说话人特征约束的多任务卷积网络语音增强

龙 华，张林濮，邵玉斌，杜庆治

（昆明理工大学 信息工程与自动化学院，云南 昆明 650500）

E-mail ：1797662488@qq.com

摘 要：针对噪声干扰环境下的说话人识别问题，提出了一种基于多任务学习的语音增强方法作为说话人识别系统的前端。在卷积神经网络（CNN）的基础上，通过构建语音增强与说话人识别的融合网络多任务学习模型，同时在输入输出端拼接梅尔频谱倒谱系数（MFCC）和基音周期特征作为辅助任务，以及利用同方差不确定性自适应调整损失权重。实验结果表明，相比只输入对数功率谱（LPS）的 CNN 以及 DNN 模型，加入辅助任务的 CNN 模型可以提高语音增强的表现。另外，语音增强与说话人识别任务的联合训练可以增强噪声干扰下的说话人识别效果，提高模型的鲁棒性。

关键词：语音增强；多任务学习；说话人识别；卷积神经网络

中图分类号：TP391.4 **文献标识码：**A **文章编号：**2020-0648

A Multi-task Convolution Network with Speaker Feature Constraint for Speech Enhancement

LONG Hua，ZHANG Lin-pu， SHAO Yu-bin，DU Qing-zhi

（Faculty of Information Engineering and Automation，Kunming University of Science and Technology, Kunming 650500, China）

Abstract: Aiming at the problem of speaker recognition in noisy environment, a speech enhancement method based on multi task learning is proposed as the front end of speaker recognition system. Based on convolutional neural network (CNN), a fusion multi-task learning model of speech enhancement and speaker recognition is built. At the same time, Mel frequency cepstrum coefficient (MFCC) and pitch features are concatenated with the input and output as auxiliary tasks, and the loss weight is adaptively adjusted by same variance uncertainty. The experimental results show that the CNN model with auxiliary tasks can improve the performances of speech enhancement compared with the CNN and DNN models which only use logarithmic power spectrum (LPS) as input. In addition, the joint training of speech enhancement and speaker recognition tasks can improve the performances of speaker recognition under noise interference and the robustness of the model.

Key words: speech enhancement; multi-task learning; speaker recognition; convolution network

1 引言

语音增强（Speech Enhancement，SE）是声学研究中的一项重要任务和课题，它的目的是在给出带噪语音的条件下，尽可能从中恢复不带噪的干净语音。语音增强的方法种类繁多，传统的单声道语音增强算法主要分为时域和频域方法。时域方法如基于参数和滤波的方法，主要利用滤波器估计发声器官的声道参数和激励参数[1]。频域方法主要基于短时谱估计，如：谱减法[2]，维纳滤波法[3]，最小均方误差法[4]等。

近年来，深度学习成为语音领域的研究热点。深度神经网络（DNN）被广泛应用于语音识别，语音合成，说话人识别等领域。因为深度学习在图像识别领域的巨大成功，研究人员开始在语音增强方向应用类似的方法并取得成功[5]。尽管神经网络在语音增强中很大程度提高了降噪音频的质量，但在这些模型中，增强过程在对背景噪声抑制的同时，也对原始信号造成了较大的破坏。

说话人识别（Speaker Recognition，SR）又称声纹识别，同样是语音识别任务中具有很大研究价值的方向。声纹识别已经成为身份认证领域的重要手段，声纹这一生物特征以经成功应用于监控、安全解锁、智能手机以及智能电器声控操作以及司法认证等领域。目前，神经网络包括 d-vector[6]，x-vector[7]等模型已经在说话人识别领域取得了成功。这些模型在理想无干扰或者高信噪比条件下均取得了比较满意的效果。但是，说话人识别任务在干扰环境特别是低信噪比干扰条件下的识别率迅速下降。虽然 x-vector 在数据集上做了加噪与加混响处理[8]，但是这种方法更多是出于数据增强方面的考虑的，而且加入的噪声能量相对较小，面对低信噪比环境效果依然不理想。所以，面对真实情况下复杂的噪声类型以及信噪比环境，对语音进行预处理以及增强，提高说话人识别系统在噪声环境下的鲁棒性是很有必要的。文献[9]

收稿日期：2020-10-1 基金项目：国家自然科学基金（No.61761025）资助 作者简介：**龙华，**女，1963 年生，博士，教授，CCF 会员，研究生导师，研究方向为无线网络及音频信号处理；**张林濮，**男，1996 年生，硕士研究生，研究方向为音频信号处理，语音识别；**邵玉斌，**男，1970 年生，硕士，教授，研究生导师，研究方向为移动通信和个人通信系统以及信号处理；**杜庆治，**男，1977 年生，硕士，高级工程师，研究生导师，研究方向为信号处理及通信与信息系统.

中利用 DNN 网络拟合干净语音和带噪语音的 i-vector 特征矢量的非线性函数关系，获得干净语音 i-vector 的近似表征，降低了噪声对说话人识别的影响。文献[10]中将带噪语音的梅尔频谱倒谱系数（Mel frequency cepstrum coefficient，MFCC）和理想二值掩蔽（Ideal Binary Mask，IBM）与对数功率谱一起拼接并输入 DNN 网络中，预测对应干净语音的这三项特征。相较于仅估计纯净语音对数功率谱的单任务模型，该框架对目标函数添加了额外的约束，提高了语音增强的效果。文献[11]设计了一种去噪自动编码器，并在其上叠加一个深度网络形成深层结构，用这个神经网络结构替换传统 i-vector 的高斯混合模型。

受上述工作启发，本文提出了一种基于多任务学习的卷积神经网络语音增强模型。在多任务框架中，该模型不只学习带噪语音与干净语音的对数功率谱之间的映射关系，同时将离散的说话人标签作为网络的另一个输出。并且把语音的连续特征（如 MFCC，基音周期）作为多任务学习中的辅助任务，希望能给网络提供更多的信息，并且作为限制项对输出的功率谱做一个约束，从而提高语音增强和说话人识别效果。

2 基于 CNN 的频谱映射模型

卷积神经网络（Convolutional Neural Network，CNN）近年来在图像识别领域取得了巨大的成功，并大规模应用于商业项目[12]。CNN 的原理是基于对生物视觉习惯和神经网络的一种模拟，并对大脑皮层中的局部感知做一种近似。相较于标准的全连接 DNN 模型，CNN 可以更好地适应音频信息中时域和频域维度的变化，并克服语音信号中不稳定环境与非平稳噪声的影响[13]。此外，由于 CNN 的卷积层采用了参数共享和稀疏连接的原理，相比于 DNN，CNN 模型的参数数量也大规模减少，提高了训练和运算速度，使得其能在性能较弱的设备上更好的运行。

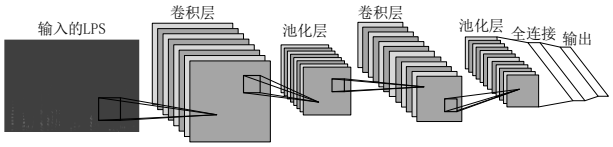

图 1 CNN 模型结构

Fig. 1 Structure of CNN model

假设有一段干净语音 和一段噪声 ，那由它们得到的加性带噪语音为：

$$y = x + n \tag{1}$$

语音增强的目标就是要在输入为带噪语音 y 的条件下，尽可能准确地估算出干净语音 x 的信号。增强模型对干净语音的预测记为 $\hat{x}$。基于频谱映射的语音增强就是根据带噪语音的幅值谱来计算干净语音的幅值谱。幅值谱代表了信号在不同时间和频率的强度变化情况，一般来说，幅值谱的横轴代表时间，纵轴代表频率，Z 轴代表相应信号的振幅。

语音信号 x 和带噪语音 y 经过分帧处理后的短时傅里叶变换（short-time Fourier transform，STFT）记为 $X(n,k)$ 和 $Y(n,k)$，其中，$n=1,2,\ldots\ldots,N$ 代表帧数。$k=1,2,\ldots\ldots,K$ 代表频带维度，傅里叶变换的点数为 D，因为傅里叶变换具有对称性，故幅值谱的有效频带维度 $K=D/2+1$。语音信号的 STFT 的复数序列形如：

$$X(n,k) = X_r(n,k) + jX_i(n,k) \tag{2}$$

X_r 和 X_i 分别为 STFT 域上的实部和虚部。则其幅值 F 和相位 φ 的计算公式为：

$$X_r(n,k) + jX_i(n,k) = \sqrt{X_r(n,k)^2 + X_i(n,k)^2}\, e^{j\theta} \tag{3}$$

$$\varphi_x(n,k) = \arctan\frac{X_i(n,k)}{X_r(n,k)} \tag{4}$$

$$F_x(n,k) = \left|X(n,k)\right| \tag{5}$$

因为人耳对音频信号的感知是非线性的，故一般对幅值谱取对数来增强振幅微弱的部分，则得到对数功率谱(Logarithmic Power Spectra, LPS):

$$P_x(n,k) = \log\left(F_x(n,k)\right) \tag{6}$$

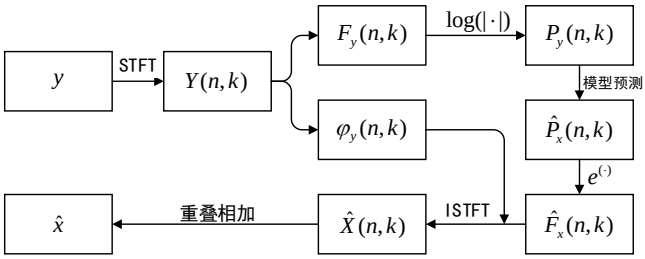

图 2 频谱特征提取及语音重构流程

Fig. 2 Process of spectrum feature extraction and speech reconstruction

$Y(n,k)$ 的 LPS 提取方法与 $X(n,k)$ 相同。为了避免极端数值的影响，提高模型的收敛速度，我们对带噪语音的对数功率谱 $P_y(n,k)$ 和干净语音的对数功率谱 $P_x(n,k)$ 都进行归一化处理。带噪语音的对数功率谱 $P_y(n,k)$ 即为 CNN 增强模型的输入，模型对干净语音的 LPS 的估计为模型的输出，记为 $\hat{P}_x(n,k)$。CNN 语音增强模型的训练目标就是最小化如下的损失函数 L：

$$L = \sum_{n=1}^{N}\sum_{k=1}^{K}\left(\hat{P}_x(n,k) - P_x(n,k)\right)^2 \tag{7}$$

首先需要对估计的 LPS 进行指数操作恢复为功率谱：

$$\hat{F}_x(n,k) = \exp\left(\hat{P}_x(n,k)\right) \tag{8}$$

因为对干净语音 LPS 的相位估计比较困难，而且实验表明人耳对相位变化的感知不明显，所以我们采用带噪语音的相位来对干净语音进行重构：

$$\hat{X}(n,k) = \hat{F}_x(n,k)e^{j\varphi_y(n,k)} \tag{9}$$

$$\hat{x}(n) = \frac{1}{K}\sum_{k=1}^{K}\hat{X}(n,k)e^{j2\pi kn/K} \tag{10}$$

其中，K 为上面提到的傅里叶变换的有效维度。

3 基于多任务学习的 CNN 声学模型

3.1 融合网络多任务学习

多任务学习（multi-task learning，MTL）概念的提出是相对于标准的单任务学习模型的。传统的单任务模型一次只优化一个目标函数，针对单个任务。而多任务学习通过共享一些隐藏层的参数，来获取多个任务之间的关联性信息[14]。

带噪语音中不只混有语音与噪声的信息，也含有不同说话人之间的音调、音色差异等信息。语音增强任务注重提取干净语音之间的相似性，并区分语音与噪声的差异。说话人识别注重区分不同说话人在语音信号中的个人特征，如发音习惯，音色等。因此，利用多任务学习的参数共享机制，通过神经网络模型让语音增强任务和说话人识别任务能够获得彼此之间的隐含信息。两个任务之间的特性与共性也通过该机制体现，并可以对各自的训练提供帮助。

区别于传统的共享所有参数的多任务神经网络模型，本文使用了一种十字绣单元[15]（Cross-Stitch Unit），通过这个单元，模型的多个任务之间可以共享隐藏层参数，并且可以通过反向传播来自适应调整共享程度。如果模型有两个任务 A 和 B，记 h_A 和 h_B 分别为模型中第 l 层经过激活函数的输出，h_A^{ij} 和 h_B^{ij} 分别为输出矩阵中位于特定位置 (i,j) 的输出。将 h_A 和 h_B 输入线性组合函数中，得到下一卷积层的输入 $h_A^{'}$ 和 $h_B^{'}$：

$$\begin{pmatrix} h_A^{'ij} \\ h_B^{'ij} \end{pmatrix} = \begin{pmatrix} \Lambda_{AA} & \Lambda_{AB} \\ \Lambda_{BA} & \Lambda_{BB} \end{pmatrix} \begin{pmatrix} h_A^{ij} \\ h_B^{ij} \end{pmatrix} \tag{11}$$

这些线性函数的参数记为 Λ，网络就可以通过将 Λ_{AB} 和 Λ_{BA} 设置为 0，来结束共享；通过将 Λ_{AB} 和 Λ_{BA} 附上更高的值来提高共享的程度。注意我们只在 CNN 的池化层或全连接层进行参数共享。

Λ 是可以通过训练学习的参数。因为单元内的函数为线性组合，所以损失函数 L 对于 Λ 的偏导可以按如下计算：

$$\begin{pmatrix} \dfrac{\partial L}{h_A^{ij}} \\ \dfrac{\partial L}{h_B^{ij}} \end{pmatrix} = \begin{pmatrix} \Lambda_{AA} & \Lambda_{BA} \\ \Lambda_{AB} & \Lambda_{BB} \end{pmatrix} \begin{pmatrix} \dfrac{\partial L}{h_A^{'ij}} \\ \dfrac{\partial L}{h_B^{'ij}} \end{pmatrix} \tag{12}$$

$$\begin{aligned} \frac{\partial L}{\partial \Lambda_{AA}} &= \frac{\partial L}{\partial h_A^{'ij}} h_A^{ij}, \frac{\partial L}{\partial \Lambda_{AB}} = \frac{\partial L}{\partial h_B^{'ij}} h_A^{ij} \\ \frac{\partial L}{\partial \Lambda_{BA}} &= \frac{\partial L}{\partial h_A^{'ij}} h_B^{ij}, \frac{\partial L}{\partial \Lambda_{BB}} = \frac{\partial L}{\partial h_B^{'ij}} h_B^{ij} \end{aligned} \tag{13}$$

融合网络多任务学习和 CNN 模型结构如图 3 所示，本文设计的 CNN 模型包含多个卷积层和全连接层，其中每个卷积层和全连接层之后需要进行非线性变换操作，我们选择线性整流函数（Rectified Linear Unit，ReLU）作为激活函数。卷积核的大小均为3×3，卷积核数量为64-64-128-128-256，步长均为 1×1。池化层均采用最大池化（Max-Pooling），大小为 2×2。卷积层之后为两个全连接层，节点数量均为 512。最后，语音增强任务的输出层为包含 101 个节点的全连接层，说话人识别任务的输出层也为全连接层，节点数量与训练集说话人数目保持一致，最终通过 softmax 函数输出。

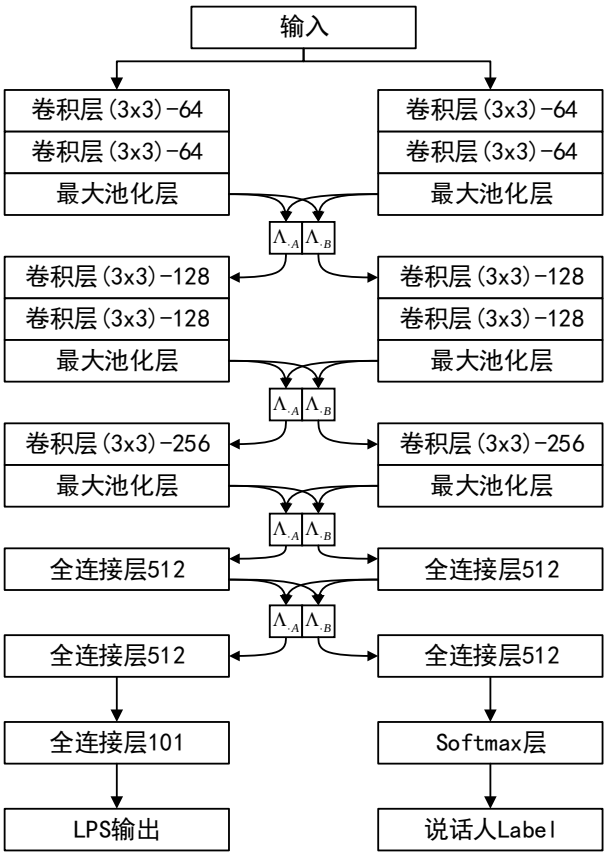

图 3 融合网络多任务学习 CNN 模型结构

Fig.3 Structure of fusion network multi-task learning CNN model

3.2 特征联合辅助训练

在基于 CNN 的语音增强中，优化的目标函数为对数功率谱上输入输出之间的最小均方误差[16, 17]。在对数功率谱域中，不同频率之间是独立假设的，各个维度的相互关系没有被考虑，模型的预测缺乏约束，也不利于对人耳的听觉特征的模拟与感知[18]。

在本节，我们通过引入辅助学习（Auxiliary Learning）的方法来间接的优化目标函数。在多任务学习框架中，如果主任务的数据维度高，不相关特征较多，会对模型的拟合造成更多的困难。辅助任务对模型的训练添加了约束，使得模型能将注意力更加集中于那些与结果紧密相关的特征。所以，辅助任务的引入让模型不仅学习干净语音的 LPS，同时学习如梅尔频谱倒谱系数和基音周期这样的连续特征。

MFCC 是常见的用于语音识别，声纹识别和情感识别等任务的语音特征。MFCC 的计算过程中，在倒谱上应用了跟人耳感知音高变化等距的梅尔滤波器组，凸显了声音中的低频部分，并且强调了相邻帧之间的联系。在 MFCC 中，梅尔三角滤波对语音频谱进行了平滑化，消除了谐波的作用，也使得语音的共振峰得到凸显。因此 MFCC 并不反映一段语音的音调或音高，所以说，如果将 MFCC 作为一个语音辨识系统的输入特征，结果并不会受到输入语音音高的影响。但实际上，音高的变化可以表示出不同说话人发音习惯上的不同，是描述语音激励的一个重要特征。

基音（pitch）反映了人在发浊音时声带振动的周期性，

而基音周期为声带振动频率的倒数。说话人声带的薄厚，韧性，长短等与基音周期有很大的关系，所以基音周期在很大程度上反映了说话人声音的个性。

本文采用小波变换法提取基音周期。由于小波变换对信号中频率和时间分辨率特性与人耳的时频分析特征极为类似，并且语音信号的小波变换极值点对应声门的开启和闭合点。所以基音周期就可以用小波变换中相邻极值点之间的距离估算。信号中的突变位置反映在零点或极值点上。于是，根据小波信号中的奇异点，就可以实现对基音周期的检测。

3.3 自适应损失

多任务学习中，多元回归和分类任务通过从共享的表示中学习多个目标来提升效率，预测精度和泛化能力。但是，不同的任务之间尺度是不同的，这就涉及到多任务学习中不同单位尺度任务的目标函数的联合学习。所以，多任务学习中很重要的一个问题是如何设计损失函数，平衡不同类型的任务，避免在训练过程中整个模型被某一个任务主导。这就涉及到为不同任务的损失函数赋上不同的权重，将不同任务的损失统一成一个损失函数。常规方法是将各任务的损失简单相加或者设置统一的权重，如下式：

$$loss=\frac{1}{N}\sum_{i=0}^{N}L_i \tag{14}$$

更进一步，可能会手动的进行权重调整，这样会造成最终模型在有些任务上表现很好，而在其他任务上效果较差。

文献[19]介绍了一种利用同方差的不确定性（Homoscedastic Uncertainty）自适应调整不同损失函数权重的方法。同方差的不确定性属于偶然不确定性，这种不确定性捕捉了不同任务之间的相关性置信度，所以这种不确定性可以作为不同任务损失权重赋值的衡量标准。

假设 $f_w(x)$ 为神经网络在输入为 x、模型参数为 w 时的输出，y 为对应的正确输出。在多任务情况下，得到 K 个离散回归任务的最大似然：

$$p(y_i \mid f_w(x))=N(f_w(x),\sigma_i^2) \tag{15}$$

$$p(y_1,y_1,\cdots,y_k \mid f_w(x))=\prod_{i=1}^{K}p(y_i \mid f_w(x)) \tag{16}$$

其中，每个模型都遵循带有噪声标量 σ 的高斯分布。对于分类问题，通常会通过 softmax 函数输出，如下式所示：

$$p(y \mid f_w(x))=\mathrm{softmax}(f_w(x)) \tag{17}$$

所以最大似然的估算可以表示为下式中最小化模型的负对数似然：

$$-\log p(y_1,y_1,\cdots,y_k \mid f_w(x)) \propto \sum_{i=1}^{K}\frac{1}{2\sigma_i^2}\| y_i-f_w(x)\|^2+\log\sigma_i \tag{18}$$

求得 softmax 分类的似然估计：

$$\log p(y=c \mid f_w(x),\sigma)=\frac{1}{\sigma^2}f_w^c(x)-\log\sum_{c'}\exp\left(\frac{1}{\sigma^2}f_w^{c'}(x)\right) \tag{19}$$

以两个输出连续性 y_1 和离散型 y_2 为例，分别用高斯分布和 softmax 建模，可得损失函数：

$$\begin{aligned}L(w,\sigma_1,\sigma_2)&=N(f_w(x),\sigma_1^2)\cdot\mathrm{softmax}(f_w(x))\\&=\frac{1}{2\sigma_1^2}L_1(w)+\frac{1}{\sigma_2^2}L_2(w)+\log\sigma_1+\log\frac{\sum_{c'}\exp\left(\frac{1}{\sigma_2^2}f_w^{c'}(x)\right)}{\left(\sum_{c'}\exp(f_w^{c'}(x))\right)^{\frac{1}{\sigma_2^2}}}\\&\approx\frac{1}{2\sigma_1^2}L_1(w)+\frac{1}{\sigma_2^2}L_2(w)+\log\sigma_1+\log\sigma_2\end{aligned} \tag{20}$$

指定每一个任务对应的损失函数 $L_i(w)=\| y_i-f_w(x)\|^2$，则最终多任务模型的联合损失为：

$$L(w,\sigma_{1:K})=\sum_{i=1}^{K}\frac{1}{2\sigma_i^2}L_i(w)+\log\sigma_i \tag{21}$$

当噪声 σ 增大时，相对应的权重就会降低；反过来，随着噪声 σ 减小，相对应的权重就要增加。

4 实验设计与结果分析

4.1 实验设置

本实验采用的语音数据为 Free ST Chinese Mandarin Corpus 中文语音库，均为重采样到 8000Hz 频率、单通道的干净语音。Free ST Chinese Mandarin Corpus 语音库包含 855 个说话人的每人 120 条，总计十余万条的中文普通话语音数据。采用的噪声数据来自 ESC-50 噪声库[20]，包括 50 类、每类 40 条的噪音。

关于带噪语音的合成，将每一条语音库中的干净语音，分别按-5dB，-2dB，0dB，5dB，10dB 与噪音混合。其中，噪音是从 ESC-50 的 50 类噪音中随机挑选一类，并随机截取与干净语音等长的片段，按照信噪比调整噪音能量，然后与原始语音混合。

语音预处理部分，先对原始语音利用谱熵法进行端点检测（Speech Activity Detection，SAD）；之后对信号进行分帧，帧长与 STFT 的点数保持一致，为 200 点（25ms）；帧移为 80 点（10ms）。之后在每一帧信号上加汉宁窗（hamming window）并计算 STFT。MFCC 变换中的 STFT 与上述保持一致，梅尔滤波器的个数为 40。

测试集中的干净语音来自 Free ST Chinese Mandarin Corpus 中与训练集不重叠的 127 个说话人的 254 条语音；噪音来自 ESC-50 中的 3 种与训练集不重叠的 6 段噪音，分别是 Laughing、Wind 和 Train 类噪音。测试集的干净语音按照-5dB，0dB 和 10dB 的信噪比分别与不同种类中随机截取的噪声混合，总计得到大小为 4572 条的带噪测试集。

本实验采用感知语音质量（Perceptual Evaluation of Speech Quality，PESQ）、短时客观可懂度（Short Time Objective Intelligibility，STOI）和分段信噪比（Segmental

SNR, SSNR）作为语音增强结果的评价指标。其中，PESQ 偏重于增强语音的总体质量，是评价语音质量常用的标准方法，得分介于-0.5 到 4.5 之间，越高代表语音质量越好；STOI 是近几年语音增强领域常用的评价语音可懂度的指标，得分介于 0 到 1 之间，越高代表可懂度越好；SSNR 代表了增强之后语音的信噪比，越高代表增强语音的干净程度越高。

4.2　实验测试

为了验证本文提出模型的效果，本实验设计了基于 DNN 的语音增强模型的基线，模型采用 LPS 作为输入输出特征，其中每个输入为包含上下文共 11 帧的 LPS，输出为对应的干净频谱的中间帧。模型共有 4 个隐藏层，每个隐藏层包含 2048 个节点，激活函数为 ReLU；只包含语音增强和说话人识别两个主任务的 CNN 模型记为 CNN-Mul；本文提出的包含两个主任务和辅助任务的 CNN 模型记为 CNN-Mul-Aux。模型 CNN-Mul-Aux 的辅助训练中，我们将带噪语音的 MFCC 和基音周期与输入端的 LPS 拼接在一起，并将干净语音的 MFCC 和基音周期拼接在语音增强的输出端。

表 1 Laughing 噪声下的语音增强性能测试

Table 1 Performance test of speech enhancement under Laughing noise

model	输入 SNR	PESQ	STOI	SSNR
DNN	-5dB	1.2368	0.4983	-2.4697
	0dB	1.6196	0.5968	-0.4677
	10dB	2.1627	0.7276	0.8542
CNN-Mul	-5dB	1.7425	0.6247	-1.3697
	0dB	2.0125	0.7607	-0.0356
	10dB	2.3647	0.8044	1.2456
CNN-Mul-Aux	-5dB	1.9465	0.6623	-1.8697
	0dB	2.2354	0.7709	0.2155
	10dB	2.4853	0.8153	1.5687

表 2 Wind 噪声下的语音增强性能测试

Table 2 Performance test of speech enhancement under Wind noise

model	输入 SNR	PESQ	STOI	SSNR
DNN	-5dB	1.8075	0.6148	-1.2716
	0dB	2.2734	0.7963	-0.2536
	10dB	2.5180	0.8461	2.3215
CNN-Mul	-5dB	1.9863	0.7035	-1.0456
	0dB	2.3058	0.7812	0.6820
	10dB	2.5266	0.8240	3.0145
CNN-Mul-Aux	-5dB	2.1969	0.7791	-0.0184
	0dB	2.3561	0.8041	0.7713
	10dB	2.6710	0.8575	3.4791

表 3 Train 噪声下的语音增强性能测试

Table 3 Performance test of speech enhancement under Train noise

model	输入 SNR	PESQ	STOI	SSNR
DNN	-5dB	1.7669	0.6572	-1.1330
	0dB	2.0001	0.7356	-0.0849
	10dB	2.3630	0.8060	3.2812
CNN-Mul	-5dB	1.8849	0.5866	-0.9616
	0dB	2.2314	0.7379	0.4135
	10dB	2.4812	0.8316	2.9145
CNN-Mul-Aux	-5dB	1.9624	0.7037	-0.2131
	0dB	2.3633	0.8258	1.1040
	10dB	2.6073	0.8464	3.7687

表 1、表 2 和表 3 分别在 Laughing、Wind 和 Train 类噪音下对模型性能进行了对比。从表 1~表 3 可以看出，CNN 增强模型在三种噪声下均实现了不错的效果，除了 10dB 的 Train 噪声下的 SSNR 外，CNN 在不同 SNR 级别以及环境下的各指标要一致的比 DNN 都要好。并且 CNN 在噪声的滤除上也比 DNN 实现了更好的效果，说明 CNN 确实更好的利用了 LPS 频域和时域的相关性，使其可以更好地估计语音中的平稳与非平稳噪声。另外，加入辅助任务的 CNN-Mul-Aux 模型在各项指标上均好于未加入辅助任务的 CNN-Mul 模型和 DNN，说明将 MFCC 特征和基音周期拼接在输入输出端确实可以显著提升模型的增强效果，并且可以避免模型在单独的 LPS 任务上过拟合，从而增强模型的鲁棒性。

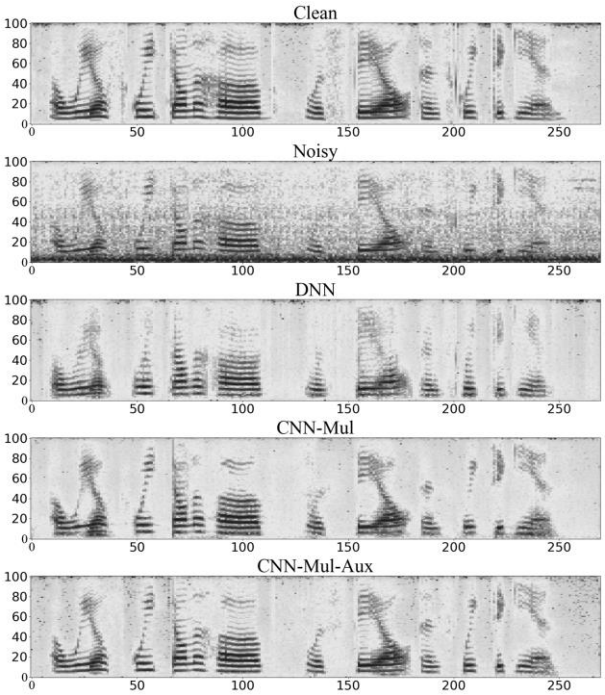

图 4 对数功率谱结果对比

Fig.4 Comparison of logarithmic power spectrum results

如图 4 所示分别为各模型对带噪语音进行增强之后的

结果的对数功率谱图对比，可以更直观的体现增强效果的差异。我们取一段加入-5dB 的 Train 噪声的语音并用各个模型进行增强，图中从上到下依次为干净语音、带噪语音以及模型增强之后的对数功率谱图。其中，横轴代表分帧之后的帧数，纵轴代表频率，Z 轴代表信号在相应频率的振幅。

从图 4 中几种算法的对数功率谱结果对比可以看出，采用 CNN 算法增强后的语音纯净度比 DNN 更高，对噪声的抑制效果更好。另外，CNN-Mul-Aux 相较于未加入辅助任务的 CNN-Mul 模型，更好的保留了原始语音的信息。

本文通过 x-vector 说话人识别算法来评估经过增强后的语音在说话人信息上的失真程度。搭建的模型框架基于 Kaldi 工具箱，训练集采用 AISHELL-1 中文语音数据集[21]。x-vector 与 CNN 说话人识别模型在测试阶段均取 softmax 之前的向量并进行概率线性判别分析（Probabilistic Linear Discriminant Analysis，PLDA）打分。注册集为 Free ST Chinese Mandarin Corpus 中与训练集不重叠的 127 个说话人的 1270 条语音，测试集来自相同说话人的另外 2540 条语音。采用的性能评价指标为等错误率（equal error rate，EER）和最小检测代价准则(minimum detection cost function，DCF)。DCF 函数计算公式如下：

$$DCF = C_{FR}E_{FR}P_{target} + C_{FA}C_{FA}(1-P_{target}) \tag{22}$$

C_{FR} 与 C_{FA} 分别为错误拒绝和错误接受的惩罚系数，P_{taraet} 与 $(1-P_{taraet})$ 分别为真实说话测试和冒充测试的先验概率。我们采用 NIST SRE 2016 设定的 $C_{FR}=1$，$C_{FA}=1$，$P_{target}=0.001$ 这组参数。当这三个值选定后，选取一组 FRR 与 FAR 的取值使得 DCF 最小，此时的 DCF 即为 minDCF。本文中采用 NIST SRE 2016 设定参数的 minDCF 记为 minDCF16。因为 minDCF16 不仅考虑了两种错误不同代价，还考虑到了测试情况的先验概率，因此在评估说话人识别模型的性能上比 EER 更加合理。

为了验证本文提出的模型的说话人识别性能，本文对比了几种不同的方法。其中，x-vector 为用 x-vector 模型直接识别带噪语音的结果；DNN-x 为经过 DNN 增强之后的语音在 x-vector 模型上的结果；CNN-ori-x 为未加入辅助任务的多任务 CNN 模型增强之后的语音在 x-vector 模型上的结果；CNN-x 为经过本文提出的 CNN 模型增强之后的语音在 x-vector 模型上的结果；CNN-direct 为本文提出的 CNN 模型直接进行说话人识别的结果。各算法的在不同信噪比下的 EER 以及 minDCF16 如表 4 所示。

表 4 说话人识别性能测试结果

Table 4 Results of speaker recognition performance test

model	输入 SNR	EER	minDCF16
x-vector	-5dB	23.12	0. 9683
	0dB	19.30	0.9473
	10dB	12.73	0.8854
DNN-x	-5dB	17.75	0.8968
	0dB	11.96	0.8303
	10dB	4.68	0.7526
CNN-ori-x	-5dB	10.56	0.8501
	0dB	6.77	0.8321
	10dB	3.94	0.7194
CNN-x	-5dB	8.63	0.8625
	0dB	5.91	0.8244
	10dB	3.66	0.6912
CNN-direct	-5dB	7.21	0.8245
	0dB	5.83	0.7569
	10dB	3.47	0.6830

由表 4 可知，噪声环境下，尤其是低信噪比情况，x-vector 的识别效果受到了较大的影响，说明对带噪语音进行预处理是有必要的。经过 DNN 增强之后的语音在 x-vector 上的识别率有了较大提高，但是效果依然不理想。另外可以看到，尤其在-5dB 情况下，CNN 增强之后的识别率要显著高于 DNN 模型，说明融合网络的说话人识别任务通过参数共享给增强模型提供了更多的说话人信息。这其中，CNN-x 的效果要好于 CNN-ori-x，说明辅助任务帮助 CNN 语音增强保留了更多的说话人信息；CNN-direct 模型的效果要略好于 CNN-x 模型，原因可能是 CNN 的识别效果要好于 x-vector 中的 TDNN、x-vector 中的统计池化层（Statistic Pooling）和因为训练集的不同所导致的信道差异。

5 结论

在本文中，针对传统语音增强模型中忽略说话人信息的问题，研究了噪声干扰下的语音增强技术及对说话人识别的影响，提出了一种基于多任务和辅助任务约束的卷积神经网络语音增强方法。通过构建语音增强与说话人识别的融合网络多任务学习模型，同时在输入输出端拼接 MFCC 和基音周期特征，以及利用同方差不确定性自适应调整损失权重，显著提高了语音增强的效果，减少了去噪语音的失真，并在噪声干扰下的说话人识别任务上取得了优秀的表现。

References:

[1] Gannot S, Burshtein D, Weinstein E. Iterative and sequential Kalman filter-based speech enhancement algorithms[J]. IEEE Transactions on speech and audio processing, 1998, 6(4): 373-385.

[2] Karam M, Khazaal H F, Aglan H, et al. Noise removal in speech processing using spectral subtraction [J]. Journal of Signal and Information Processing, 2014, 5(2): 32

[3] Xia B, Bao C. Wiener filtering based speech enhancement with weighted denoising auto-encoder and noise classification[J]. Speech Communication, 2014, 60(7): 13-29.

[4] Ephraim Y, Malah D. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator[J]. IEEE Transactions on acoustics, speech, and signal

processing, 1984, 32(6): 1109-1121.

[5] Xu Y, Du J, Dai L R, et al. An experimental study on speech enhancement based on deep neural networks[J]. IEEE Signal processing letters, 2013, 21(1): 65-68.

[6] Variani E, Lei X, McDermott E, et al. Deep neural networks for small footprint text-dependent speaker verification[C]. 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014: 4052-4056.

[7] Snyder D, Garcia-Romero D, Povey D, et al. Deep Neural Network Embeddings for Text-Independent Speaker Verification[C]. Interspeech, 2017: 999-1003.

[8] Snyder D, Garcia-Romero D, Sell G, et al. X-vectors: Robust dnn embeddings for speaker recognition[C].2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018: 5329-5333.

[9] Zhang H R. Research on Robustness of Speaker Recognition in Noisy Environment [D]. Jiangsu: Nanjing University of Posts and Telecommunications, 2018.

[10] Xu Y, Du J, Huang Z, et al. Multi-objective learning and mask-based post-processing for deep neural network based speech enhancement[C]. Interspeech, 2015: 1508-1512.

[11] Tan Z, Mak M W, Mak B K W, et al. Denoised senone i-vectors for robust speaker verification[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 26(4): 820-830.

[12] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]. Advances in neural information processing systems, 2012: 1097-1105.

[13] Park S R, Lee J W. A Fully Convolutional Neural Network for Speech Enhancement[J]. Evaluation, 2017, 10(2): 5.

[14] Argyriou A, Evgeniou T, Pontil M. Multi-task feature learning[C]. Advances in neural information processing systems, 2007: 41-48.

[15] Misra I, Shrivastava A, Gupta A, et al. Cross-stitch networks for multi-task learning[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 3994-4003.

[16] Xu Y, Du J, Dai L R, et al. A regression approach to speech enhancement based on deep neural networks[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014, 23(1): 7-19.

[17] Fu S W, Tsao Y, Lu X. SNR-Aware Convolutional Neural Network Modeling for Speech Enhancement[C]. Interspeech, 2016: 3768-3772.

[18] Bregman A S. Auditory scene analysis: The perceptual organization of sound[M]. Massachusetts: MIT press, 1994.

[19] Kendall A, Gal Y, Cipolla R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018: 7482-7491.

[20] Piczak K J. ESC. Dataset for environmental sound classification[C]. Proceedings of the 23rd ACM international conference on Multimedia, 2015: 1015-1018.

[21] Bu H, Du J, Na X, et al. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline[C]. 2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment, 2017: 1-5.

附中文参考文献：

[9] 张洪冉. 噪声环境下说话人识别的鲁棒性研究[D]. 江苏: 南京邮电大学, 2018.