20)代入式(18)中，可得

$$\tilde{\psi}^{(2)}(l)=\frac{\alpha_y}{\tilde{\alpha}_v^{(\mathrm{l})}(l)}\left(\tilde{\psi}(l-1)+\frac{\frac{\tilde{\alpha}_v^{(\mathrm{l})}(l)}{\alpha_y}(1-\alpha_y)\psi^{(\mathrm{l})}(l)-(1-\tilde{\alpha}_v^{(\mathrm{l})}(l))\psi'}{\tilde{\alpha}_v^{(\mathrm{l})}(l)+(1-\tilde{\alpha}_v^{(\mathrm{l})}(l))\psi^{(\mathrm{l})}(l)}\right) \tag{21}$$

实际中 α_v、α_y 和 α_p 均取较大值，在我们的实验中取 $\alpha_v=\alpha_y=0.95\to 1$，$\alpha_p=0.6$。当 $l-1$ 时刻存在语音，$p(l-1)\approx 1$。根据式 $\hat{p}(l)=\alpha_p p(l-1)+(1-\alpha_p)p^{(\mathrm{l})}(l)$，$\alpha_p$ 较大时，经过平滑后的 $\hat{p}(l)\approx 1$。将 $\hat{p}(l)\approx 1$ 代入 $\tilde{\alpha}_v^{(\mathrm{l})}(l)=\alpha_v+(1-\alpha_v)\hat{p}(l)$，有 $\tilde{\alpha}_v^{(\mathrm{l})}(l)\approx 1$ 及 $1-\tilde{\alpha}_v^{(\mathrm{l})}(l)\approx 0$。将 $\tilde{\alpha}_v^{(\mathrm{l})}(l)\approx 1$ 及 $1-\tilde{\alpha}_v^{(\mathrm{l})}(l)\approx 0$ 代入式(21)中得到

$$\frac{\frac{\tilde{\alpha}_v^{(\mathrm{l})}(l)}{\alpha_y}(1-\alpha_y)\psi^{(\mathrm{l})}(l)-(1-\tilde{\alpha}_v^{(\mathrm{l})}(l))\psi'}{\tilde{\alpha}_v^{(\mathrm{l})}(l)+(1-\tilde{\alpha}_v^{(\mathrm{l})}(l))\psi^{(\mathrm{l})}(l)}\approx 0 \tag{22}$$

所以有 $\tilde{\psi}^{(2)}(l)\approx\alpha_y\tilde{\psi}(l-1)$。

参考文献

[1] 王文益，伊雪. 基于改进语音存在概率的自适应噪声跟踪算法[J]. 信号处理，2020，36(1)：32-41.
WANG Wenyi，YI Xue. An adaptive noise tracking algorithm using improved speech presence probability[J]. Journal of Signal Processing，2020，36(1)：32-41.（in Chinese）

[2] 成帅，张海剑，孙洪. 结合时变滤波和时频掩码的语音增强方法[J]. 信号处理，2019，35(4)：601-608.
CHENG Shuai，ZHANG Haijian，SUN Hong. Joint time-varying filtering and masking for speech enhancement[J]. Journal of Signal Processing，2019，35(4)：601-608.（in Chinese）

[3] ZHANG Xuefeng，JIA Ying. A soft decision based noise cross power spectral density estimation for two-microphone speech enhancement systems[C]//IEEE International Conference on Acoustics，Speech，and Signal Processing. Philadelphia，PA. IEEE，2005：I/813-I/816.

[4] RAHMANI M，AKBARI A，AYAD B，et al. A modified coherence based method for dual microphone speech enhancement[C]//IEEE International Conference on Signal Processing and Communications. Dubai，United Arab Emirates. IEEE，2008：225-228.

[5] FREUDENBERGER J，STENZEL S，VENDITTI B. A noise PSD and cross-PSD estimation for two-microphone speech enhancement systems[C]//IEEE/SP 15th Workshop on Statistical Signal Processing. Cardiff，UK. IEEE，2009：709-712.

[6] KALLEL F，GHORBEL M，FRIKHA M，et al. A noise cross PSD estimator based on improved minimum statistics method for two-microphone speech enhancement

dedicated to a bilateral cochlear implant[J]. Applied Acoustics, 2012, 73(3): 256-264.
[7] RAHMANI M, AKBARI A, AYAD B, et al. Noise cross PSD estimation using phase information in diffuse noise field[J]. Signal Processing, 2009, 89(5): 703-709.
[8] YOUSEFIAN N, LOIZOU P C. A dual-microphone speech enhancement algorithm based on the coherence function [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2011, 20(2): 599-609.
[9] ITO N, VINCENT E, NAKATANI T, et al. Blind suppression of nonstationary diffuse acoustic noise based on spatial covariance matrix decomposition[J]. Journal of Signal Processing Systems, 2015, 79(2): 145-157.
[10] HENDRIKS R C, GERKMANN T. Noise correlation matrix estimation for multi-microphone speech enhancement[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2012, 20(1): 223-233.
[11] SOUDEN M, CHEN Jingdong, BENESTY J, et al. Gaussian model-based multichannel speech presence probability[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2010, 18(5): 1072-1077.
[12] 鲍长春, 项扬. 基于深度神经网络的单通道语音增强方法回顾[J]. 信号处理, 2019, 35(12): 1931-1941.
BAO Changchun, XIANG Yang. Review of monaural speech enhancement based on deep neural networks[J]. Journal of Signal Processing, 2019, 35(12): 1931-1941. (in Chinese)
[13] XU Yong, DU Jun, DAI Lirong, et al. A regression approach to speech enhancement based on deep neural networks[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2015, 23(1): 7-19.
[14] TAN Ke, WANG Deliang. A convolutional recurrent neural network for real-time speech enhancement[C]// Interspeech. ISCA: ISCA, 2018: 3229-3233.
[15] ZHAO Shengkui, MA Bin. D2Former: A fully complex dual-path dual-decoder conformer network using joint complex masking and complex spectral mapping for monaural speech enhancement[C]//IEEE International Conference on Acoustics, Speech and Signal Processing. Rhodes Island, Greece. IEEE, 2023: 1-5.
[16] 王骞, 何培宇, 徐自励. 利用奇异谱分析的深度神经网络语音增强方法[J]. 信号处理, 2020, 36(6): 902-910.
WANG Qian, HE Peiyu, XU Zili. Deep neural network speech enhancement method based on singular spectrum analysis[J]. Journal of Signal Processing, 2020, 36(6): 902-910. (in Chinese)
[17] 时文华, 张雄伟, 邹霞, 等. 利用深度全卷积编解码网络的单通道语音增强[J]. 信号处理, 2019, 35(4): 631-640.
SHI Wenhua, ZHANG Xiongwei, ZOU Xia, et al. Single channel speech enhancement based on deep fully convolutional encoder-decoder neural network[J]. Journal of Signal Processing, 2019, 35(4): 631-640. (in Chinese)
[18] TOLOOSHAMS B, GIRI R, SONG A H, et al. Channel-attention dense u-net for multichannel speech enhancement[C]//IEEE International Conference on Acoustics, Speech and Signal Processing. Barcelona, Spain. IEEE, 2020: 836-840.
[19] OLIVIERI M, COMANDUCCI L, PEZZOLI M, et al. Real-time multichannel speech separation and enhancement using a beamspace-domain-based lightweight CNN [C]// IEEE International Conference on Acoustics, Speech and Signal Processing. Rhodes Island, Greece. IEEE, 2023: 1-5.
[20] ZHANG Qiquan, QIAN Xinyuan, NI Zhaoheng, et al. A time-frequency attention module for neural speech enhancement[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023, 31: 462-475.
[21] HU Yanxin, LIU Yun, LV Shubo, et al. DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement[C]//Interspeech. ICSA: ICSA, 2020: 2472-2476.
[22] SOUDEN M, CHEN Jingdong, BENESTY J, et al. An integrated solution for online multichannel noise tracking and reduction[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2011, 19(7): 2159-2169.
[23] SOUDEN M, BENESTY J, AFFES S. On optimal frequency-domain multichannel linear filtering for noise reduction[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2010, 18(2): 260-276.
[24] WANG Deliang, CHEN Jitong. Supervised speech separation based on deep learning: An overview[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 26(10): 1702-1726.
[25] LUO Yi, MESGARANI N. Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 27(8): 1256-1266.
[26] BARKER J, MARXER R, VINCENT E, et al. The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines[C]//IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). Scottsdale, AZ, USA. IEEE, 2016: 504-511.
[27] RIX A W, BEERENDS J G, HOLLIER M P, et al. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs[C]//IEEE International Conference on Acoustics, Speech, and Signal Processing. Salt Lake City, UT, USA. IEEE, 2002: 749-752.

[28] CHEN Jingdong, BENESTY J, HUANG Yiteng, et al. New insights into the noise reduction Wiener filter[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2006, 14(4): 1218-1234.

[29] TAAL C H, HENDRIKS R C, HEUSDENS R, et al. A short-time objective intelligibility measure for time-frequency weighted noisy speech [C]//IEEE International Conference on Acoustics, Speech and Signal Processing. Dallas, TX, USA. IEEE, 2010: 4214-4217.

[30] HU Yi, 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.

作者简介

雷　菁　女，1999 年生，河南信阳人。中国科学院声学研究所博士生，主要研究方向为语音增强、深度学习。
E-mail：leijing20@mails.ucas.ac.cn

王劲夫　男，1997 年生，山西忻州人。中国科学院声学研究所博士生，主要研究方向为阵列信号处理、语音增强。
E-mail：wangjingfu19@mails.ucas.edu.cn

杨飞然　男，1982 年生，山东泰安人。中国科学院声学研究所研究员，博士生导师，主要研究方向为音频信号处理。
E-mail：feirany.ioa@gmail.com

杨　军　男，1968 年生，安徽安庆人。中国科学院声学研究所研究员，博士生导师，主要研究方向为通信声学、3D 音频系统、音频信号处理、声场控制和非线性声学。
E-mail：jyang@mail.ioa.ac.cn

（责任编辑：边熙淳）