ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 34,No. 4,Jul. 2019,pp. 615－624
DOI:10. 16337/j. 1004-9037. 2019. 04. 005

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

一种基于LSTM-RNN的喉振传声器语音盲增强算法

郑昌艳　张雄伟　曹铁勇　杨吉斌　孙　蒙　邢益搏

(陆军工程大学,南京,210007)

摘　要:喉振传声器以其优良的抗噪声特性已在多种强噪声场景中得到应用,但其产生的语音尚存在着中频成份厚重、高频成份缺失等问题,严重影响了语音的清晰度和可懂度。为改善喉振传声器的语音质量,本文提出了一种基于长短时记忆递归神经网络(Long short term memory recurrent neural networks, LSTM-RNN)的喉振传声器语音盲增强算法。与基于低维的谱包络特征估计算法不同,该算法首先利用LSTM-RNN对喉振传声器语音与空气传导语音的高维对数幅度谱之间的转换关系进行建模,能有效捕捉上下文信息实现语音幅度谱的重构,然后采用非负矩阵分解(Non-negative matrix factorization, NMF)对估计出的语音幅度谱进行处理,有效抑制了过平滑问题,进一步提高了语音质量。仿真实验得到的LLR,LSD,PESQ性能指标表明,该算法可有效改善喉振传声器的语音质量。

关键词:喉振传声器;语音盲增强;递归神经网络;长短时记忆;非负矩阵分解

中图分类号:TN912.3　　**文献标志码**:A

Blind Enhancement Algorithm for Throat Microphone Speech Based on LSTM Recurrent Neural Networks

Zheng Changyan, Zhang Xiongwei, Cao Tieyong, Yang Jibin, Sun Meng, Xing Yibo

(Army Engineering University, Nanjing, 210007, China)

Abstract: Throat microphones have been used in a variety of strong noise scenarios due to their excellent anti-noise characteristics. However, the generated speech has some shortcomings such as much higher energy in middle frequency and severe loss of high frequency, which have greatly affected the speech quality and intelligibility. In order to improve the speech quality, a blind speech enhancement algorithm based on long short memory recurrent neural networks (Long short term memory recurrent neural networks, LSTM-RNN) is proposed. In contrast to previous estimation algorithms based on low-dimensional spectral envelope features, this algorithm first models the relationship of the high-dimensional logarithmic amplitude spectrum between the throat and air-conducted microphone speech directly, and this kind of neural networks can impressively capture the context information to reconstruct the signal. Secondly, the estimated speech amplitude spectrum is processed by non-negative matrix factorization (Non-negative matrix factorization, NMF), which can effectively suppress the over-smoothing problem and further improve the speech quality. The simulation results of LLR, LSD, PESQ show that this algorithm can effectively improve the speech quality of throat microphones.

Key words: throat microphone; speech blind enhancement; RNN; LSTM; NMF

基金项目:国家自然科学基金(61471394,61402519)资助项目。
收稿日期:2018-03-18;**修订日期**:2018-05-07

引 言

人体传声器(Body-conducted microphone, BCM)[1-2]是一种利用人体骨头或者组织的振动产生语音信号的设备。现有的BCM设备包括喉振传声器(Throat microphone,TM)、头骨传声器(Headset microphone,HM)以及利用耳后组织的非声耳语传声器(Nonaudible murmur microphone, NAM)等。与常见的空气传导麦克风(Air-conducted microphone, ACM)不同,BCM采集的信号基本不受环境噪声干扰,具有很强的抗噪性能,因此常被应用于军事、工厂、极限运动、医疗等强噪声场合。例如,文献[2]利用NAM实现咽喉受损患者语音交流,文献[3]利用HM协助战场士兵通信,文献[4]利用TM实现鲁棒的语音识别。

虽然BCM具有很强的抗噪性能,但是由于人体信号传导的低通性,其语音高频成份衰减严重,截止频率通常在2.5 kHz左右。并且由于声音不再经过口腔、鼻腔等传播路径,爆破音、擦音、鼻音等成份丢失。再加上设备机械振动的固有特性,语音的中频成份相比于ACM语音厚重[5-6]。这些问题使得BCM语音听起来比较沉闷,语音质量达不到人耳舒适度需求,从而在一定程度上影响了BCM的进一步推广应用。

近年来,诸多学者开展了与BCM语音相关的语音增强算法的研究,但是在多数情况下,BCM只是作为ACM语音增强的辅助。例如,文献[7]通过设计自适应的线性与非线性相结合的滤波,融合BCM语音与带噪ACM语音,文献[8]通过线性融合ACM与TM的声学特征来提高语音识别率。上述增强算法在增强阶段必须同时具有TM与ACM语音信息,在强噪声环境下,带噪ACM语音可能完全不可用,并且一些BCM设备并未配置ACM,因此存在较大的应用局限性。

BCM语音盲增强(Blind enhancement),原称盲恢复(Blind restoration)[9],是指在增强阶段直接从已有的BCM语音中推断出纯净ACM语音信号,而不需要ACM语音信息作为辅助。现有的BCM语音盲增强算法大都是通过转换语音谱包络特征达到增强目的。例如文献[10]利用简单神经网络建立BCM到ACM语音加权线性感知倒谱系数(Weighted linear predictive cepstrum coefficient,wLPCC)之间的转换关系;文献[9]认为线谱频率LSF比LPC特征拥有更好的稳定性,并且利用浅层递归神经网络实现特征的转换;文献[11]采用深度玻尔兹曼机神经网络,建BCM语音到ACM语音的LSF参数转换关系;文献[12]首先利用K-means聚类算法将TM语音的梅尔广义倒谱系数(Mel generalized cepstral coefficients,MGC)分为10类,每一类分别建立简单神经网络映射MGC特征关系,以实现语音特征更精细的转换;文献[2]利用语音转换中常用的语音分解合成模型STRAIGHT[13](Speech transformation and representation using adaptive interpolation of weighted spectrum),将语音分解为谱包络特征、基音周期和非周期成份,利用GMM建立NAM与ACM语音梅尔倒谱系数之间的转换关系。上述增强算法可以较好改善BCM语音谱包络特征,但是由于特征维数较低,谱的细节信息不能很好恢复,因此增强效果与人耳舒适度需求仍有较大差距。

本文提出了一种基于特定说话人的喉振传声器语音盲增强算法,该算法利用长短时记忆递归神经网络模型[14](Long short term memory recurrent neural networks,LSTM-RNN)直接建模TM语音和ACM语音高维对数幅度谱特征之间的映射关系,这种神经网络能够有效利用上下文信息实现特征学习,然后针对神经网络输出过平滑问题,利用非负矩阵分解(Non-negative matrix factorization,NMF)[15]算法对估计出的幅度谱进行抑制平滑处理。

1 算法思路

TM语音和ACM语音可看成由同一激励源(人的喉头)产生的信号,那么TM语音盲增强的关键就是要找TM语音到ACM语音声道特征之间的转换关系。显然,这是一种复杂的非线性转换关系,由于

TM语音丢失了经过口腔、鼻腔等辐射的语音音素,并且不同人的身体传导特性也不尽相同,因此这种转换关系不仅基于语音音素,而且也基于特定说话人。

1.1 基于高维特征转换的TM语音盲增强

以往的TM语音盲增强算法均是基于语音源-滤波器模型,将语音分解为激励(源)特征和声道(滤波器)特征,在假定激励特征不变的情况下,对低维的声道参数特征(如LSF、MGC)进行映射以实现语音增强。这些低维参数特征能够反映出语音谱包络的变化趋势,但对谱的细节信息描述不够,因而增强效果有限。为获取更高的增强语音质量,本文提出了一种基于高维谱特征转换的TM语音盲增强算法模型,并利用深度学习技术实现了TM与ACM语音高维特征间的有效转换,算法的总体思路如图1所示。

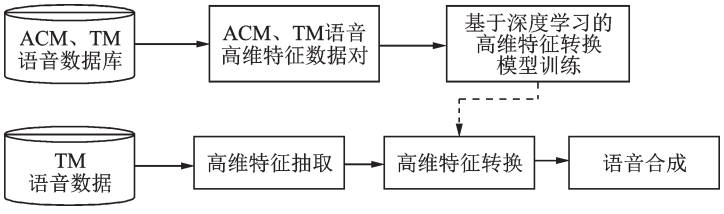

图1 TM语音盲增强总体思路

Fig.1 Framework of TM speech blind enhancement algorithm

该算法首先根据基于信号的语音分解合成模型,将语音分解为高维幅度谱与相位谱,通过转换高维幅度谱实现TM语音盲增强。考虑到对数幅度谱能够对幅度谱进行有效压缩,减少动态范围,易于神经网络训练,算法最终选取了对数幅度谱作为转换特征。

1.2 特征映射模型选取

TM与ACM语音在幅度谱上的差异主要表现为高频成份的严重丢失。从信息丢失严重的TM语音中恢复出高频信号并非易事,也可将这种恢复视为一种人工频谱扩展(Artificial bandwidth extension),简称频谱扩频。但是传统的频谱扩频的目的是将原始语音信号从0.3~3.7 kHz扩展到0.3~8 kHz左右,关注的是电信网络传输语音信号的音质,而TM语音截止频率约为2.5 kHz,不仅是人耳的听觉感知受到了影响,很多与内容相关的信息也丢失了。这种丢失的信息并不能简单地从单个语音帧的低频信息推断出,而是必须结合上下文信息,从语境中“猜测”丢失的信号。

深度学习强大的非线性映射能力使得高维特征之间的建模成为了可能。递归神经网络模型能够利用其内部的递归结构实现上下文信息的建模,因而更适合建模TM语音的“频谱扩展”问题。LSTM-RNN引入了精心设计的记忆单元结构解决了传统递归神经网络梯度爆炸和消失的问题,使得学习序列长时信息成为了可能。本文正是利用LSTM-RNN强大的序列学习能力,实现TM语音丢失信息的恢复。

1.3 神经网络输出的后处理

神经网络在训练过程中,会依据TM和ACM语音的对数幅度谱之间的距离调整网络参数,调整中默认每个频点差距对距离的贡献是相同的。这种平均贡献会产生数据过平滑问题,因为语音数据的结构特点并未体现其中。

NMF是一种经典的字典学习方法,它能够将一个非负矩阵分解为两个非负矩阵的乘积,其中一个矩阵反映原矩阵的局部特征(又称为字典矩阵),另一个则反映这些特征的大小与增益称之为激活矩阵。由于字典基的数量远远小于原始数据的个数,为尽可能地还原原始信息,NMF能够有效地捕捉数据的结构特点[16]。本文利用NMF的这一优点缓解神经网络输出数据过平滑问题,这种后处理方法已在语噪分离[17]、频谱扩展[18]中得到成功应用。

2 算法实现

2.1 算法实现流程

算法的具体实现分为训练阶段和增强阶段。训练阶段主要包括：TM与ACM语音的特征抽取、基于LSTM-RNN的特征转换模型训练以及基于NMF的ACM语音特征字典学习。增强阶段主要包括：TM语音特征的提取、基于LSTM-RNN模型的特征转换、基于NMF的神经网络输出过平滑处理，以及最终的增强语音合成。需指出的是，为使神经网络更好地收敛，需要对神经网络的输入数据进行高斯归一化[19]。

算法在训练阶段的具体步骤为：

步骤1 对训练集的TM语音$x(n)$和ACM语音$s(n)$分帧加窗并进行短时傅里叶变换，分别得到TM与ACM语音幅度谱特征X与S；

步骤2 对幅度谱特征X与S进行对数变换得到对数幅度谱$\log(X)$与$\log(S)$，计算出对数幅度谱每一维的均值与方差，记为$\bar{X}$，σ_X和$\bar{S}$，σ_S；

步骤3 对数幅度谱$\log(X)$与$\log(S)$分别进行高斯归一化，计算公式为

$$\log_{\text{Norm}}(X)=[\log(X)-\bar{X}]/\sqrt{\sigma_X} \tag{1}$$

$$\log_{\text{Norm}}(S)=[\log(S)-\bar{S}]/\sqrt{\sigma_S} \tag{2}$$

步骤4 将$\log_{\text{Norm}}(X)$作为输入，$\log_{\text{Norm}}(S)$作为输出目标，训练LSTM-RNN模型，得到训练好的模型，记为G；

步骤5 利用NMF对ACM语音幅度谱S进行分解，得到字典矢量基$\boldsymbol{D}_A$。

算法在增强阶段的具体步骤为：

步骤1 对待增强的TM语音$t(n)$分帧加窗并进行短时傅里叶变换，得到TM语音幅度谱特征T与相位谱特征P；

步骤2 对幅度谱特征T取对数得到对数幅度谱$\log(T)$，并根据训练集得到的TM语音对数幅度谱均值与方差进行高斯归一化，计算公式为

$$\log_{\text{Norm}}(T)=[\log(T)-\bar{X}]/\sqrt{\sigma_X} \tag{3}$$

步骤3 利用训练好的LSTM-RNN模型对特征进行转换，得到输出$\log(\hat{S})=G(\log_{\text{Norm}}(T))$；

步骤4 依据训练集ACM语音特征的均值、方差进行反归一化，并进行指数计算，得到估计的幅度谱为

$$\hat{S}=\exp([\log(\hat{S})\cdot\sqrt{\sigma_S}+\bar{S}]) \tag{4}$$

步骤5 根据训练阶段得到的字典$\boldsymbol{D}_A$，对$\hat{S}$进行过平滑处理，得到最终估计的幅度谱$\hat{S}_{\text{NMF}}$；

步骤6 利用反傅里叶变换(Inverse short time Fourier transform, ISTFT)将$\hat{S}_{\text{NMF}}$与P合成增强语音$t_E(n)$。

整个算法的具体实现流程如图2所示，为简练起见，数据的高斯归一化与反归一化过程未在图中体现。

2.2 基于LSTM-RNN的特征映射

设TM的第i帧语音的对数幅度谱特征为X_i，相对应的ACM语音对数幅度谱特征为$\boldsymbol{S}_i$，并且均已经过高斯归一化。LSTM-RNN需联立多帧语音信息建模上下文关系，联立的帧数称为迭代步长，设为$2m+1$，其中m为整数，$0\leqslant m<N$。联立形式通常为开窗，即连接前后m帧信息推断中间帧信息。因

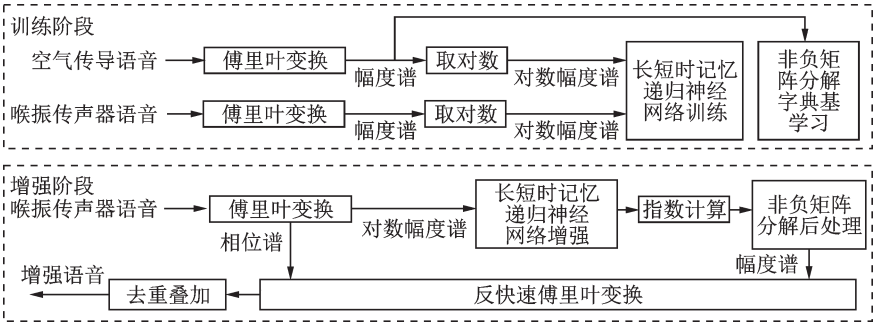

图2 算法实现流程

Fig.2 Flowchart of the proposed algorithm

此,LSTM-RNN的输入$\boldsymbol{x}_n$可表示为如下形式

$$\boldsymbol{x}_n=[X_{i-m},X_{i-m+1},\cdots,X_i,X_{i+1},\cdots,X_{i+m}]$$

式中:n为输入样本个数索引。对应的目标输出$\boldsymbol{y}_n=\boldsymbol{S}_i$,网络的训练目标函数为均方误差函数,如式(5)所示。

$$J_{\mathrm{MSE}}(\boldsymbol{W},b)=\frac{1}{N}\sum_{n=1}^{N}\frac{1}{2}\left\|\tilde{\boldsymbol{y}}_n-\boldsymbol{y}_n\right\|^2 \tag{5}$$

$$\tilde{\boldsymbol{y}}_n=f_{\boldsymbol{W},b}(\boldsymbol{x}_n) \tag{6}$$

式中:N为样本总数;$\tilde{\boldsymbol{y}}_n$为对数幅度谱$\boldsymbol{y}_n$的估计;$f_{\boldsymbol{W},b}$指经过LSTM-RNN的非线性转换函数;$\boldsymbol{W}$为神经网络的权值矩阵;b为神经网络偏置值。LSTM-RNN根据目标函数计算出估计的对数幅度谱与目标对数幅度谱之间的误差,并根据此误差利用基于时间的反向传播算法(Back propagation through time)更新神经网络参数。

与受限玻尔兹曼机-深度置信网络不同的是,LSTM-RNN输入信息并不是多帧语音信息的简单联合,它通过在激活单元中设计了3种门结构即输入门、遗忘门、输出门和一个记忆状态,实现了无用信息的丢弃和有用信息的保留,从而控制了信息流在神经网络中的有效流动。若没有丢弃无用信息的过程,则过多的信息会导致神经网络难以拟合,从理论上也可证明语音的前后帧信息对于推断当前帧信息并非都是有用的。

LSTM-RNN中,输入门i_t、遗忘门f_t和输出门o_t以及当前时刻记忆单元的状态值c_t的计算过程如下

$$i_t=\delta(\boldsymbol{W}_{xi}x_t+\boldsymbol{W}_{hi}h_{t-1}+\boldsymbol{W}_{ci}c_{t-1}+b_i) \tag{7}$$

$$f_t=\delta(\boldsymbol{W}_{xf}x_t+\boldsymbol{W}_{hf}h_{t-1}+\boldsymbol{W}_{cf}c_{t-1}+b_f) \tag{8}$$

$$c_t=f_tc_{t-1}+i_t\tanh(\boldsymbol{W}_{xc}x_t+\boldsymbol{W}_{hc}h_{t-1}+b_c) \tag{9}$$

$$o_t=\delta(\boldsymbol{W}_{xo}x_t+\boldsymbol{W}_{ho}h_{t-1}+\boldsymbol{W}_{co}c_t+b_o) \tag{10}$$

$$h_t=o_t\tanh(c_t) \tag{11}$$

式中:x_t为当前时刻的输入值,对应的是x_n中的一帧;h_t是隐藏层输出;$\boldsymbol{W}$为权重矩阵,例如$\boldsymbol{W}_{xf}$指输入x_t与遗忘门f之间的权重矩阵;b为偏置值,例如b_f为遗忘门偏置值;δ为激活函数。

上述公式清楚地展现了当LSTM-RNN接收到一帧数据后,会保留该帧中的有用信息,丢弃无用信息,并且更新记忆状态值,而此记忆状态存储着该帧之前所有的有用信息,由此上下文信息得到了联系。再输入下一帧数据,LSTM-RNN重复同样的动作,直到达到最大的迭代步长,即完成了上下文所有信息$\boldsymbol{x}_n$的输入,才可得到最终的输出。

神经网络的最终输出$\boldsymbol{y}_n$需经过以下反归一化变换得到估计的对数幅度谱

$$\boldsymbol{y}_n'(k)=\boldsymbol{y}_n(k)\times\sqrt{v(k)}+m(k) \tag{12}$$

式中：k表示第k维；$\boldsymbol{y}_n'$为重构的对数幅度谱；$v(k)$、$m(k)$分别为ACM语音对数幅度谱的第k维的方差与均值。

2.3 过平滑处理

LSTM-RNN虽然能够很好地建模高维数据之间的相关关系，但是其输出$\boldsymbol{y}_n'$存在过平滑的问题，利用NMF来缓解这个问题。

首先将训练集的ACM语音幅度谱S经过非负矩阵分解算法得到字典$\boldsymbol{D}_A$及其激活矩阵$\boldsymbol{H}_A$，选取KL散度距离作为优化目标函数，如式(13)所示。式(14)，(15)分别为字典矩阵和激活矩阵计算的迭代过程。

$$\min\sum_{ij}\left[S_{ij}\ln\frac{1}{(\boldsymbol{D}_A\boldsymbol{H}_A)_{ij}}+(\boldsymbol{D}_A\boldsymbol{H}_A)_{ij}\right] \tag{13}$$

$$(\boldsymbol{H}_A)_{kj}\leftarrow(\boldsymbol{H}_A)_{kj}\frac{\sum_i(\boldsymbol{D}_A)_{ik}S_{ij}/(\boldsymbol{D}_A\boldsymbol{H}_A)_{ij}}{\sum_i(\boldsymbol{D}_A)_{ik}} \tag{14}$$

$$(\boldsymbol{D}_A)_{ik}\leftarrow(\boldsymbol{D}_A)_{ik}\frac{\sum_j(\boldsymbol{H}_A)_{ik}S_{ij}/(\boldsymbol{D}_A\boldsymbol{H}_A)_{ij}}{\sum_j(\boldsymbol{H}_A)_{kj}} \tag{15}$$

式中：字典矩阵$\boldsymbol{D}_A$大小为$K\times T$，K等于幅度谱特征维度；T为字典基矢量个数，激活矩阵$\boldsymbol{H}_A$大小为$T\times N$，N为训练集样本个数；i，j分别为矩阵的行、列索引。

在得到$\boldsymbol{D}_A$后，固定字典矩阵对神经网络估计的幅度谱$\hat{S}$进行分解，可得到激活矩阵$\boldsymbol{H}_T$，最终得到抑制平滑后的幅度谱$\hat{S}'=\boldsymbol{D}_A\times\boldsymbol{H}_T$。

非负矩阵分解可对神经网络输出的幅度谱特征进行稀疏化重构，因而可以抑制过平滑问题。将经过NMF处理后的幅度谱与TM语音的相位谱经过反傅里叶变换并进行去重叠加操作，得到重构的增强语音。

3 仿真实验及结果分析

3.1 实验数据及评估方法

目前，国内外没有公开可用的数据库，本文首先制作了某型号的TM设备语音与ACM语音的平行语音数据库。该数据库包括800个语句，由2男2女录制完成。录制时，每个人需同时佩戴喉振传声器和普通空气传导麦克风，并在声暗室中进行标准普通话录制。采用Cooledit软件录制，采样率为32 kHz，采用16 bit量化。录音语料来源于报纸、网络以及一些人为构造的音素平衡语句。每人共录制200句语音，每句话时长约在3～4 s，200句语音被分为160句作为训练集，40句作为测试集，训练集与测试集中没有重复语料。

在模型训练前，首先对TM与ACM语音降采样到8 kHz，然后进行能量归一化，使得两者语音能量在相近的动态范围内。语音特征提取时，帧长设为32 ms，帧移设为10 ms，STFT频点设为256，即得到的幅度谱维度实际为129维，幅度谱取对数后进行高斯归一化处理。

在评价指标中，采用了3种客观评价指标：对数谱距离(Log-spectral distance, LSD)、感知语音质量评估方法(Perceptual evaluation of speech quality, PESQ)和对数似然比(Log-likelihood ratio, LLR)。LSD反映增强语音与理想ACM语音之间的对数幅度谱距离，其值越小表明语音质量越高。LLR是衡

量语音线性预测系数距离的一种指标,其值越小表明语音质量越好。PESQ是一种能够很好评价语音主观试听效果的评价指标,其得分越高,表明语音质量越好。

3.2 LSTM-RNN与NMF模型参数设置

通过参数调整实验,本文得到的最优LSTM-RNN模型参数设置如下:2个隐层,每个隐层的单元个数为512,隐层的激活函数为正切(tanh)函数,输出层激活函数为线性函数,隐层丢弃正则化比率为0.2,迭代步长为23帧。

在LSTM-RNN训练过程中,随机选取10%的训练数据作为验证集,每批次(Batchsize)送入的数据数量为128,采用均方根传播(Root mean square propagation, RMSProp)算法更新网络参数,初始学习率设为0.01,当验证集误差不再减少时则学习率降为原来一半,直到验证集误差连续2次不再减少,则停止训练。比较不同LSTM-RNN参数设置下的验证集误差值,验证集误差最小的模型参数即为本文选取的最优LSTM-RNN模型参数。

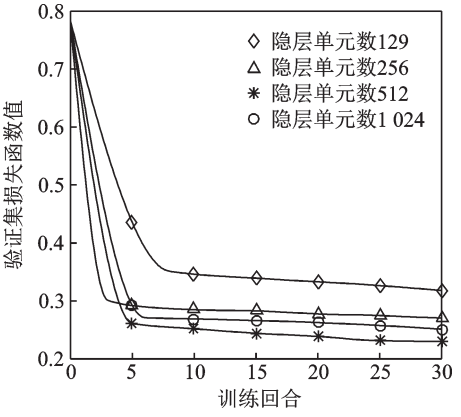

图3 LSTM-RNN不同隐层单元个数下的验证集误差

Fig.3 Validation loss of LSTM-RNN with different numbers of hidden units

图3为女声1数据训练时,固定隐层数为2,迭代步长为23,不同的隐层单元个数下的验证集误差值,横轴为训练的回合数。从图3中可看出当隐层单元数为512时,验证集损失函数值最小。对比隐层单元数为129,256时验证集损失函数值,可看出,随着隐层单元数的增加,验证集损失函数值降低明显,说明只有隐层单元数达到一定个数时,才能充分实现LSTM-RNN的拟合性能。对比隐层单元数为512和1 024时验证集损失函数值,可看出,隐层单元个数并非越大越好,过大的隐层单元数会增加模型复杂度,也会影响LSTM-RNN的拟合。

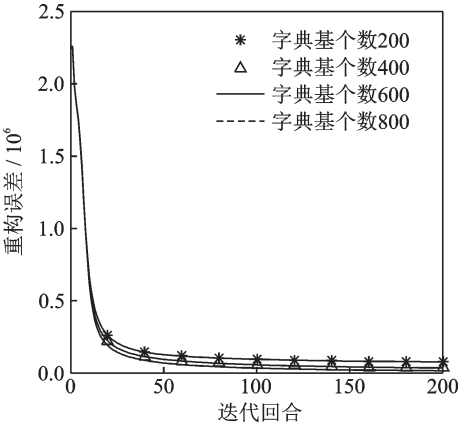

图4 NMF不同字典基个数下的重构误差

Fig.4 Reconstructed error of NMF with different numbers of dictionary atoms

本文依据NMF训练集ACM语音幅度谱分解时的重构误差值,选取最优的NMF字典个数,实验结果如图4所示。从图4中可看出,随着字典个数的增加,重构误差值逐渐减小,当字典个数达到600时,再增加字典个数,重构误差值已无明显降低,因此,本文选取的最优NMF字典个数为600。

3.3 实验结果分析

为对比不同神经网络结构对高维特征转换的效果,本文将基于LSTM-RNN的特征转换算法(未经过NMF后处理,记为LSTM)与基于受限玻尔兹曼机-深度置信神经网络(记为DNN)的特征转换算法进行了比较,将基于DNN、LSTM-RNN的特征转换并且经过NMF后处理的算法分别记为DNN-NMF、LSTM-NMF。

实验结果如表1—3所示,测试结果为每人40句测试语句的评价指标平均得分值,其中,TM指原始TM语音与ACM语音的对比结果,3种增强算法均为增强后的语音与ACM语音的对比结果。

由表1可看出,无论是经过DNN还是LSTM增强,增强后的语音LSD都明显减小,说明神经网络

表 1 对数谱距离比较(LSD)

Tab.1 Comparisons of LSD

说话人	TM	增强算法			
		DNN	DNN-NMF	LSTM	LSTM-NMF
男 1	1.48	1.13	1.07	0.97	0.89
男 2	1.44	1.15	1.08	0.97	0.90
女 1	1.47	1.15	1.05	1.01	0.91
女 2	1.50	1.14	1.02	0.99	0.87

表 2 对数似然比距离比较(LLR)

Tab.2 Comparisons of LLR

说话人	TM	增强算法			
		DNN	DNN-NMF	LSTM	LSTM-NMF
男 1	1.18	0.57	0.56	0.43	0.42
男 2	1.32	0.63	0.61	0.48	0.47
女 1	1.33	0.63	0.59	0.53	0.48
女 2	1.41	0.65	0.62	0.53	0.50

能够很好拟合高维特征。LSTM 的拟合效果明显优于DNN,说明LSTM-RNN的神经网络结构更适合于 TM 语音盲增强。DNN、LSTM 的输出经过 NMF 处理后 LSD 进一步减小,验证了不同的神经网络结构下,NMF 均能够有效抑制神经网络输出过平滑问题。表 2 的对数似然比评价指标与 LSD 结果类似。

表 3 感知语音质量比较(PESQ)

Tab.3 Comparisons of PESQ

说话人	TM	增强算法			
		DNN	DNN-NMF	LSTM	LSTM-NMF
男 1	2.27	2.53	2.55	2.86	2.88
男 2	2.10	2.29	2.31	2.71	2.73
女 1	1.97	2.13	2.23	2.38	2.44
女 2	1.92	2.15	2.22	2.32	2.39

由表 3 可知,相比于 DNN,LSTM 在 PESQ 值有了较大提升,证明了这种递归神经网络结构能够有效提高 TM 语音的感知语音质量。男声数据提升效果明显优于女声,原因是 TM 语音的高频成份丢失,而男声语音高频成份远少于女声,因此恢复相对较为容易。

图 5,6 展示女、男声的语谱图,其中图 5(a),6(a)为 ACM 语音,图 5(b),6(a)为 TM 语音,图 5(c),6(c)为经 LSTM 算法增强后的语音,图 5(d),6(d)为 LSTM-NMF 算法增强后的语音。对比图 5(a),5(b)以及图 6(a),6(b)可看出,相比于 ACM 语音,TM 语音 2.5 kHz 以上的能量几乎已完全衰减,并且中频谐波能量没有起伏,这也就是 TM 语音听起来沉闷、不自然的原因;对比图 5(a)与图 6(a)可知,女声 ACM 语音高频成份明显多于男生 ACM 语音,在客观指标的分析中指出,这是女声相对于男声较难恢复的原因。由图 5 与图 6 的(c)和(d)可观察出,LSTM、LSTM-NMF 增强算法都较好恢复了 TM 语音高频丢失的成份,证明了增强算法的有效性;由图 5,6 虚线椭圆中的成份可看出,LSTM-NMF 增强算法相比于 LSTM,可

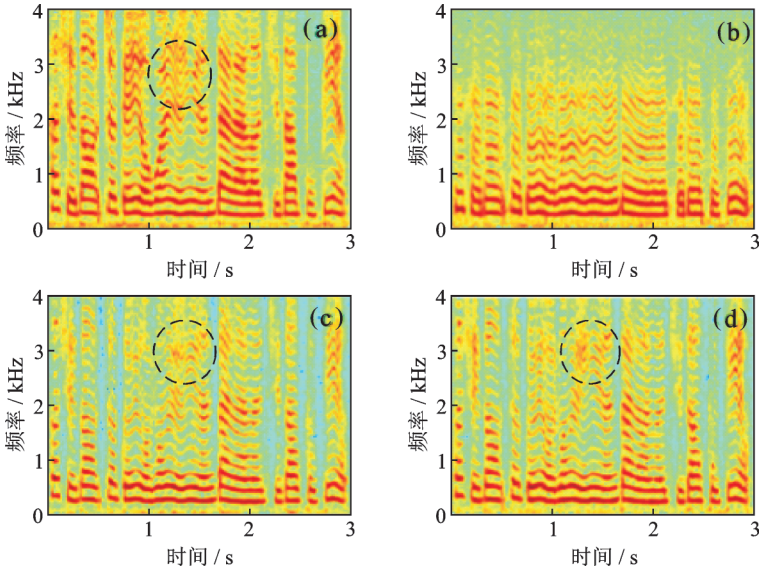

图 5 女声语谱图

Fig.5 Spectrograms of a female speaker

获得更接近ACM语音的数据,验证了NMF能够有效抑制神经网络输出过平滑问题。

由以上结果可得出以下结论:深度神经网络可有效建模TM与ACM语音高维特征之间的相关关系;相比于DNN,能够实现长时序列有效建模的LSTM-RNN可得到更好的TM语音增强效果,NMF能够有效抑制神经网络输出过平滑问题。

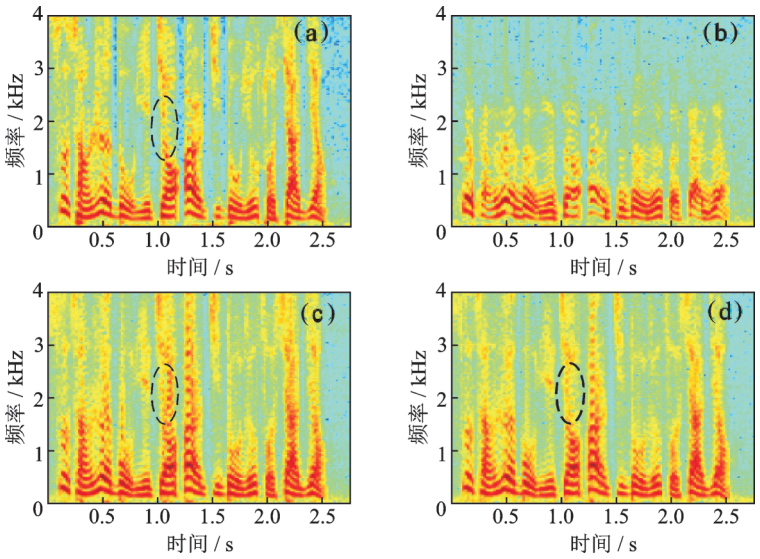

图6 男声语谱图

Fig.6 Spectrograms of a male speaker

4 结束语

本文提出了一种基于LSTM-RNN的喉振传声器语音盲增强算法。该算法首先利用LSTM-RNN建模喉振传声器语音与空气传导语音高维对数幅度谱特征之间的相关关系,然后利用NMF对估计出的幅度谱进行处理以抑制神经网络输出过平滑问题。实验结果表明,该算法能有效提高特定说话人的喉振传声器语音质量,增强效果优于受限玻尔兹曼机-深度置信神经网络以及单一的长短时记忆递归神经网络。该算法对男声的增强效果明显优于女声,经分析是由于该算法生成的高频成份与真实数据分布间存在偏差,而女声的高频成份较多,因此不易恢复。下一步将针对高频成份的生成问题,拟通过生成式对抗神经网络[20]进一步对生成的数据分布进行修正,以缩小生成的高频成份与真实数据分布间的差异。

参考文献:

[1] Shin H S, Kang H G, Fingscheidt T. Survey of speech enhancement supported by a bone conduction microphone[C]//Speech Communication, 10 ITG Symposium 2012. Braunschweig, Germany:IEEE, 2012: 1-4.

[2] Toda T, Nakagiri M, Shikano K. Statistical voice conversion techniques for body-conducted unvoiced speech enhancement[J]. IEEE Transactions on Audio Speech & Language Processing, 2012, 20(9): 2505-2517.

[3] Tran P, Letowski T, Mcbride M. Bone conduction microphone: Head sensitivity mapping for speech intelligibility and sound quality[C]// International Conference on Audio, Language and Image Processing 2008. [S.l.]: [s.n.], 2008: 107-111.

[4] Erzin E. Improving throat microphone speech recognition by joint analysis of throat and acoustic microphone recordings[J]. IEEE Transactions on Audio Speech & Language Processing, 2009, 17(7): 1316-1324.

[5] 李静. 基于骨导信号的语音重构技术[D]. 西安:西北工业大学, 2004.
Li Jing. Speech reconstruction technology based on bone-conducted signal [D]. Xi′an: Northwestern Polytechnical University, 2004.

[6] Turan M A T, Erzin E. Source and filter estimation for throat-microphone speech enhancement[J]. IEEE/ACM Transactions on Audio Speech & Language Processing, 2016, 24(2): 265-275.

[7] 李敏杰. 骨导和气导结合的语音增强系统搭建[D]. 哈尔滨:哈尔滨工业大学, 2016.
Li Minjie. Bone-conducted and air-conducted combined speech enhancement system[D]. Harbin: Harbin Institute of Technology, 2016.

[8] Jou S C S, Schultz T, Waibel A. Adaptation for soft whisper recognition using a throat microphone[C]//Interspeech. [S.l.]: [s.

n.], 2004: 189-194.

[9] Thang Tat Vu, Unoki M, Akagi M. A blind restoration model for bone-conducted speech based on a linear prediction scheme [C]// International Symposium on Nonlinear Theory and Its Applications. [S.l.]: [s.n.], 2007.

[10] Shahina A, Yegnanarayana B. Mapping speech spectra from throat microphone to close-speaking microphone: A neural network approach[J]. Eurasip Journal on Advances in Signal Processing, 2007(1): 1-10.

[11] Huang B, Gong Y, Sun J, et al. A wearable bone-conducted speech enhancement system for strong background noises[C]// International Conference on Electronic Packaging Technology. [S.l.]: [s.n.], 2017: 1682-1684.

[12] Vijayan K K S. Comparative study of spectral mapping techniques for enhancement of throat microphone speech[C]// Communications (NCC), Twentieth National Conference. [S.l.]: [s.n.], 2014.

[13] Kawahara J E A O. Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system straight[C]// 2nd International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications. [S.l.]: [s.n.], 2001.

[14] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[15] Hennequin R, Badeau R, David B. NMF with time-frequency activations to model non stationary audio events[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. [S.l.]: IEEE, 2014: 445-448.

[16] 黄建军,张雄伟,张亚非,等. 时频字典学习的单通道语音增强算法[J]. 声学学报, 2012(5): 539-547.
Huang Jianjun, Zhang Xiongwei, Zhang Yafei, et al. Time-frequency dictionary learning for single channel speech enhancement [J]. Acoustics Journal, 2012(5): 539-547.

[17] Williamson D S, Wang Y, Wang D L. A two-stage approach for improving the perceptual quality of separated speech[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. [S.l.]: IEEE, 2014: 7034-7038.

[18] Liu B, Tao J. A novel research to artificial bandwidth extension based on deep BLSTM recurrent neural networks and exemplar-based sparse representation[C]//Interspeech. [S.l.]: [s.n.], 2016: 3778-3782.

[19] Wang Y, Zhao S, Qu D, et al. Using conditional restricted Boltzmann machines for spectral envelope modeling in speech bandwidth extension[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. [S.l.]: IEEE, 2016: 5930-5934.

[20] 王坤峰,苟超,段艳杰,等. 生成式对抗网络GAN的研究进展与展望[J]. 自动化学报, 2017, 43(3): 321-332.
Wang Kunfeng, Xun Chao, Duan Yanjie, et al. Generative adversarial networks: The state of the art and beyond[J]. Acta Automatica Sinica, 2017, 43(3): 321-332.

作者简介:

郑昌艳(1990-),女,博士研究生,研究方向:语音处理、深度学习,E-mail:echoaimaomao@163.com。

张雄伟(1965-),男,博士,教授,研究方向:语音与图像处理、多媒体信息处理,E-mail:xwZhang9898@163.com。

曹铁勇(1971-),男,教授,研究方向:智能信息处理、图像处理,E-mail:cty_ice@sina.com。

杨吉斌(1978-),男,副教授,研究方向:语音信号处理、语音编码,E-mail:yjbice@sina.com。

孙蒙(1984-),男,副教授,研究方向:语音与图像处理,机器学习,深度学习,E-mail:sunmengccjs@163.com。

邢益搏(1994-),男,硕士研究生,研究方向:语音增强、智能信息处理,E-mail:18252059100@163.com。

(编辑:张彤)