文章编号：2096-1618(2024)03-0275-08

听觉模型鲁棒性特征研究及应用

王文华， 夏秀渝
(四川大学电子信息学院，四川 成都 610064)

摘要：人类的听觉系统具有非常精细而巧妙的结构，即使在嘈杂的环境中，也能准确地理解语音。采用精细的耳蜗模型作为前端处理可以实现更好的语音处理。利用快速压缩的非对称谐振器级联(CARFAC)作为人耳外周模型，结合听觉稳定图像得到精确的皮层前听觉模型。在听觉模型的基础上提取较准确的基音轮廓，利用基音信息进行声场景分析，合成鲁棒性语音特征，并将其送入神经网络进行监督训练，以实现语音增强。实验结果表明，噪声条件下，由听觉模型提取的特征在各语音评价指标下都有较好的体现，可以更好表征语音信号，具有一定的鲁棒性。

关 键 词：CARFAC 模型；听觉稳定图像；语音增强系统；基音提取
中图分类号：TP391.4 **文献标志码**：A
doi：10.16836/j.cnki.jcuit.2024.03.003

0 引言

语音是人类最重要的交流手段。由于目标语音常会受到其他背景噪声信号的破坏，因此将语音和背景干扰噪声分离至关重要。人类通过十分精细的听觉系统，可以轻松地进行语音分离，具有从多个混合声源中提取一个声源的能力。事实证明，构建一个与人类听觉系统相匹配的自动系统是非常具有挑战性的。近几十年来，随着机器学习的发展，众多学者致力于实现与人类听觉系统功能相近的听觉模型，并在信号处理的语音分离方向进行了广泛研究。

目前，实现单通道语音增强或语音分离的方法有：语音增强、计算听觉场景分析法 CASA、基于监督学习的方法等。经典语音增强方法[1]一般是通过噪声估计，从有噪的语音中估计出干净的语音，目前应用最广泛的语音增强方法有谱减法和维纳滤波法。人耳听觉感知过程较复杂，可分为分组和重组 2 个阶段。在此基础上，提出了计算听觉场景(CASA)[2]，将听觉场景应用到语音处理中，开始了无监督式的语音分离算法。随着深度学习技术的发展，将监督学习技术应用到语音分离算法中[3-4]，利用深度神经网络学习，从带噪语音信号特征到分离目标的非线性映射关系进而实现语音分离。Chen J 等[5]采用 RNN 和 LSTM 模型实现语音分离算法。

Chen J 等[5]提出，输入特征和网络模型是可以互补的。如果提出代表性强区分度高的特征，那么对整个增强系统可以起到锦上添花的作用。因此，将较为精细的人耳听觉模型与语音增强系统结合，可作为一个突破口。

本文采用速动压缩非对称谐振器级联模型(CARFAC)[6]作为耳蜗外周模型，听觉稳定图像(SAI)[7]作为从耳蜗到脑干的神经传输模型，构建外周到听觉皮层级别的听觉模型[8]。在 SAI 的基础上进行听觉场景分析，提取语音特征，并将其送入神经网络，组成语音增强系统。在语音增强的实验中，SAI 的相关特征相比常用的语音特征，在各指标中都有一定优势。低噪声环境下，表现其较强的鲁棒性。

1 听觉模型及听觉特征

人类的听觉系统可以在噪声下进行有效的声源分离和辨识[9]，因此模拟听觉系统进行声音分析及特征提取将更加高效、鲁棒。近些年，研究出模拟人类听觉外周系统的人耳模型[10]，如 Gammatone 滤波器组、DRNL 模型以及 CARFAC 模型等。在几种模型中，CARFAC 可以更好地匹配生理模型，且具有较强的鲁棒性。因此，本文采用 CARFAC 模型实现听觉外周系统，并结合听觉稳定图像实现较完整的听觉模型，进而提取鲁棒的语音特征。

1.1 速动压缩非对称谐振模型

速动压缩非对称谐振器模型(CARFAC)可以更好地模拟人耳基本功能，匹配心理滤波器及心理冲激响应。CARFAC 模型包括基底膜模型、内毛细胞、外毛细胞模型以及担负了模型大部分压缩任务的耦合自动增益控制模块。CARFAC 数字耳蜗模型如图 1 所示。

收稿日期：2023-06-11
通信作者：夏秀渝. E-mail：xiaxy@163.com

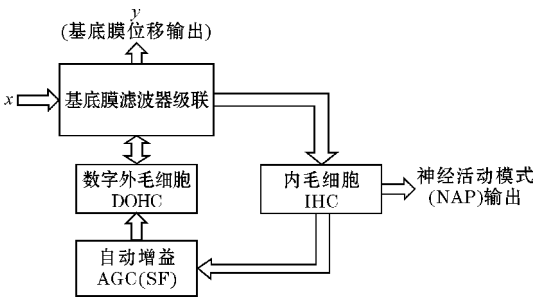

图 1 CARFAC 数据耳蜗模型

基底膜模型由双极点、双零点的非对称谐振器级联组成,单个谐振器的传递函数:

$$H(z)=\frac{Y}{X}=g\left[\frac{(z^2+(-2a_0+hc_0)rz+r^2)}{(z^2-2a_0rz+r^2)}\right]$$

式中,r 为极点半径;g 用于调整总增益,$g=(1-2a_0r+r^2)/(1-(2a_0-hc_0)r+r^2)$;参数 $a_0=\cos\theta_R=\cos(2\pi f_c/f_s)$,$c_0=\sin\theta_R=\sin(2\pi f_c/f_s)$,$f_s$ 和 f_c 分别为采样频率和截止频率;h 用于控制零点与极点的频率比率,要求 $h<(2+2a_0)/c_0$,令 $h=c_0$。改变 r 的值就可以同时移动极点和零点,从而改变阻尼实现基底膜模型的非线性变化。

基底膜中的变化受数字外毛细胞(DOHC)调控。DOHC 模块主要的实现依靠其中的非线性(NLF)函数:

$$\mathrm{NLF}(v)=\frac{1}{1+(v\cdot\mathrm{scale}+\mathrm{offset})^2}$$

式中,scale = 0.1,offset = 0.04,v 是 2 个相邻样本的差值计算所得。NLF 和 AGC 环路滤波器的反馈信号共同控制负阻尼。

利用 2 个相邻样本的差值计算变化速率,结合速率的局部瞬时非线性以及自动增益(AGC)模块的传出反馈 b,计算出相应的系数 r 非线性函数:

$$r=r_1+d_{\mathrm{rz}}(1-b)\mathrm{NLF}(v)$$

式中,r_1 原始值;d_{rz}为阻尼变化参数;b 为 AGC 增益反馈信号;v 为局部基底膜上的位移速率。

在人耳中,内毛细胞是一种传感器,可感知声音在耳蜗隔膜上产生的运动,并将结果作为传入信号传递给神经系统。数字内毛细胞(DIHC)是检测器或整流器。数字内毛细胞框图如图 2 所示。

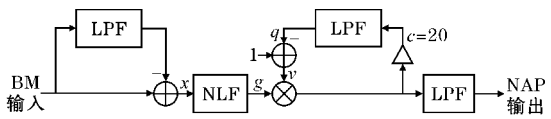

图 2 数字内毛细胞框图

基底膜的输出通过 DIHC 模型后,可得到神经活动图,用来表征 DIHC 的瞬时放电速率。DIHC 的短时平均输出控制自适应增益和 AGC 反馈环路,主要采用非线性函数来检测非线性,计算方法:

$$u=\mathrm{HWR}(x+0.175)$$

$$g=\frac{u^3}{u^3+u^2+0.1}$$

式中,HWR 表示正幅半波整流函数,x 是基底膜运动经高通滤波后的输出。

CARFAC 中一个重要的模块就是自动增益(AGC)模块。AGC 模块的每通道采用 4 个单极平滑滤波器,将其输出组合在一起;4 个滤波器的每一级在相邻通道之间都有耦合,通过这样的空间耦合可以实现侧向抑制,即相邻感受器之间互相抑制的现象。

通过 CARFAC 模型可得到神经活动模式图(NAP),该图具有精细的时序结构,有噪声情况下具备一定的鲁棒性[11],可以更好地表征声音信号的各类特征。

1.2 听觉稳定图像

听觉稳定图像(SAI)表征了从耳蜗到脑干的神经传输模式,该图像是从脑干和中脑提取并投射到听觉皮层所得到的。SAI 图像模拟发生在耳蜗和听觉皮层之间神经中枢内的转换,即将耳蜗输出转换为声音初始听觉图像,以完成对人耳听觉皮层前的模拟。此处使用触发式时序积分来实现,简单来说就是将信号与自身稀疏化后做互相关操作。稀疏化是指选取输入片段中前 n 个最高值标记为触发点,其余值都置零。将触发点与时延原点对齐,进行时序积分后即可获得输入对应的听觉稳定图像。

$$g(t,\tau)=(\hat{f}(t)f(t-\tau))*w(t)$$

式中,$\hat{f}(t)$是触发事件的稀疏序列,当且仅当在 t_{trigger} 时非零;w 是应用于乘积过往值的加权函数;$*$ 是卷积运算。为了表示不同通道,使用 x 作为通道索引,将 $f(t)$记作 $f(x,t)$显性的表示通道维数。每当触发事件出现时,对输出图像 $I(x,\tau)$的行实施离散更新规则。

$$I(x,\tau)\leftarrow af(x,t_{\mathrm{tigger}}-\tau)+\beta I(x,\tau)$$

式中,参数 α 和 β 与上次触发时间有关或与触发时的幅度有关。

图 3 是一帧浊音的 SAI 图像。图中横坐标表示时延样点数,其中 300 所在点为零时延;纵坐标表示通道数,此处共采用 64 通道。听觉图像中的两个维度分别可代表频率和基音,二维的图像展示了两者之间的相互作用。

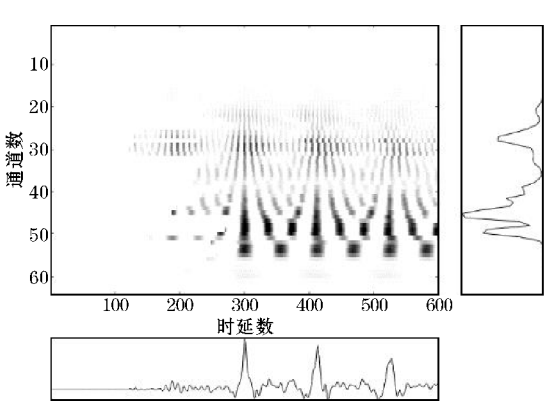

图 3 一帧浊音语音 SAI 图像

由图3可见,一帧浊音SAI图像的行均值可以表征共振峰信息,列均值可表征基音信息。每帧的行均值通过时间堆叠后可以形成类似于听觉谱图或耳蜗谱图,列均值经过时间堆叠后就是基音谱图,可得到语音的基音特征,图4和图5是一条纯净语音的耳蜗谱图及基音谱图。

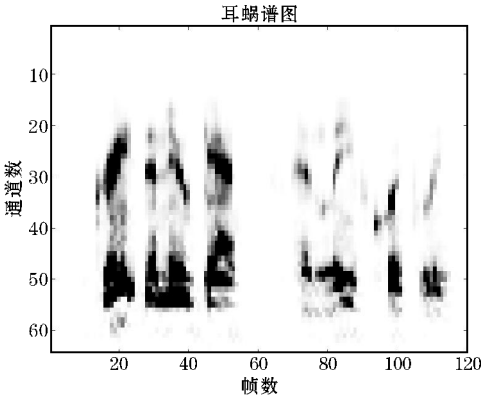

图4　纯净语音的耳蜗谱图

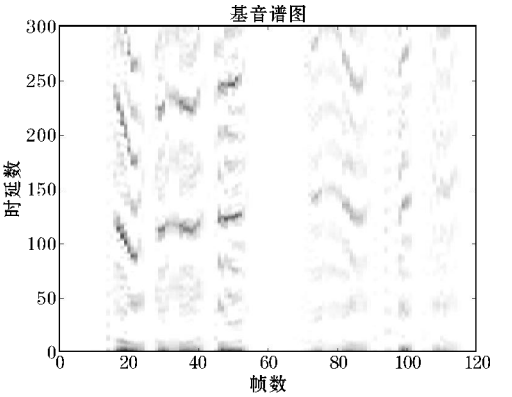

图5　纯净语音的基音谱图

随机选取一条纯净语音,在它的基础上添加噪声信号,信噪比分别为-5 dB、0 dB、5 dB。在不同信噪比下,混合信号与纯净语音信号的同一帧浊音的SAI图像对比如图6所示。

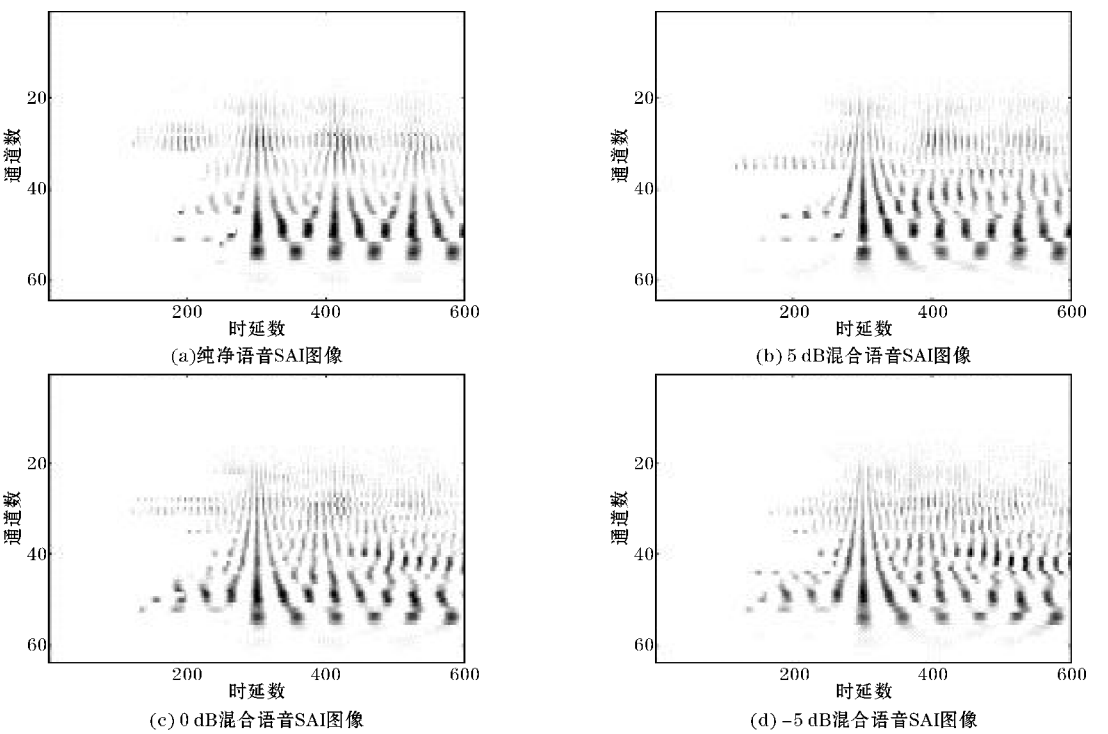

图6　不同信噪比下同一帧浊音SAI图像对比

由图6可得,纯净语音的SAI图像呈现周期性,而混合语音的听觉图像相比纯净语音图像有一定的偏移,不具有周期性,且信噪比越低,偏移越大,图像越混乱。

由于人体声带的结构,语音信号的浊音具有一定的周期性,而噪声信号属于无序信号,不存在周期性。因此,可以通过提取基音处的SAI数据拼接来获取语音的特征,实现一定的听觉场景分析,得到一种鲁棒性特征,并将其应用于语音增强系统中。

本文基于SAI图像提出3种听觉特征开展研究。一是由SAI图像行均值所得的耳蜗谱图作为特征;二是SAI图像经过PCA降维后作为特征;三是在SAI上进行场景分析,提取SAI图像基音处的数据作为特征。

2　语音增强系统

本文设计了一个基于监督性学习构架的语音增强系统,框架如图7所示。该系统的目标是对混合语音做增强处理。首先将混合语音通过耳蜗模型获取对应的特征参数,作为神经网络的输入,通过已训练好的网络来预测理想比率掩码(IRM)。再由IRM进行语音合成,得到干净的语音信号。

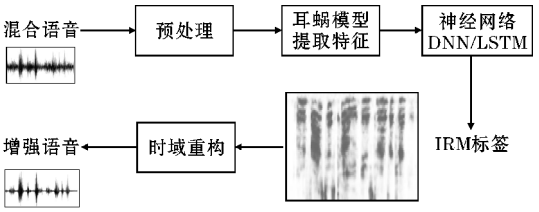

图 7 语音增强系统整体框图

2.1 预处理

预处理主要包括预加重、分帧加窗等步骤,具体实现如图 8 所示。

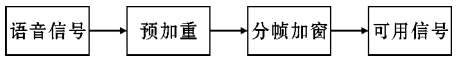

图 8 预处理框图

混合语音信号以 16 k 的频率进行采样,为补偿高频分量的损失,对其进行预加重处理。预加重滤波器设为

$$H(z)=1-az^{-1}$$

式中,a 是预加重系数,$0.9<a<1$。

语音信号是近似短时平稳的,在语音处理前将语音做分帧加窗处理。帧长一般取 10 ~ 30 ms,帧移和帧长的比值为0 ~ 0.5。窗函数一般具有低通特性,目前使用较多的是汉宁窗。

2.2 听觉特征

本文主要目的是基于 SAI 图像提取出具有鲁棒性的听觉特征,与目前常使用的听觉特征进行对比,分析不同语音分离系统的性能。

目前,常用的听觉特征包括:MFCC、RASTA-PLP、GFCC 以及 MRCG 等特征。其中,MFCC 是最普遍使用的声学特征。经过预处理后的信号通过 Mel 滤波器获得梅尔频率倒谱系数,最后经过离散余弦变换获得 MFCC。RASTA-PLP 是一种经过修正的线性预测频谱分析。GFCC 特征将预处理后的语音信号通过 64 通道的 Gammatone 滤波器组,再进行 DCT 变换得到,更加符合人类听觉模型。多分辨率耳蜗特征 MRCG[12] 将 4 种不同分辨率的耳蜗图结合。一个高分辨率的耳蜗捕捉局部信息,3 个低分辨率的耳蜗捕捉不同光谱的上下文。

基于 SAI 图像,本文考虑了 3 种提取特征的方式。

SFCC 特征。该特征通过 SAI 图像作行均值,获取语音的耳蜗谱图,再经过 DCT 倒谱运算,观察听觉模型作用下的特征表征能力。

SAI-PCA 特征。该特征将 SAI 图像直接进行 PCA 降维。SAI 图像是从脑干和中脑中提取并映射到听觉皮层的图像,模拟了听觉皮层前的人耳功能。对 SAI 图像通过 PCA 降维,提取每帧图像中最有代表性的信息作为特征。

SAI-SPE 特征。该特征是针对 SAI 图像进行声场景分析得到的一种特征。首先在基音谱图上获取语音信号的基音轮廓,在基音轮廓的基础上,提取 SAI 基音处和零延迟处的 SAI 图像,按帧拼接得到频谱信号,最后做 DCT 处理得到特征。

语音信号的浊音部分具有周期性,可提取基音;噪声信号不存在基音,因此可在基音处从混合信号中提取纯净信号的特征。本文采用 3 条规则在基音谱图中提取语音基音轮廓。

$$\text{pitch\_m}(i)>\text{pitch}_{\min}$$
$$\text{Corr}(i)>C_{\max}$$
$$\text{SNR}(i)>S_{\min}$$

式中,pitch_m 为基音谱图的幅值;$\text{pitch}_{\min}$表示浊音帧的基音处最低幅值;$\text{Corr}(i)$表示第帧语音 SAI 图像中心列和基音列的相关系数,当大于 $C_{\max}$时可认为是浊音帧;$\text{SNR}(i)$表示 SAI 零延迟处和基音处的能量比,当能量大于 $S_{\min}$时为浊音帧。经过 3 个门限的筛选,可以确定较准确的基音轮廓,不满足 3 个门限的帧表示非浊音帧,基音值置零。根据基音轮廓提取每帧 SAI 的基音列,浊音帧直接提取基音列,非浊音帧不进行提取,由此得到对应基音的频谱。基音处频谱和中心列频谱拼接,再进行 DCT 处理,得到 SAI_SPE 特征。

2.3 神经网络及掩模标签

采用 LSTM 循环神经网络,以理想比率掩码(IRM)作为对应的输出标签进行训练,实现语音的增强效果。

2.3.1 LSTM 循环神经网络

LSTM 循环神经网络[13]是一种具有记忆功能的网络。LSTM 在普通 RNN 的基础上进行改进,可以解决普通 RNN 网络梯度爆炸或梯度消失的问题。LSTM 结构如图 9 所示。

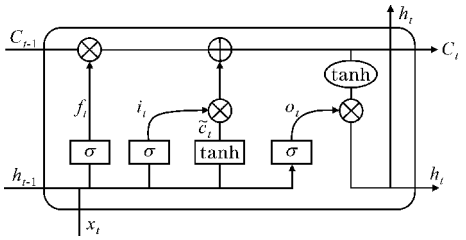

图 9 LSTM 结构图

LSTM 实现了长距离的信息记忆,主要通过遗忘门、输入门和输出门 3 种门结构来保护和控制细胞状态。

遗忘门根据上个单元的细胞状态决定丢弃哪些信息,保留重要信息,抛弃无关紧要的信息,计算公式如下:

$$f_t=\sigma(W_f\cdot[h_{t-1},x_t]+b_f)$$

输入门确定哪些新信息会被存放在细胞状态中,计算公式如下:

$$i_t=\sigma(W_i[h_{t-1},x_t]+b_i)$$

$$\tilde{C}_t=\tanh(W_c[h_{t-1},x_t]+b_c)$$

$$C_t=f_t\times C_{t-1}+i_t\times\tilde{C}_t$$

输出门输出细胞最新状态,计算公式如下:

$$O_t=\sigma(W_O[h_{t-1},x_t]+b_O)$$

$$h_t=O_t\times\tanh(C_t)$$

式中,h_{t-1}和C_{t-1}表示上一个时刻隐藏层输出和细胞状态;h_t和C_t表示当前时刻隐藏层输出和细胞状态;$\tilde{C}_t$表示细胞状态候选值;W和b表示权值偏置。通过3个门的控制,LSTM可以实现长短时记忆的功能,可以捕捉到长序列之间的语义关联,适用于处理语音信号。

2.3.2　掩码标签

常见的掩码标签有理想二值掩码(IBM)[14]和理想比率掩码(IRM)。理想二值掩蔽是指对每一个时频单元,如果其局部信噪比大于某一阈值,对应的掩蔽矩阵标记为1,否则标记为0。

理想二值掩蔽将语音分离问题转化为一个二元分类问题,有一定的局限性。因此本文采用理想比率掩模作为分离目标。在时频谱中,设$Y(t,f)$、$S(t,f)$、$N(f)$分别为混合语音信号$y(t)$、纯净语音信号$s(t)$、噪声信号$n(t)$在时间帧t和频率f的时频表示。

$$\mathrm{IRM}(t,f)=\left(\frac{S^2(t,f)}{S^2(t,f)+N^2(t,f)}\right)^{\beta}$$

式中,$S^2(t,f)$和$N^2(t,f)$分别定义了混合语音中时间帧为t,和频率f的时频单元的语音和噪音的能量,β是一个可调节的尺度因子,其最佳值为0.5[15]。

语音分离系统就是将各种听觉特征作为神经网络的输入,以IRM作为输出标签,通过训练后得到稳定的网络结构,由此可得到预测的Gammatone域的IRM。计算$Y(t,f)\times\mathrm{IRM}(t,f)$,对有噪信号的Gammatone谱图进行增强,然后增强后的语音进行时域重构,得到时域增强语音。

3　实验结果及分析

3.1　SAI场景分析结果

SAI图像的列均值经过时间堆叠可得到基音谱图,在基音谱图的基础上提取基音轮廓。实验中,$\mathrm{pitch}_{\min}=14.25$、$C_{\max}=0.54$、$S_{\min}=0.38$来获取混合语音的基音轮廓。5 dB混合语音的基音轮廓如图10所示,可较为准确地提取纯净语音的基音轮廓。

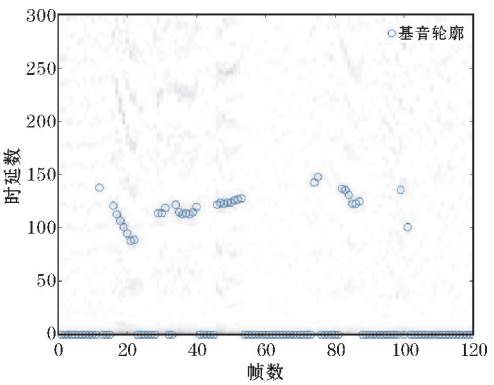

图10　5 dB混合语音的基音轮廓图

随机选取一条纯净语音,在它的基础上添加噪声信号,信噪比分别为-5 dB、0 dB、5 dB。不同信噪比下,SAI图像在中心处和基音处提取的频谱信号如图11和图12所示。

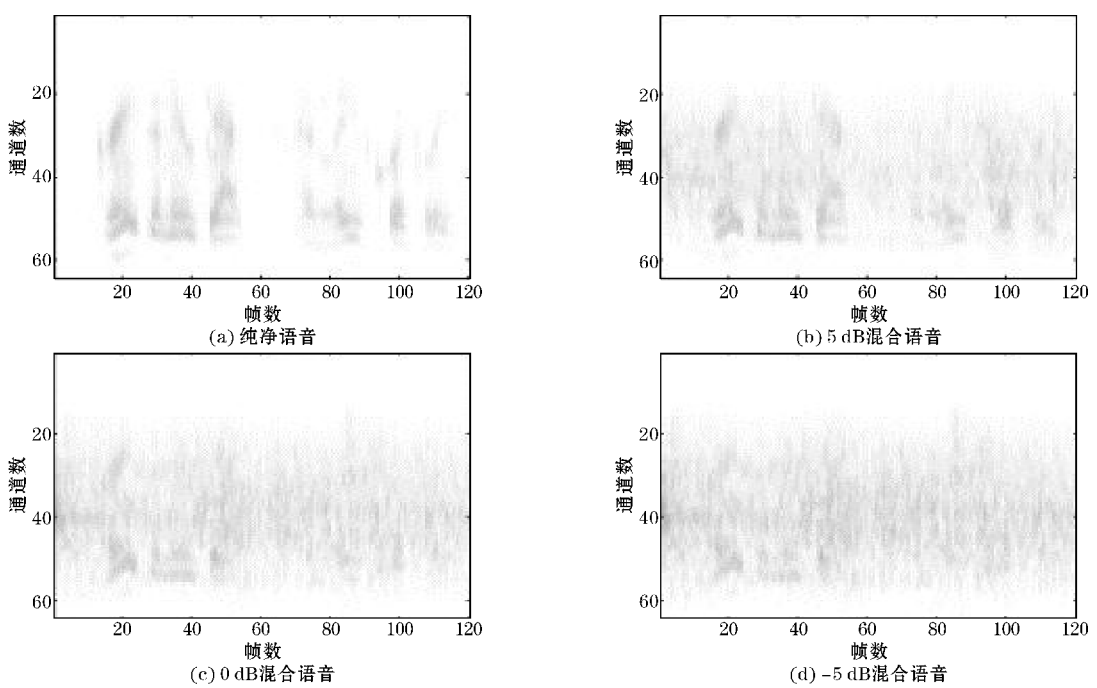

图11　SAI中心处提取频谱图像

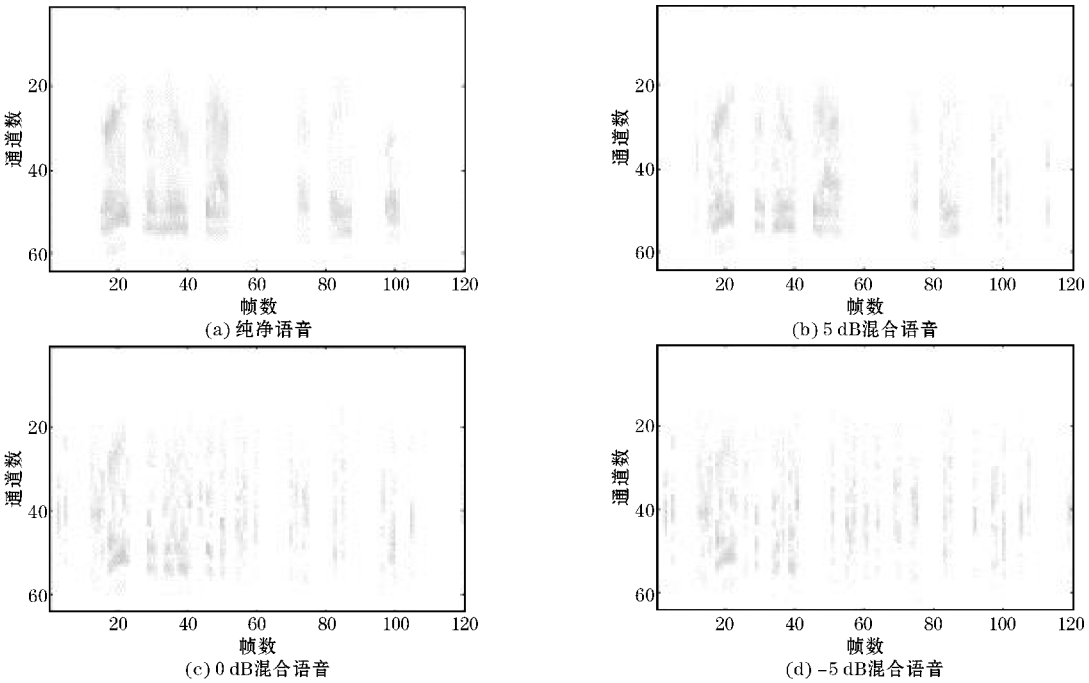

图 12 SAI 基音处提取频谱图像

将 SAI 基音处和中心处提取的频谱特征与纯净语音的频谱特征进行比较,观察不同信噪比下,基音处频谱的降噪效果。具体数据如表 1 所示。

表 1 SAI 提取频谱与纯净语音频谱相似度

信噪比/dB	中心处相似度	基音处相似度
-5	0.6702	0.7805
0	0.7071	0.9035
5	0.8164	0.9875

从表 1 可以看出,SAI 基音处频谱的提取有一定降噪效果。在0 dB和5 dB混合信号中,基音处频谱与纯净语音相似度比中心处频谱的相似度提高了约 15% 以上,由于-5 dB信噪比过低,所以相似度的提升不是那么明显。

3.2 语音增强实验

实验采用 IEEE Corpus 语音数据集,随机选取 300 条语音作为训练集,60 条语音作为测试集,选用 Noise92 噪声库中的工厂噪声作为噪声信号。将纯净语音和噪声按-5 dB、0 dB、5 dB的信噪比混合。

首先对混合信号进行预处理,分别提取混合语音的 MFCC、GFCC、RASTA-PLP、MRCG、张涛等[16] 提出的自编码特征 IF、SFCC、SAI-PCA、SAI-SPE 等特征,将特征输入到神经网络中训练。此处采用双层 LSTM,节点数都为 2048,为防止模型过拟合,将 Dropout 值设为0.2,学习率设置为0.005,采用 Adam 方法来进行模型的训练。将相邻两帧的数据与本帧数据融合,共 5 帧特征数据作为输入,输出为该帧对应的 IRM 掩码。实验中采用 4 个指标作为语音增强的评价指标,STOI、PESQ、SNR 以及输出与理想比率掩模的相似度 R。4 个指标都是数值越高,语音增强效果越好。

各种特征作为输入的语音增强各指标输出如表 2 所示。其中△表示各指标的变化量。

表 2 LSTM 网络下各特征的语音增强指标输出

特征	△STOI			△PESQ			△SNR			R		
	-5 dB	0 dB	5 dB	-5 dB	0 dB	5 dB	-5 dB	0 dB	5 dB	-5 dB	0 dB	5 dB
MFCC	0.031	0.061	0.069	0.406	0.333	0.567	8.009	5.984	5.049	58.51	73.67	82.71
GFCC	0.059	0.091	0.075	0.386	0.550	0.545	8.055	6.457	5.025	69.48	81.26	83.52
RASPLP	0.016	0.068	0.047	0.385	0.479	0.549	8.272	6.205	4.994	63.52	77.35	81.51
MRCG	0.080	0.073	0.068	0.511	0.475	0.539	8.273	6.331	4.839	71.79	77.63	82.96
IF	0.087	0.092	0.065	0.520	**0.559**	0.549	8.427	6.504	4.913	72.81	80.34	83.16
SFCC	0.090	0.089	0.068	0.526	0.503	0.555	8.358	6.549	4.971	72.64	80.81	83.04
SAI-PCA	0.085	0.083	0.068	0.527	0.519	0.545	8.256	6.459	4.924	71.65	79.66	82.94
SAI-SPE	**0.095**	**0.100**	**0.080**	**0.528**	0.554	**0.581**	**8.663**	**6.733**	**5.124**	**75.37**	**82.81**	**84.89**

由表 2 可见,SFCC 特征相比 MFCC、GFCC 等特征有一定的性能提升,可以表现皮层前模型的有效性,但相比 MRCG 特征、IF 特征优势并不明显。SAI-PCA 特征只进行降维操作,实验性能不如 SFCC,但依旧高于其他常用的听觉特征。而 SAI-SPE 特征,在听觉模型后加入了听觉场景分析,表现出更好的性能。在0 dB下,SAI-SPE 比起 MFCC 特征有较大的提升,STOI 提升了约0.04、PESQ 提升了约0.2、SNR 提升了约1 dB、相似度 R 提升了约 8%。5 dB信噪比下,STOI 提升了约0.02,SNR 提升了约0.8 dB。数据表明,在低信噪比的情况下,SFCC 和 SAI-SPE 特征可以表现更好的语音增强结果。此次实验验证了将人类听觉模型、听觉场景分析与神经网络相结合的好处,人耳听觉模型特征可表征鲁棒性,应用于语音处理系统中,可以表现更好的性能。

4　结束语

在较完备的听觉模型的基础上进行声场景分析,得到鲁棒性语音特征并将其应用到语音增强系统中。其中 CARFAC 作为耳蜗外周模型,SAI 作为从耳蜗到脑干的神经传输模型构成皮层前的听觉模型。结合场景分析,将 SAI 图像基音处的频谱作为特征,应用于语音增强系统中,在低信噪比的情况下,相比其他常用特征的增强效果有一定的提升。实验结果表明,将人耳听觉模型应用于机器听觉系统中,可表现一定的鲁棒性。下一步的研究可针对人耳听觉模型的简化以及研究更贴近人耳功能的特征。

参考文献:

[1] Das N, Chakraborty S, Chaki J, et al. Fundamentals, present and future perspectives of speech enhancement[J]. International Journal of Speech Technology, 2021, 24: 883-901.

[2] Computational Auditory Scene Analysis: Proceedings of the Ijcai-95 Workshop[M]. CRC press, 2021.

[3] Wang D L, Chen J. Supervised speech separation based on deep learning: An overview[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 26(10): 1702-1726.

[4] 孙林慧,王灿,梁文清,等. 基于深度学习特征融合和联合约束的单通道语音分离方法[J]. 电子与信息学报,2022,44(9):1-11.

[5] Chen J, Wang D L. Long short-term memory for speaker generalization in supervised speech separation[J]. The Journal of the Acoustical Society of America, 2017, 141(6): 4705-4714.

[6] Xu Y, Afshar S, Singh R K, et al. A binaural sound localization system using deep convolutional neural networks[C]. 2019 IEEE International Symposium on Circuits and Systems (ISCAS) IEEE, 2019: 1-5.

[7] Lyon R F, Ponte J, Chechik G. Sparse coding of auditory features for machine hearing in interference[C]. 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) IEEE, 2011: 5876-5879.

[8] Lyon R F. Human and machine hearing: extracting meaning from sound[M]. Cambridge University Press, 2017.

[9] Virtanen T, Plumbley M D, Ellis D. Computational analysis of sound scenes and events[M]. Cham: Springer International Publishing, 2018.

[10] Saremi A, Beutelmann R, Dietz M, et al. A comparative study of seven human cochlear filter models[J]. The Journal of the Acoustical Society of America, 2016, 140(3): 1618-1634.

[11] Islam M A, Xu Y, Monk T, et al. Noise-robust text-dependent speaker identification using cochlear models[J]. The Journal of the Acoustical Society of America, 2022, 151(1): 500-516.

[12] Peng Z, Dang J, Unoki M, et al. Multi-resolution modulation-filtered cochleagram feature for LSTM-based dimensional emotion recognition from speech[J]. Neural Networks, 2021, 140: 261-273.

[13] Yu Y, Si X, Hu C, et al. A review of recurrent neural networks: LSTM cells and network architectures[J]. Neural computation, 2019, 31(7): 1235-1270.

[14] Kolbæk M, Tan Z H, 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.

[15] Wang Y, Narayanan A, Wang D L. On training targets for supervised speech separation[J].

IEEE/ACM transactions on audio, speech, and language processing, 2014, 22(12): 1849-1858.

[16] 张涛, 任相赢, 刘阳, 等. 基于自编码特征的语音增强声学特征提取[J]. 计算机科学与探索, 2019, 13(8): 1341.

Research and Application of Robust Characteristics of Auditory Models

WANG Wenhua, XIA Xiuyu

(School of Electronic Information, Sichuan University, Chengdu 610064, China)

Abstract: The human auditory system has a very fine and ingenious structure, and it can accurately understand speech even in a noisy environment. Using a fine cochlea model as front-end processing allows for better speech processing. In this paper, a rapidly compressed asymmetric resonator cascade (CARFAC) is used as a peripheral model of the human ear, combined with an auditory stabilization image (SAI) to obtain an accurate precortical auditory model. Based on the auditory model, a more accurate pitch contour is extracted, the pitch information is used to analyze the acoustic scene, and robust speech features are synthesized, which are sent to the neural network for supervised training to achieve speech enhancement. Experiments show that under noise conditions, the features extracted by the auditory model are better reflected in various speech evaluation indicators, which can better characterize the speech signal and have certain robustness.

Keywords: CARFAC model; auditory stabilization image; speech enhancement system; pitch extraction