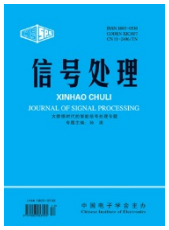


信号处理
Journal of Signal Processing
ISSN 1003-0530,CN 11-2406/TN

《信号处理》网络首发论文

题目：基于多目标联合优化的语音增强方法研究
作者：谢福仕，康迁勇，施明月，郑能恒
网络首发日期：2021-05-31
引用格式：谢福仕，康迁勇，施明月，郑能恒．基于多目标联合优化的语音增强方法研究．信号处理.
https://kns.cnki.net/kcms/detail/11.2406.TN.20210531.1009.016.html

网络首发时间：2021-05-31 11:37:00
网络首发地址：https://kns.cnki.net/kcms/detail/11.2406.TN.20210531.1009.016.html

基于多目标联合优化的语音增强方法研究*

谢福仕 康迂勇 施明月 郑能恒
（深圳大学电子与信息工程学院，深圳 518060）

摘 要：语音增强旨在从受噪声干扰的语音中提取目标语音，目前基于神经网络的语音增强方法在提升语音质量和可懂度方面已被证明是有效的。通过多目标联合优化，利用不同特征之间的互补性，可以提升基于神经网络的语音增强方法的性能。然而，这类多目标学习的语音增强方法在网络优化过程中，通常分别对单个输出目标进行损失函数的计算，多目标之间是并行的，并没有充分利用多目标之间可能存在的关联。为了在网络训练过程中增加输出目标间的关联，本文利用长短时记忆网络构建一种双输出系统框架，设计一种多目标损失函数计算策略用于网络训练。该框架估计出目标语音和噪声，基于此得到估计的带噪语音，然后对这三部分进行联合优化。实验结果表明，所提方法可以提高网络对噪声抑制能力，通过该策略可以获得质量更高，噪声残留更少的增强语音。

关键词：长短时记忆网络；多目标；损失函数；语音增强
中图分类号： TN912.35 **文献标识码：**A

Speech Enhancement Method Based on Multi-Objective Joint Optimization、

Xie Fushi Kang Yuyong Shi Mingyue Zheng Nengheng
College of Electronics and Information Engineering, Shenzhen University, Shenzhen 518060, China

Abstract: Speech enhancement aims to extract target speech from noisy speech. Recently, neural networks (NN) have been effectively implemented for speech enhancement. In particular, network training with multi-objective joint optimization techniques, aiming to take advantage of the complementarity between different features, can significantly improve the quality and intelligibility of the target speech. However, in the network optimization of the multi-objective learning speech enhancement method, the loss function is usually calculated for a single output target separately, and the multiple targets are parallel, but the possible associations between the multiple targets are not fully utilized. This paper presents a speech enhancement framework using long-short term memory networks (LSTMs) with a dual-target output architecture. A multi-objective loss function is proposed for network training such that a balance between the global and local optima can be achieved. The framework estimates the target speech and noise to get the estimated noisy speech, and then optimizes the three parts jointly. Experimental results demonstrate the proposed method can effectively improve the noise suppression ability of the NNs. Through this strategy, enhanced speech with shigher quality and less noise residue can be obtained.

Keywords: long-short term memory network; multiple targets; loss function; speech enhancement

基金项目：广东省重点领域研发计划（2018B030338001）国家自然科学基金面上项目（61771320）

1 引言

语音增强的目的是从受噪声干扰的语音中提取目标语音，提高语音质量和可懂度，是语音识别[1]、移动语音通信和人工耳蜗[2]等应用中一个重要组成部分，在学术界和工业界受到广泛研究。在现实声学环境中噪声种类繁多，除目标语音外，其他任何声音包括人声都可以成为干扰噪声。因此，想要完全去除干扰噪声是非常困难的。

目前语音增强算法主要分为传统算法和基于深度学习的算法。传统方法通常不依赖语音的先验信息，需要的计算量较小，在实际应用中容易实现，但需要对噪声模型进行假设，常常会导致较大偏差，而且无法适应随时间变化的噪声，尤其在非平稳的噪声环境中。典型的传统语音增强算法有谱减法[3]、维纳滤波[4]和子空间法[5]等。近年来，随着计算机运算能力的提高和各类数据的爆炸性增长，基于神经网络的监督学习算法在语音增强任务中取得重大突破。与传统语音增强算法相比，基于神经网络的语音增强算法不需要对噪声模型进行假设，其算法核心是利用大量的训练数据对神经网络进行训练，使模型具有从带噪语音特征估计出目标语音特征的能力，在平稳和非平稳的噪声中能获得更好的语音质量和可懂度，成为当下语音增强任务的重点研究方向[6]。

基于神经网络的语音增强策略大致可以分为两类：Mapping 和 Masking。Mapping 策略是直接利用目标语音谱作为网络训练目标，训练好的网络将带噪语音映射到目标语音[7]；Masking 策略类似于传统的基于掩膜的维纳滤波器，不同之处是将掩膜系数如理想二值掩膜（Ideal Binary Mask, IBM）、理想比率掩膜（Ideal Ration Masking, IRM）和幅度谱掩膜（Spectral Magnitude Mask, SMM）等作为网络输出目标[8]，训练好的网络能够直接由输入带噪语音特征获得掩膜系数。为了进一步提升增强算法的性能，目前的工作主要在网络模型、输入特征和训练目标三个方向进行优化，在训练目标的研究中，多目标学习是被研究的方向之一。

2015 年 Xu 等人[9]采用深度神经网络（Deep Neural Network, DNN）同时输出目标语音的对数功率谱（Log-Power Spectral, LPS）、梅尔频率倒谱系数（Mel Frequency Cepstrum Coefficient, MFCC）和 IBM，并对三者进行联合优化，这种多特征联合优化相比于只优化 LPS 增加了更多的约束，从而提高网络对首要目标特征 LPS 的学习能力。2017 年 Sun 等人[10]发现直接估计 IRM 在高信噪比可获得较好的结果，相反地，直接估计目标语音的 LPS 在低信噪比下获得了更好的结果。为了充分利用这种互补性，其采用长短时记忆（Long-Short Term Memory, LSTM）网络同时对 IRM 和目标语音的 LPS 进行预测。该方法保留了两种输出目标各自的优势，进一步提升了语音增强效果。带噪语音是由目标语音和噪声两部分组成，前述方法虽然采用了多种特征联合优化网络，但所采用的输出特征仍然是目标语音相关的，并没有对噪声信息进行充分利用。2017 年 Wang 等人[11]提出一种联合噪声和掩膜的感知训练策略。其所提系统使用两个 DNN，第一个 DNN 利用前导无声段的噪声和带噪语音的 LPS 作为网络的输入特征，估计出动态噪声信号[12]的 LPS 和目标语音的 IRM，并将这两个特征连同带噪语音的 LPS 作为第二个 DNN 的输入，估计出目标语音 LPS。实验结果表明，动态噪声估计与 IRM 信息有很强的互补性，可以让网络更好的预测目标语音的 LPS。上述多目标学习的语音增强方法在网络优化过程中，通常分别对单个输出目标进行损失函数的计算，多目标之间是并行的，并没有充分利用多目标之间可能存在的关联。

针对上述问题，本文进一步提出基于语音、噪声、带噪语音三者联合优化的策略，设计了一个双输出（语音和噪声）系统框架，并提出一种新的多目标（语音、噪声和带噪语音）损失函数计算策略。该策略利用网络估计的语音和噪声计算出带噪语音，充分考虑了目标语音、噪声、带噪语音以及三者间的关联信息，在训练时对网络输出目标增加了更多的约束。本文构建了一个基于 LSTM 网络的语音增强系统，利用均方误差（Mean-Square error, MSE）和尺度不变信号失真率[13]（Scale-Invariant Signal-to-Distortion Ration, SISDR）作为损失函数对网络参数进行优化，验证所提方法的有效性。实验结果表明，所提方法可有效提高网络对噪声的抑制能力，获得质量更高，噪声残留更少的增强语音。

2 算法描述

给定信号模型：

$$y(n) = x(n) + d(n) \tag{1}$$

其中$x(n)$、$d(n)$和$y(n)$分别为目标语音、噪声和带噪语音信号。语音增强的目的就是从带噪语音$y(n)$中尽可能的提取出目标语音$x(n)$，从而降低噪声$d(n)$对语音质量和可懂度的影响。假设噪声和语音之间是不相关的，对公式（1）中各信号进行短时离散傅里叶变换（Short-time Fourier Transform, STFT）可得：

$$Y(t,f) = X(t,f) + D(t,f) \tag{2}$$

式中$X(t,f)$、$D(t,f)$和$Y(t,f)$分别表示各自时域信号的谱。t、f分别为时间帧与频率通道的索引。同样，增强的目标是从$Y(t,f)$中尽可能估计出$X(t,f)$。

2.1 基于 LSTM 网络的语音增强系统

LSTM 网络是为了缓解传统循环神经网络（Recurrent Neural Network，RNN）的梯度消失和爆炸而提出的，通过输入门、遗忘门和输出门对信息流进行动态的更新、丢弃和输出，可以有效利用信号自身的时序相关信息[14]。由于语音信号具有很强的时间相关性，LSTM 网络可以很好地捕捉语音和噪声的统计特性，在语音增强中有较好的表现[15, 16, 17, 18]。本文将利用 LSTM 网络在 Masking 策略下进行语音增强。为了在网络训练时得到最佳的 Mask，可采用两种损失函数计算方式：其一是直接利用网络估计 Mask 系数如 IBM、IRM 和 SMM 等（与训练数据已知的相应系数值对比）计算损失函数；其二是将网络输出的 Mask 系数与带噪语音幅度谱相乘得到增强的幅度谱，再与目标语音幅度谱计算损失函数（称为信号逼近法[19] Signal Approximation, SA）。方法二的优势在于不用考虑使用何种掩膜对网络学习是最优的，因为其直接以目标语音的幅度谱为代价函数，本文将利用该方法进行实验。图 1 所示为本文所采用的语音增强系统框图。带噪语音的 LPS 作为特征输入给 LSTM 网络，输出的 Mask 与带噪语音幅度谱相乘得到增强语音幅度谱。在训练过程中，增强语音幅度谱与目标语音幅度谱计算损失函数，并更新网络参数，使网络具有估计目标语音的能力。在测试阶段，利用训练好的模型从带噪语音估计得到的增强语音幅度谱，再利用带噪语音的相位，进行时域波形的重构，从而实现语音增强。

2.2 基于多目标损失函数的网络训练

语音增强系统的目标是输出目标语音，在基于神经网络的语音增强系统中，通常只会输出目标语音的谱特征或者掩膜系数。在文献[10]中，则将目标语音的谱特征和掩膜系数作为网络输出。受其启发，为了能在网络优化过程中考虑到噪声信息，本文设计了一个双输出系统框架，如图 2 所示。LSTM 网络输出分为两个掩膜系数 Mask1 和 Mask2，分别对目标语音和噪声的掩膜进行估计。将 Mask1 和 Mask2 分别与带噪语音幅度谱 $|\hat{Y}(t,f)|$ 相乘获得估计的目标语音幅度谱 $|\hat{X}(t,f)|$和噪声幅度谱$|\hat{D}(t,f)|$。将二者分别与目标语音和原噪声的幅度谱计算网络估计的损失函数，分别记为 Loss1、Loss2。在此基础上，本文

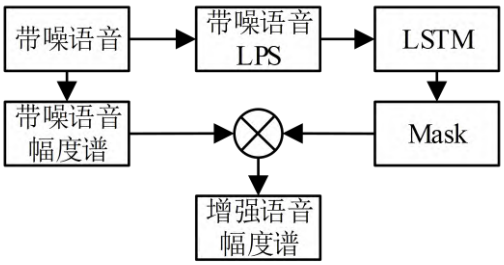


图 1　基于神经网络的语音增强系统框图

Fig.1 Structure of speech enhancement system based on neural network

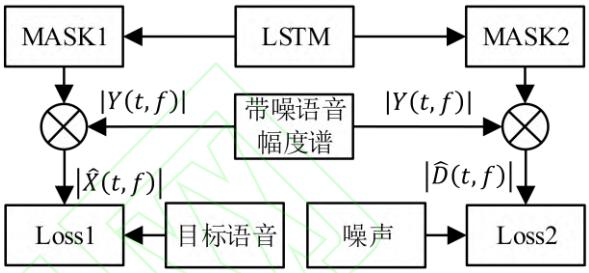


图 2　双输出系统框架

Fig.2 Dual-target output structure

引入第三个损失函数：根据幅度谱减法[3]，计算 $|\hat{Y}(t,f)| = |\hat{X}(t,f)| + |\hat{D}(t,f)|$ 与原带噪语音幅度谱$|Y(t,f)|$的损失函数 Loss3。最后得到联合优化损失函数：

$$\mathcal{L}_{joint} = Loss1 + Loss2 + \alpha Loss3 \quad (3)$$

其中α值依赖于损失函数的选取，当损失函数为 MSE 时，在训练过程中发现取值在 0.5~4 可以提升增强效果，取值在 2 左右时效果最好。当损失函数为 SISDR 时，α值小于 0.02 时可以提升增强效果，并在 0.01 左右时效果最好。引入 Loss3 的逻辑在于：虽然由 Loss1 和 Loss2 进行联合优化，并通过网络权值共享方式使语音和噪声在网络优化中相互影响，但实验结果发现，该方法并不能有效提升网络的增强性能。通过增加 Loss3 来优化网络，迫使估计得到的带噪语音$|\hat{Y}(t,f)|$与原始带噪语音$|Y(t,f)|$的损失函数最小。此时，由于估计带噪语音是由估计的语音和噪声构成的，优化$|\hat{Y}(t,f)|$时同时也对估计的语音和噪声进行了优化，通过这种方式，将语音、噪声和带噪语音三者之间紧密联合，从而使网络能够更好的分离出噪声和语音，提升语音增强效果。

3 实验设置

3.1 实验数据

使用 THCHS-30 数据集[20]对 2 中所提算法进行性能验证。THCHS-30 数据集包含训练、验证和

测试三个子集，分别包含 10000、893 和 2495 句语音。本实验中，训练集和验证集的数据用于训练网络，测试集中随机选取 50 句用于测试。训练噪声数据使用 NOISEX-92 数据集[21]，该数据集共包含 15 种不同的环境噪声。合成的带噪语音共使用 5 种不同的信噪比，分别是-5dB、0dB、5dB、10dB 和 20dB。训练集和测试集中每句语音只混合一种噪声和一个信噪比，故训练集和测试集的带噪语音数量分别为 10000 句和 893 句。用于测试的噪声包含训练集中的 15 种噪声作为可见噪声，还包含来自 MS-SNSD[22]中的两种噪声 Station_1 和 Cafe_1 作为不可见噪声。用于测试的每一句语音分别与所有 17 种噪声在上述 5 种 SNR 下构成带噪语音，故共有 50*17*5=4250 句带噪测试语音。

3.2 特征提取和网络设置

所有数据重采样率均为 16kHz。对语音进行 512 点短时傅里叶变换（使用汉宁窗，帧长 32ms，帧移 16ms），于是每一帧语音可得到 257 个点的幅度谱，计算其 LPS 作为网络的输入特征。网络结构设置如图 3 所示，257 维 LPS 为输入网络的特征，中间两个 LSTM 层和线性层 1 的输出都是 512 维，上一层的输出作为下一层的输入。当预测单目标时，线性层 2 输出维度为 257，对应 Mask1；当预测为多目标时，线性层 2 输出维度为 514，对应 Mask1 和 Mask2。训练网络时，初始学习率设置为 0.01，当上一个 epoch 验证集的 Loss 小于当前 epoch 验证集的 Loss，学习率则以 0.8 的指数衰减率进行降低。批大小（batch size）不是固定的，训练时进行逐句训练，批大小根据每一句语音的长度而确定，并使用 Adam 优化器对网络进行优化，每个模型训练 60 个 epoch。

3.3 基线系统和所提算法

根据所采取的优化目标和损失函数不同，本实验对比六种不同的增强方法。方法一：采用 MSE（记为：$\mathcal{L}_{MSE}$）的单目标优化训练网络，网络输出只有语音，并只计算目标语音对应的损失函数，简写为 ST_MSE（Single-Target MSE）。方法二：采用 MSE 的双目标联合优化训练网络，网络输出有语音和噪声，并分别对目标语音和噪声计算相应的 MSE，计算二者之和作为损失函数进行联合优化，简写为 DT_MSE（Dual-Target MSE）。方法三：采用 MSE 的多目标联合优化训练网络，网络输出有语音和噪声，并分别计算语音、噪声、带噪语音的 MSE，按公式（3）所给损失函数联合优化网络，其中α设置为 2，简写 TT_MSE（Tri-Target MSE）。方法四、五、六与前三种方法相对应，但损失函数从 MSE 改为尺度不变信号失真率 SISDR（记为：$\mathcal{L}_{SISDR}$），SISDR 近年来被用来作为端到端语音增强系统的损失函数[16, 23]并获得较好的增强结果。在计算 SISDR 时，将估计的幅度谱（采用带噪语音的相位）先转换为时域波形再进行计算。方法六中的α设置为 0.01。方法四、五、六分别简写为 ST_SISDR、DT_SISDR 和 TT_SISDR。六种方法中除了上述的区别外，其余的设置都是相同的。在测试时，网络输出结果只考虑目标语音部分。

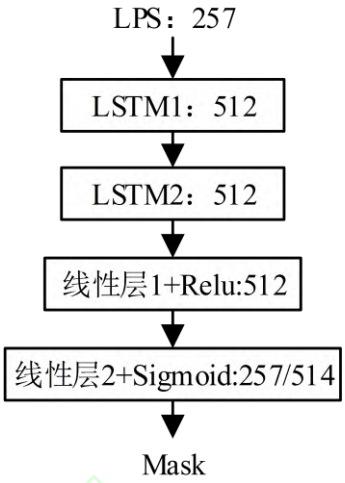


图 3 系统网络结构

Fig.3 System network structure

4 结果分析

分别使用语音质量感知评估[24]（Perceptual Evaluation of Speech Quality, PESQ）和短时客观可懂度[25](Short-Time Objective Intelligibly, STOI)对语音质量和可懂度进行评估。PESQ 通过听觉变换计算响度谱，将增强语音与参考语音的响度谱进行比较，得到-0.5 到 4.5 的语音质量分数值。文献[22]中证明，PESQ 与主观听力测试 MOS 分有较大的相关性。STOI 通过计算增强语音与参考语音的短时包络之间的相关性，得到 0 到 1 的分数值，通常被用来评价增强语音的可懂度。这两种客观指标数值越高，说明增强效果越好。

表 1，给出了六种方法在可见噪声和不可见噪声下的 PESQ 得分结果，对每种噪声分别计算五种不同信噪比的得分均值和所有信噪比下的全局平均得分。对比 ST_MSE 和 ST_SISDR，在可见噪声中，全局平均值没有太大差别，在 20dB 的信噪比下 ST_SISDR 有一定的优势，在不可见噪声中，ST_SISDR 在所有信噪比下呈现出轻微的提升，这个结果说明，在时域使用$\mathcal{L}_{SISDR}$相比在幅度谱上

表 1　PSEQ 指标评估结果

Tab.1 PESQ evaluation results

噪声类型	可见噪声						不可见噪声					
信噪比/dB	20	10	5	0	-5	平均值	20	10	5	0	-5	平均值
Noisy	3.094	2.391	2.030	1.653	1.278	2.089	3.383	2.654	2.308	1.928	1.539	2.362
ST_MSE	3.707	3.247	2.966	2.639	2.235	2.959	3.719	3.127	2.848	2.502	**2.126**	2.864
DT_MSE	3.704	3.248	2.971	2.648	2.248	2.964	3.700	3.129	2.846	2.495	2.115	2.857
TT_MSE	**3.789**	**3.340**	**3.043**	**2.698**	**2.277**	**3.029**	**3.758**	**3.198**	**2.907**	**2.524**	2.123	**2.902**
ST_SISDR	3.733	3.245	2.961	2.627	2.220	2.957	3.733	3.155	2.865	2.512	**2.135**	2.880
DT_SISDR	3.845	3.357	3.008	2.589	2.117	2.983	3.787	3.218	2.895	2.444	1.995	2.868
TT_SISDR	**3.892**	**3.417**	**3.092**	**2.723**	**2.278**	**3.080**	**3.805**	**3.252**	**2.947**	**2.547**	2.123	**2.935**

表 2　STOI 指标评估结果

Tab.2 STOI evaluation results

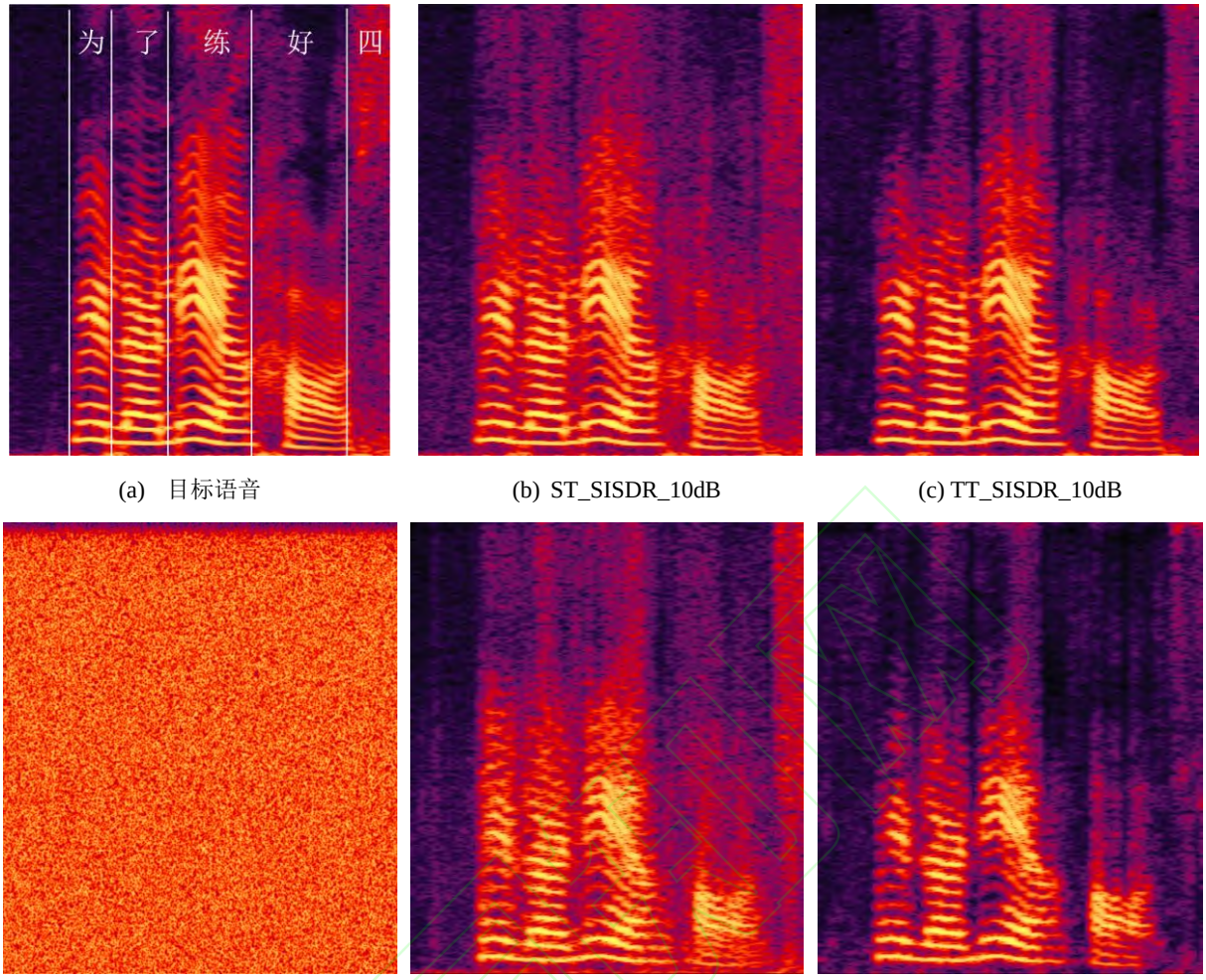
噪声类型	可见噪声						不可见噪声					
信噪比/dB	20	10	5	0	-5	平均值	20	10	5	0	-5	平均值
Noisy	0.880	0.808	0.749	0.675	0.594	0.741	0.891	0.833	0.785	0.714	0.637	0.772
ST_MSE	**0.907**	**0.871**	**0.841**	**0.796**	0.727	0.828	**0.906**	**0.863**	**0.830**	**0.775**	**0.699**	**0.814**
DT_MSE	0.906	**0.871**	**0.841**	**0.796**	**0.728**	**0.829**	0.905	0.861	0.828	0.774	0.695	0.812
TT_MSE	0.906	0.870	0.838	0.791	0.721	0.825	**0.906**	**0.863**	0.829	0.770	0.693	0.812
ST_SISDR	**0.910**	**0.875**	**0.845**	**0.800**	**0.731**	**0.832**	**0.907**	**0.864**	**0.832**	**0.776**	**0.702**	**0.816**
DT_SISDR	0.909	0.871	0.837	0.784	0.702	0.821	0.905	0.861	0.826	0.765	0.680	0.807
TT_SISDR	0.907	0.865	0.829	0.776	0.696	0.815	0.905	0.859	0.823	0.760	0.677	0.805

使用$\mathcal{L}_{MSE}$具有一定的优势。DT_MSE 对比 ST_MSE 整体在可见噪声会有轻微的提升，在不可见噪声则有轻微的下降，所以对于$\mathcal{L}_{MSE}$，通过双目标联合优化的方法并不会对网络估计目标语音的能力产生太大影响。不过在 DT_SISDR 对比 ST_SISDR 中，在可见噪声和不可见噪声中的不同信噪比中的结果都有明显的提升或下降，DT_SISDR 在高信噪比下有明显的提升，但在低信噪比下则会明显的下降。这说明在$\mathcal{L}_{SISDR}$下，双目标联合优化对网络估计目标语音可以产生一定的效果，但会导致信噪比低时性能下降。

TT_MSE 相比于 ST_MSE 和 DT_MSE，除了在-5dB 不可见噪声外，其余条件下 TT_MSE 都有不同程度上的提升，尤其是在高信噪比的情况下，提升较为明显。TT_SISDR 相比于 ST_SISDR 和 DT_SISDR，在可见噪声中,所有信噪比都获得到了明显的提升，在不可见噪声下，除了-5dB 相比于 ST_SISDR 有轻微的下降，其余情况也是有稳定的提升，这里的结果并没有展现出 DT_SISDR 与 ST_SISDR 对比时产生高低信噪比两种方法各占优势的趋势。对比 TT_SISDR 和 TT_MSE，在利用了所提的多目标损失函数计算策略后，可以克服信噪比低时$\mathcal{L}_{SISDR}$不如$\mathcal{L}_{MSE}$的情况，并且在高信噪比的条件下能更大程度的优于$\mathcal{L}_{MSE}$的增强结果。上述结果说明，本文所提出的方法能进一步提升网络的语音增强性能，提高增强语音的语音质量。

表 2 给出 STOI 得分结果。在利用$\mathcal{L}_{MSE}$时，在可见噪声和不可见噪声情况下，三种方法的结果都比较接近，但 TT_MSE 在低信噪比的情况下，则有一定的下降。在使用$\mathcal{L}_{SISDR}$时，DT_SISDR 和 TT_SISDR 在信噪比越低时，性能下降越严重，TT_SISDR 下降程度比 DT_SISDR 更加明显，而 ST_SISDR 在可见和不可见噪声中都得到最好的结果。上述结果说明，所提算法在 STOI 指标上并没有有效的提升。

为了进一步分析得出以上结果的原因，图 4 给出目标语音、噪声、ST_SISDR 和 TT_SISDR 分别对 10dB 和-5dB 带噪语音的增强结果语谱图，混合噪声为白噪声。图中(b)、(c)、(e)、(f)的 PESQ 得分结果分别为 3.212、3.376、2.221、2.288，STOI 得分结果分别为 0.858、0.846、0.743、0.713。图中两个增强方法降噪效果明显，但在更低的信噪比下，增强结果会造成更大的语音失真。对比两个信噪比下 ST_SISDR 和 TT_SISDR 的增强结果，可以观察到 ST_SISDR 的增强结果在高频区域和辅音谐波之间存在更多非目标语音能量，而 TT_SISDR 在清音的位置，如“四”，语音的能量



(a) 目标语音　(b) ST_SISDR_10dB　(c) TT_SISDR_10dB

(d) 白噪声　(e) ST_SISDR_-5dB　(f) TT_SISDR_-5dB

图 4 语谱图

Fig.4 Spectrogram

则被明显的去除，在-5dB 信噪比下，“好”字的声母“h”也被基本去除。在非正式的主观听力测试中，TT_SISDR 的增强结果相比于 ST_SISDR 可以明显感受到更少的噪声残留，并具有更舒适的听音体验，而在低信噪比条件下会感受到更多的语音失真。

综上，本文所提双输出系统框架和多目标损失函数计算策略，能提高网络对噪声的抑制能力，在高信噪比条件下，可以获得残留噪声更少的增强结果，同时不会导致过多的语音失真。但在低信噪比条件下，噪声对语音信号有很强干扰与破坏，导致去除噪声的同时也去除更多的语音成分，尤其是清音部分，造成降噪后语音信号产生更多失真。因此，本文所提方法在要求语音质量较高、噪声残留较少的语音通话和人工耳蜗等应用中具有一定优势，而不适用于语音识别、声纹识别等对语音失真敏感的系统。

5 结论

本文在现有基于神经网络的语音增强算法基础上，提出一种新的双输出系统结构，并构造一个多目标损失函数计算策略，将目标语音、噪声和带噪语音三部分的信息加入到网络优化过程中，充分考虑目标语音、噪声、带噪语音以及三者间的关系，从而对网络输出增加更多的约束。通过实验验证，所提方法能提高网络对噪声的抑制能力，从而获得更好的语音质量。在接下来的工作中，如何提升低信噪比条件下的语音可懂度是主要研究的方向。此外，如何将网络估计的目标语音和噪声进行更合理的结合成带噪语音，使系统具有更好的增强效果，也是值得考虑的方向之一。

参 考 文 献

[1] LI Jinyu, DENG Li, GONG Yifan, et al. An overview of noise-robust automatic speech recognition[J]. IEEE/ACM Transactions on Audio,

Speech, and Language Processing, 2014, 22(4): 745-777.
[2] BOLNER F, GOEHRING T, MONAGHAN J, et al. Speech enhancement based on neural networks applied to cochlear implant coding strategies[C]//2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Shanghai, China. IEEE, 2016: 6520-6524.
[3] BOLL S. Suppression of acoustic noise in speech using spectral subtraction[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1979, 27(2): 113-120.
[4] SCALART P, FILHO J V. Speech enhancement based on a priori signal to noise estimation[C]//1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings. Atlanta, GA, USA. IEEE, 1996: 629-632.
[5] EPHRAIM Y, VAN TREES H L. A signal subspace approach for speech enhancement[J]. 1993 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1993, 2: 355-358.
[6] 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)
[7] XU Yong, DU Jun, DAI Lirong, et al. An experimental study on speech enhancement based on deep neural networks[J]. IEEE Signal Processing Letters, 2014, 21(1): 65-68.
[8] WANG Yuxuan, NARAYANAN A, WANG Deliang. On training targets for supervised speech separation[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014, 22(12): 1849-1858.
[9] XU Yong, DU Jun, HUANG Zhen, et al. Multi-objective learning and mask-based post-processing for deep neural network based speech enhancement[C]//Interspeech2015, 2015: 1508-1512.
[10] SUN Lei, DU Jun, DAI Lirong, et al. Multiple-target deep learning for LSTM-RNN based speech enhancement[C]//2017 Hands-free Speech Communications and Microphone Arrays (HSCMA). San Francisco, CA, USA. IEEE, 2017: 136-140.
[11] WANG Qing, DU Jun, DAI Lirong, et al. Joint noise and mask aware training for DNN-based speech enhancement with SUB-band features[C]//2017 Hands-free Speech Communications and Microphone Arrays (HSCMA). San Francisco, CA, USA. IEEE, 2017: 101-105.
[12] XU Yong, DU Jun, DAI Li rong, et al. Dynamic noise aware training for speech enhancement based on deep neural networks[J]. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2014: 2670-2674.
[13] ROUX J L, WISDOM S, ERDOGAN H, et al. SDR – half-baked or well done?[C]//ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, UK. IEEE, 2019: 626-630.
[14] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.
[15] TU Yanhui, DU Jun, GAO Tian, et al. A multi-target SNR-progressive learning approach to regression based speech enhancement[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 28: 1608-1619.
[16] HU Yanxin, LIU Yun, LV S, et al. DCCRN: deep complex convolution recurrent network for phase-aware speech enhancement[C]//Interspeech 2020. ISCA: ISCA, 2020: 2472-2476.
[17] TANG Xin, DU Jun, CHAI Li, et al. Geometry constrained progressive learning for lstm-based speech enhancement[C]//ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain. IEEE, 2020: 7514-7518.
[18] GAO Tian, DU Jun, DAI Lirong, et al. Densely connected progressive learning for LSTM-based speech enhancement[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, AB, Canada. IEEE, 2018: 5054-5058.
[19] 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.
[20] WANG Dong, ZHANG Xuewei. Thchs-30: A free chinese speech corpus. arXiv preprint arXiv:1512.01882, 2015.
[21] 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.
[22] REDDY C K A, BEYRAMI E, POOL J, et al. A scalable noisy speech dataset and online subjective test framework[C]//Interspeech 2019. ISCA: ISCA, 2019: 1816-1820.
[23] KOLBÆK M, TAN Zhenghua, JENSEN S H, et al. On loss functions for supervised monaural time-domain speech enhancement[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 28: 825-838.
[24] 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]//2001 IEEE International Conference on Acoustics, Speech, and Signal Processing(ICASSP). Salt Lake City, UT, USA. IEEE, 2001: 749-752.
[25] TAAL C H, HENDRIKS R C, HEUSDENS R, et al. An algorithm for intelligibility prediction of time–frequency weighted noisy speech[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2011, 19(7): 2125-2136.

作者简介：

谢福仕 男，1995 年生，广西南宁人。深圳大学硕士研究生，主要研究方向为语音增强。
E-mail：1800261039@email.szu.edu.cn

康迁勇 男，1995 年生，江西赣州人。深圳大学硕士研究生，主要研究方向语音信号处理。
E-mail：1810262077@email.szu.edu.cn

施明月 女，1993 年生，安徽省亳州人。深圳大学博士研究生，主要研究方向为语音增强、人工耳蜗信号处理等。
E-mail：2050432012@email.szu.edu.cn

郑能恒（通讯作者） 男，1974 年生，福建福州人。深圳大学电子与信息工程学院副教授，主要研究方向为语音信号处理、人工耳蜗言语处理策略、人工听觉等。

E-mail：nhzheng@szu.edu.cn