基于时频单元选择的双耳目标声源定位

李如玮*① 李 涛① 孙晓月① 杨登才② 王 琪①

①(北京工业大学信息学部人工智能研究院和信息与通信工程学院 北京 100124)

②(北京工业大学科技发展研究院 北京 100124)

摘 要：针对复杂声学环境下，现有目标声源定位算法精度低的问题，该文提出了一种基于时频单元选择的双耳目标声源定位算法。该算法首先利用双耳目标声源的频谱特征训练1个基于深度学习的时频单元选择模型，然后使用时频单元选择器从双耳输入信号中提取可靠的时频单元，减少非目标时频单元对定位精度的负面影响。同时，基于深度神经网络的定位系统将双耳空间线索映射到方位角的后验概率。最后，依据与可靠时频单元相对应的后验概率完成目标语音的声源定位。实验结果表明，该算法在低信噪比和各种混响环境，特别是存在与目标声源类似的噪声环境下目标声源的定位精度得到明显改善，性能优于对比算法。

关键词：目标声源定位；深度学习；时频单元选择

中图分类号：TN912.3 **文献标识码**：A **文章编号**：1009-5896(2019)12-2932-07

DOI: 10.11999/JEIT181127

Binaural Target Sound Source Localization Based on Time-frequency Units Selection

LI Ruwei① LI Tao① SUN Xiaoyue① YANG Dengcai② WANG Qi①

①(*Laboratory of Speech and Audio Signal Processing and Institute of Artificial Intelligence, Faculty of Information Technology, Beijing University of Technology, Beijing* 100124, *China*)

②(*Institute of Science and Technology Development, Beijing University of Technology, Beijing* 100124, *China*)

Abstract: The performance of the existing target localization algorithms is not ideal in complex acoustic environment. In order to improve this problem, a novel target binaural sound localization algorithm is presented. First, the algorithm uses binaural spectral features as input of a time-frequency units selector based on deep learning. Then, to reduce the negative impact of the time-frequency unit belonging to noise on the localization accuracy, the selector is emploied to select the reliable time-frequency units from binaural input sound signal. At the same time, a Deep Neural Network (DNN)-based localization system maps the binaural cues of each time-frequency unit to the azimuth posterior probability. Finally, the target localization is completed according to the azimuth posterior probability belonging to the reliable time-frequency units. Experimental results show that the performance of the proposed algorithm is better than comparison algorithms and achieves a significant improvement in target localization accuracy in low Signal-to-Noise Ratio(SNR) and various reverberation environments, especially when there is noise similar to the target sound source.

Key words: Target sound localization; Deep learning; Time-frequency units selection

1 引言

双耳的目标声源定位具有较大的潜在应用价值，它可应用于诸如双耳数字助听器、语音识别和空间音频重建等领域，是计算听觉场景分析和信号处理中的关键技术。在复杂的声学环境下如鸡尾酒会环境等，人类对于目标声源的定位几乎没有困难，人们可以十分轻松地解混到达每只耳朵的混合声音，并找到感兴趣声源的位置。但是在低信噪比和混响环境中，尽管研究者们在声源定位方面进行了许多研究，现有的目标声源定位算法的定位精确度与人类的听觉系统仍存在着较大的差别。

心里声学的研究指出，在自由声场下，对声源的方向定位因素包括双耳时间差(Interaural Time

收稿日期：2018-12-06；改回日期：2019-05-21；网络出版：2019-06-04

*通信作者： 李如玮 liruwei@bjut.edu.cn

基金项目：国家自然科学基金(51477028)，北京市教委科技计划面上项目(KM201510005007)

Foundation Items: The National Natural Science Foundation of China(51477028), The Scientific Research Program of Beijing Municipal Commission of Education (KM201510005007)

Differences, ITD)、双耳声级差(Interaural Level Differences, ILD)等。许多研究人员也已经提出了基于双耳时间差、广义互相关函数(Cross-Correlation Function, CCF)和双耳声级差的声源定位算法[1–4]。这些算法都是将双耳信号首先经过听觉滤波器，如Gammatone滤波器等，将双耳信号分解为若干个子带，在每个子带中提取ITD或者CCF和ILD作为声源定位的特征，文献[2]中，根据Gammatone子带能量的不同，减少ITD和ILD的维度，最后使用高斯混合模型(Gaussian Mixture Model, GMM)定位。文献[4]中，提取每个时频单元的CCF和ILD，随后使用深度学习模型将CCF和ILD映射为每个时频单元所属方位角的后验概率，最后，每个时频单元所属方位角的后验概率在频率上相乘，在时间上平滑后选取最大后验概率所对应的方位角，便是声源的方向。但是在低信噪比下，干扰源的能量将大于目标语音的能量，干扰源的空间信息对定位的贡献远远大于目标声源导致定位精度低。为了解决该问题，文献[5]提出了结合噪声源先验信息模型的声源定位方法，该方法使用GMM分别对噪声信号和语音信号进行建模，得出每个时频单元属于语音信号的概率作为权重，利用该权重对每个时频单元所属方位角的后验概率进行指数加权。在低信噪比和具有类似语音的频谱的噪声的情况下，由GMM模型估计的权重并不十分稳健，即会产生将属于噪声的时频单元的定位信息用于目标声源定位的情况。因此这种结合已知噪声模型的概率加权定位方法，在恶劣的声学环境下并不能消除噪声对目标声源定位的影响。

分析可知语音信号具有时频稀疏性和短时正交特性(W-Disjoint Orthogonality, W-DO)，每个时频单元仅有1个主导的声源，即该时频点属于目标语音信号或者噪声信号，属于目标语音的时频单元即为可靠的时频单元。为此，依据以上分析，本文提出了一种基于时频单元选择的双耳目标声源定位算法。该算法首先利用Gammatone滤波器对信号进行分解，然后使用深度神经网络(Deep Neural Network, DNN)将每个时频单元的双耳空间特征CCF和ILD作为输入，预测每个时频单元方位角的后验概率分布。同时，使用二进制掩码将属于语音信号的时频单元标记为1，其他的噪声标记为0，并采用一个隐藏层为2层的DNN对每个时频单元进行分类，并选择属于目标声源的时频单元。最后，根据可靠时频单元选择相对应的后验概率分布完成目标语音的声源定位。实验结果表明，在复杂的声学条件下与对比算法相比，本文所提出的目标声源定位算法具有低信噪比下更强的鲁棒性以及更高的定位准确度。

2　双耳目标声源定位算法

本文所提出的目标声源定位系统如图1所示。系统的双耳输入信号由特定方向的目标语音与目标语音方向不同的噪声信号在混响环境中混合而成。首先，双耳输入信号先经Gammatone滤波器进行频域分析，提取双耳的空间特征作为定位模型的输入，以获得每个时频单元所属方位角的后验概率分布。然后提取频谱特征作为时频单元选择DNN的输入以获得可靠时频单元(图1中阴影区域)的索引。最后，从全部后验概率分布中选择出属于目标语音时频单元的后验概率，即使用可靠时频单元提供的目标语音方向信息获得的后验概率完成目标语音的声源定位。

2.1　Gammatone滤波器组分解

Gammatone滤波器可以模拟人耳基底膜的特

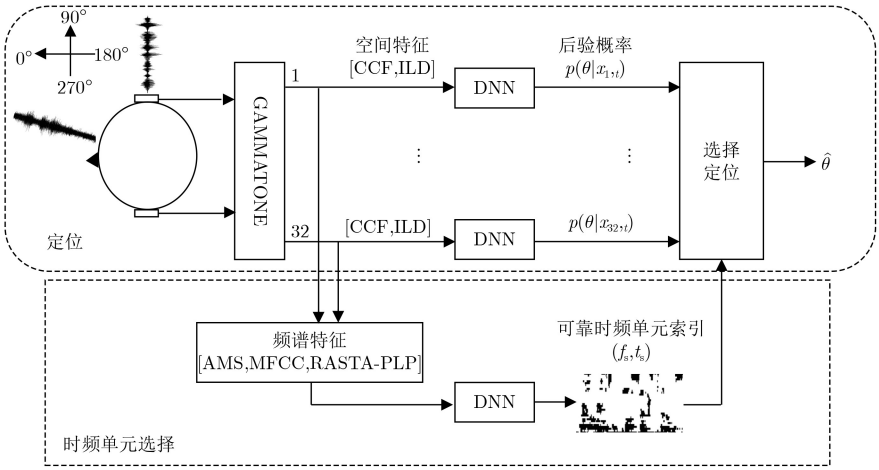

图 1 本文算法原理框图

性并符合人耳的听觉感知特性，本文使用Gammatone滤波器组对输入语音信号进行时频分析。通过使用中心频率为80～8000 Hz的32通道的Gammatone滤波器组，将每个通道的输入信号分解为32个子带信号。随后，每个通道共32个子带信号使用汉宁窗进行分帧加窗，帧长取20 ms，帧移取10 ms，得到双耳输入信号的时频域表示$x_{ft,\mathrm{l}}(n)$, $x_{ft,\mathrm{r}}(n)$，其中l, r为左右耳的表示，t为时间帧的索引，f为频率的索引。以下特征都在使用Gammatone滤波器组时频分析后的基础上提取。

2.2 特征提取

2.2.1 双耳空间特征

双耳的空间特征包括ITD和ILD是影响对声源方向定位精确度的重要因素。ITD和ILD同时也是一对互补的特征，因为ITD是低频定位(<1.5 kHz)的主要特征，ILD是高频定位(≥1.5 kHz)的主要特征。左右耳之间的归一化互相关函数(CCF)是计算ITD的主要方法，每个时频单元的CCF的计算如式(1)所示。

$$\mathrm{CCF}(t,f,\tau)=\frac{\sum\limits_i x_{ft,\mathrm{l}}(i)x_{ft,\mathrm{r}}(i-\tau)}{\sqrt{\sum\limits_i x_{ft,\mathrm{l}}^2(i)}\sqrt{\sum\limits_i x_{ft,\mathrm{r}}^2(i-\tau)}} \tag{1}$$

式中，i为样点数的索引，τ为延时，范围取–1～1 ms，且在16 kHz的采样率下，每个时频单元的CCF的维度为33维。但是对于较复杂恶劣的声学环境下计算ITD时的峰值拾取往往不够稳健[4]，因此本文使用CCF来代替只含有互相关最大值信息的ITD。

当声源偏离双耳连线的中垂面时，由于头部对声波的阴影和散色作用，特别在高频，与声源异侧耳处的声压受到衰减，而与声源同侧的声压有一定的提升，因而形成与声源方向有关的双耳声级差

$$\mathrm{ILD}(t,f)=10\lg\frac{\sum\limits_i x_{ft,\mathrm{l}}^2(i)}{\sum\limits_i x_{ft,\mathrm{r}}^2(i)} \tag{2}$$

将33维的CCF与1维的ILD结合成一个34维特征用作声源定位模型的输入。

2.2.2 频谱特征

在对属于语音信号的时频单元选取过程中需要对每个时频单元的所属做出判断，因此需要为每个目标时频单元选取合适的特征表示。在语音识别和语音增强领域，已经有许多优秀的特征被提取出来。幅度调制谱(Amplitude Modulation Spectrogram, AMS)同时有着清音特征和浊音特征已被成功应用于语音分离中, 但是AMS的泛化能力较差。结合相对谱变换(Relative Spectral Transform, RASTA)的感知线性预测(Perceptual Linear Prediction, PLP)是一种基于听觉模型的语音特征参数，它模拟了人的听觉感知机理且具有对噪声的鲁棒性，适用于噪声不匹配环境下的语音分离。梅尔倒谱系数(Mel Frequency Cepstrum Coefficient, MFCC)是基于人的听觉机理的，根据人的听觉实验结果来分析语音的频谱，也是增强和音频重建等常用的语音特征。针对各个特征含有不同的信息，Wang等人[6]利用group LASSO的特征选择算法得到了最优的特征组合AMS+RASTA-PLP+MFCC，这个互补的特征组合在各种声学环境下取得了很好的效果，显著优于单个特征的分离算法。

为了更可靠地从混合语音中选取属于语音的时频单元，本文先将两耳的幅度谱进行平均，随后提取了包括MFCC, AMS和RASTA-PLP的1组互补特征集。并使用自回归移动平均滤波器(Auto Regressive Moving Average, ARMA)来对邻近帧的特征进行平滑后作为时频单元选择模型的输入[7,8]。

2.3 语音时频单元的选择

2.3.1 时频单元的分类

为了精确地定位目标语音，需要选取属于目标语音的时频单元，并使用目标语音时频单元提供目标语音的空间信息，排除噪声时频单元的干扰信息，完成精确的目标声源定位。选取属于目标语音的时频单元的过程可以看作是对每个时频单元先分类再选取的过程，即首先将属于语音的时频单元标记为1，属于其他干扰源的时频单元标记为0，随后选取为1的时频单元用作目标声源的定位。对于分类过程，本文选用理想二值掩码(Ideal Binary Mask, IBM)作为训练目标训练深度神经网络[9]。理想二值掩码的定义如式(3)所示

$$\boldsymbol{M}(t,f)=\begin{cases}1, & \mathrm{SNR}(t,f)>\mathrm{LC}\\ 0, & \text{其他}\end{cases} \tag{3}$$

在式(3)中，$\mathrm{SNR}(t,f)$为某个时频单元的局部信噪比。在语音分离任务中，需要选取合适阈值LC，以免噪声去除过度导致语音舒适度不足。但是在这里，最主要的是保证选取的时频单元的可靠性，即选取的是属于语音信号的时频单元。随着所取阈值的提高，可选取的时频单元数量减少鲁棒性减弱；随着所取阈值的降低，导致残余噪声时频单元不利于目标声源的定位。因此本文设置阈值LC = 0 dB，语音能量刚刚大于噪声能量，此时既可以保证选取由目标语音主导的时频单元，也可以保证选取的可靠时频点的数量较多。

2.3.2 可靠时频单元的选择

由于深度学习在数据学习方面表现出很大的优势。利用深度学习强大的学习能力，从含噪语音频谱中估计语音频谱，以获得可靠语音时频单元的方法是目前最先进的方法之一。

本文搭建1个包含2个隐藏层的DNN，其中每层有1000个神经单元。采用线性整流函数(Rectified Linear Unit, ReLU)作为深度神经网络的隐藏层的激活函数，在隐藏层与输出层之间使用sigmoid函数作为激活函数。使用最小均方误差函数作为损失函数，自适应梯度算法用来最小化损失函数。为了避免出现过拟合，对隐藏层中神经单元采用丢弃法，丢弃率选取经验值0.5，即此时随机生成的网络结构最多，更好地避免过拟合。以混合语音的MFCC, AMS和RASTA-PLP频谱特征作为输入，预测包含每个时频单元所属类别的二进制掩码矩阵$\hat{\boldsymbol{M}}(t,f)$，并依此选择属于语音的时频单元用作目标声源定位的依据。

2.4 目标声源定位

虽然目标语音与噪声在时间上重叠，但是在局部时间里的每个频带中主要由单个声源主导，这允许仅使用单声源数据来训练定位的深度模型。为了证明时频单元选择对目标声源的定位的贡献是有效的，本文使用最先进的声源定位系统[4]作为时频单元选择的承载。

针对32个频带中的每一个频率训练单独的DNN，每个DNN由输入层、2个隐藏层和输出层组成。每个隐藏层具有128个隐藏单元，采用sigmoid激活函数，输出层为具有72个结点，采用softmax激活函数。本文把语音的双耳空间线索CCF和ILD作为输入，输出层得到时频单元所属角度的后验概率$p(\theta|\boldsymbol{v}_{f,t})$，其中$\boldsymbol{v}_{f,t}=[\mathrm{CCF}(t,f)\mathrm{ILD}(t,f)]$，即34维双耳空间特征向量。随后将每个时频单元的后验概率在频率通道上整合，得到每一帧语音所属方位角的概率，如式(4)所示

$$p(\theta|\boldsymbol{v}_t)=\frac{\prod\limits_f p(\theta|\boldsymbol{v}_{f,t})}{\sum\limits_\theta\prod\limits_f p(\theta|\boldsymbol{v}_{f,t})} \tag{4}$$

其中$\boldsymbol{v}_t=[v_{1,t}\ v_{2,t}\cdots v_{32,t}]$，即特征在时间$t$和全部频率范围上的表示。这里所有的时频单元的信息都参与声源定位，这对于高信噪比下是精确的，但是在低信噪比下却是不可靠的，因为此时噪声的时频单元数量和能量都占优，噪声提供更多的方位角信息，导致由目标语音时频单元所提供的信息被掩盖，目标语音的定位不够精确。因此引入可靠时频单元进行目标声源的定位，时频单元选择如式(5)所示

$$(t_\mathrm{s},f_\mathrm{s})=\mathrm{find}(\hat{\boldsymbol{M}}(t,f)==1) \tag{5}$$

式中，表示选取二进制矩阵$\hat{\boldsymbol{M}}(t,f)$中元素值为1的索引值即为可靠时频单元的索引。find(·)表示选择索引的函数，$(t_\mathrm{s},f_\mathrm{s})$为选择出的可靠时频单元的索引。于是，使用可靠时频单元的定位可定义为式(6)

$$p(\theta|\boldsymbol{v}_{t_\mathrm{s}})=\frac{\prod\limits_{f_\mathrm{s}} p\left(\theta|\boldsymbol{v}_{f_\mathrm{s},t_\mathrm{s}}\right)}{\sum\limits_\theta\prod\limits_{f_\mathrm{s}} p\left(\theta|\boldsymbol{v}_{f_\mathrm{s},t_\mathrm{s}}\right)} \tag{6}$$

式(6)表示从承载系统获得的所有时频单元的后验概率中选择出基于目标语音时频单元的后验概率。之后在每帧上求取了方位角的概率分布$p\left(\theta|\boldsymbol{v}_{t_\mathrm{s}}\right)$，然后在输入信号的所有帧间求平均，可得输入信号方位角的概率分布如式(7)所示

$$p\left(\theta\right)=\frac{1}{T}\sum_{t_\mathrm{s}}^{t_\mathrm{s}+T-1} p\left(\theta|\boldsymbol{v}_{t_\mathrm{s}}\right) \tag{7}$$

其中，T为帧总数，使后验概率最大的θ便是目标的方位角，如式(8)所示

$$\hat{\theta}=\arg\max_\theta(p(\theta)) \tag{8}$$

3 实验设置与结果分析

3.1 双耳信号的生成

通过将单通道语音与头相关冲缴响应(Head Related Impulse Responses, HRIR)或者双耳房间冲缴响应(Binaural Room Impulse Response, BRIR)进行卷积生成双耳信号。在训练定位模型时，使用基于KEMAR人工头部的头相关冲缴函数生成无混响的单源双耳信号用作训练数据，HRIR数据集包括0～360°以5°为间隔的共72个冲缴响应[10]。在测试时双耳信号生成的数据取自Surrey BRIR数据库[11]，Surrey BRIR数据库是使用头部和躯干模拟器(HATS)在房间A, B, C, D 4个大小不同的房间中测量的，房间的特性参数如表1所示。声源放置于以躯干模拟器为圆心，半径为1.5 m的前半圆弧上，即声源与人耳的距离为1.5 m。每个房间数据集内包括–90～90°以5°为间隔的共37个冲激响应。单通道语音与双耳房间冲激响应卷积生成带有混响的双耳信号用来模拟真实的房间混响环境。

表 1 房间特性参数

房间	A	B	C	D
T60(s)	0.32	0.47	0.68	0.89
DRR(dB)	6.09	5.31	8.82	6.12

3.2 模型的训练设置

对于定位模型的训练，对于每个方位角从TIMIT数据集中随机选取30句语音信号作为为原始的单通道语音信号。随机选取的单通道语音信号与其对应的方位角的HRIR卷积生成双耳信号。最后，从训练信号中提取双耳特征用于声源定位深度神经网络的输入。

对于选择时频单元模型的训练，用于生成训练集的目标语音来自于TIMIT数据集中为每个方位角随机选取的32句语音信号，它们分别于对应房间A中方位角的BRIR卷积生成带混响的目标语音信号。本文使用来自于文献[12]的部分噪声和选取自TIMIT的32条语音混合生成的babble噪声以及白噪声(如表2所示)，随机截取与目标语音相同长度的噪声信号，与不同于目标语音方位角对应的房间A的BRIR卷积生成混响中的噪声，最后与目标语音混合生成训练集，剩余3个房间混响环境不参与训练。

表 2 噪声类型与描述

噪声	噪声描述
symphony	弦乐器的声音，频率范围分布较广且与目标语音频率范围重叠
baby	与目标语音相比有更高的共振峰，且频率范围重叠
babble	随机选取于TIMIT数据集的32条语音混合形成，与目标语音频谱类似
alarm	信号能量大部分集中于2 kHz左右的窄带噪声
telephone	窄带噪声，能量集中于1 kHz与2 kHz附近
white	白噪声，能量在整个频带内均匀分布

3.3 测试

测试集的目标语音来自GRID数据集[13]的50条语音，测试集的语音信号未参与训练。生成双耳信号作为目标源后与双耳噪声信号相加生成含噪语音，信噪比设置为6 dB, 0 dB, –6 dB和–12 dB用于测试算法的有效性。目标源的方位角的变化范围在–90°～90°之间以步长为10°变化。噪声的方位角与目标源的方位角之间的角距离≥10°。

为了验证算法的有效性，本文选择最先进的基于DNN的声源定位系统[4]作为对比算法1，该算法使用没有任何关于时频单元的信息。结合频谱特征建模的目标声源定位算法[5]作为对比算法2。在本文所提算法和参考算法中，对数据集的预处理方法都相同。本文通过双耳声源定位精度作为指标表示3种算法的优劣，其中精度由声源的正确估计方位角所占的百分比表示，如果估计的源空间方向偏离小于5°的真实方向，则认为是正确的，如式(9)所示

$$\mathrm{Acc}=\frac{N_{\mathrm{dist}\left(\phi,\hat{\phi}\right)\le 5^\circ}}{N} \tag{9}$$

其中，N表示测试集中所有测试样例的总数，$N_{\mathrm{dist}\left(\phi,\hat{\phi}\right)\le 5^\circ}$表示估计方位角与真实方位的角距离小于5°的测试样例总数，即算法正确估计的测试样例的总数。角距离的计算如式(10)所示

$$\mathrm{dist}\left(\phi,\hat{\phi}\right)=180^\circ-\left||\phi-\hat{\phi}|-180^\circ\right| \tag{10}$$

3.4 实验结果与分析

表3显示了在6 dB, 0 dB, –6 dB以及–12 dB下各种混响条件和噪声条件下本文算法和两种对比算法的结果统计。

首先从表3中可以看出没有利用任何关于时频单元信息的对比算法1受噪声和混响的影响较大，目标声源定位准确度最差，因为由于信噪比低时噪声能量占优，噪声时频单元数量占优，定位精度受此影响较大，随着信噪比的降低，目标声源定位精度大幅降低。另外，混响环境下也会破坏双耳空间线索，不能提供足够的方位角信息。对比算法2和本文算法，在低信噪比和混响情况下比对比算法1的定位精度都有显著的性能提升，这说明着基于目标语音时频单元的定位是十分有效的。

当干扰噪声为babble时，对比算法2目标定位的精度较差，因为噪声信号的频谱与目标语音的频谱过于相似以及与目标语音在频谱能量的分布区间上的重叠。虽然对比算法2利用了GMM来建模频谱特征改善了定位性能，但是由于噪声频谱的类语音性质，导致效果并不理想。在baby与symphony噪声环境中也存在类似情形，虽然在较高信噪比下不如babble噪声下明显，但是随着信噪比降低这种性质依旧为目标声源定位产生较大的负面影响。综合对比可以看出本文提出的算法在这3种噪声环境下平均精度提升在6%左右，在低信噪比(–6 dB和–12 dB)下平均精度提升在9%左右，这改善了定位的精度特别是在更低的信噪比下，这显现出了本文算法的鲁棒性和可靠性。算法显现的优越性在于首先是选取1组互补的特征更好地表达了目标语音，这对于时频单元的分类选择是十分重要的，另外深度神经网络强大分类能力也是性能提升的另一个重要原因。

与此类噪声不同的是能量集中在2 kHz以上的窄带噪声，在如telephone与alarm的噪声条件下，本文算法与对比算法2取得了类似的定位效果，且定位精度都较高，这是因为在窄带噪声环境下都可以有效地利用频谱特征应用于目标声源的定位，而两者的误差都来自于最初提取的双耳空间特征，如前面所说，ILD在频率大于1.5 Hz时效果占优，而

表 3　目标声源定位精度

噪声	信噪比(dB)	房间A(%)			房间B(%)			房间C(%)			房间D(%)		
		对比算法1	对比算法2	本文算法	对比算法1	对比算法2	本文算法	对比算法1	对比算法2	本文算法	对比算法1	对比算法2	本文算法
symphony	6	85.3	97.9	96.8	83.2	97.8	98.0	82.1	100.0	97.9	85.3	98.9	98.8
	0	69.8	95.7	96.8	71.6	96.8	97.9	78.8	96.8	96.8	74.7	98.9	96.9
	–6	40.0	90.5	92.6	35.3	88.4	94.7	57.7	92.6	96.8	51.2	93.3	96.8
	–12	13.7	64.8	77.9	8.4	55.2	72.9	11.6	67.1	75.8	21.1	70.7	76.8
baby	6	94.7	100.0	100.0	94.8	97.9	98.9	92.6	93.7	100.0	94.8	100.0	100.0
	0	83.1	97.9	100.0	88.7	97.4	100.0	84.6	95.8	100.0	89.5	97.9	100.0
	–6	75.7	95.2	96.8	80.6	92.6	95.8	79.8	92.6	94.9	80.0	94.7	97.9
	–12	56.8	73.7	87.4	61.1	78.9	85.7	71.6	80.5	87.4	74.7	81.5	91.2
babble	6	81.1	92.6	98.9	85.2	93.7	100.0	82.1	89.5	97.9	85.3	94.7	97.8
	0	66.2	87.4	97.8	68.7	88.4	97.9	73.6	88.4	98.6	72.5	92.6	98.9
	–6	47.6	64.2	87.5	44.2	57.9	78.7	60.5	72.2	88.4	61.2	76.8	87.6
	–12	44.5	58.2	67.4	34.7	53.4	75.7	57.9	64.7	71.6	52.6	61.2	78.9
alarm	6	95.8	100.0	98.9	94.7	100.0	97.9	91.6	98.9	94.7	94.7	98.9	98.9
	0	89.7	98.9	98.9	87.4	96.8	97.9	84.2	97.9	94.7	90.5	98.9	98.9
	–6	69.9	93.7	95.8	63.9	94.7	95.8	66.1	94.7	95.3	86.8	95.2	95.8
	–12	21.1	86.3	84.2	30.5	85.2	91.6	43.2	83.2	88.4	64.2	85.3	83.2
telephone	6	69.5	100.0	98.9	75.8	100.0	98.7	70.5	100	100.0	83.2	100.0	100.0
	0	48.3	91.5	91.6	52.4	99.1	97.8	46.9	95.8	96.8	62.4	97.9	98.9
	–6	29.5	88.4	89.5	25.1	84.7	89.1	26.2	88.7	91.6	41.3	90.5	92.6
	–12	16.8	77.6	88.4	10.5	81.1	86.7	14.7	77.9	89.7	21.1	84.2	85.9
white	6	64.2	85.3	89.5	60.0	90.5	95.8	54.7	84.2	84.2	66.3	83.2	93.7
	0	41.1	78.9	82.1	38.9	82.1	83.2	26.3	76.8	83.2	45.3	82.6	86.4
	–6	18.9	48.7	61.1	16.8	44.2	64.2	5.3	31.9	59.6	19.5	41.1	57.9
	–12	10.5	18.9	44.2	7.4	14.7	35.9	2.1	7.4	12.6	8.4	13.7	27.4

此时噪声的能量大都集中在此频率带，导致可利用的特征信息存在偏差。本文算法较对比算法2平均高出2%左右，这得益于DNN的低信噪比下数据建模能力较GMM强。

最后，白噪声经常作为多条件训练[14,15]的添加噪声，在训练过程中用来模拟房间混响对双耳空间线索的影响。低信噪比下白噪声作为干扰源导致混响作用进一步加重，噪声能量分布于全频带，目标双耳空间线索的方位角信息被噪声掩盖，导致3种算法定位精度都比较低。但是，2种利用频谱信息的定位方法都有很大程度性能的提升。对比算法2利用GMM对噪声和语音分别建模，但是在更低的信噪比下对随机白噪声建模的准确度不高，导致定位所使用的权重不可靠，随着信噪比的降低定位精度大幅下降。而本文算法直接利用深度学习在混合语音中直接估计所使用的时频单元，避免了需要对随机白噪声的建模的问题，针对性更强，鲁棒性更强，取得了较好的精度，在低信噪比下较对比算法2提升17%左右，在4个信噪比下较对比算法2平均高出6.3%。

综上所述，本文提出的基于深度学习和时频单元选择的算法通过含噪语音的频谱特征使用深度神经网络从含噪语音频谱中直接估计目标语音的时频单元，避免了在低信噪比下对某些噪声建模不准确的问题，提升了鲁棒性，有效地改善了各种噪声下的目标声源定位精度，特别是在较低信噪比下较对比算法1和对比算法2的性能改善较大。

4　结束语

基于时频单元选择的目标声源定位算法对于低信噪比和高混响中的目标语音定位精度较高。与对比算法相比，对比算法虽然能够在较高信噪比下有效地进行目标声源的定位，但是，在低信噪比下很大程度上由于目标声源的信息被掩盖定位精度较低。而本文算法首先利用具有强大非线性映射能力的深度学习，将双耳的目标声源的频谱特征作为输

入训练了一个基于深度学习的时频单元选择模型，然后使用时频单元选择模型直接从双耳输入信号中提取可靠的时频单元，以减少非目标时频单元对定位精度的负面影响，利用语音信号的短时正交特性，使用可靠的语音时频单元提供的空间信息进行定位，不仅对低信噪比下复杂的声学环境具有更强的鲁棒性，而且进一步提高了与目标语音类似的噪声环境中目标声源的定位精度，可为语音识别等提供可靠的前端处理。

参 考 文 献

[1] MAY T, VAN DE PAR S, and KOHLRAUSCH A. A probabilistic model for robust localization based on a binaural auditory front-end[J]. *IEEE Transactions on Audio, Speech, and Language Processing*, 2011, 19(1): 1–13. doi: 10.1109/TASL.2010.2042128.

[2] 李如玮, 潘冬梅, 张爽, 等. 基于Gammatone滤波器分解的HRTF和GMM的双耳声源定位算法[J]. 北京工业大学学报, 2018, 44(11): 1385–1390. doi: 10.11936/bjutxb2017090015.
LI Ruwei, PAN Dongmei, ZHANG Shuang, *et al*. Binaural sound source localization algorithm based on HRTF and GMM Under Gammatone filter decomposition[J]. *Journal of Beijing University of Technology*, 2018, 44(11): 1385–1390. doi: 10.11936/bjutxb2017090015.

[3] WOODRUFF J and WANG Deliang. Binaural localization of multiple sources in reverberant and noisy environments [J]. *IEEE Transactions on Audio, Speech, and Language Processing*, 2012, 20(5): 1503–1512. doi: 10.1109/TASL.2012.2183869.

[4] MA Ning, MAY T, and BROWN G J. Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments[J]. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2017, 25(12): 2444–2453. doi: 10.1109/TASLP.2017.2750760.

[5] MA Ning, GONZALEZ J A, and BROWN G J. Robust binaural localization of a target sound source by combining spectral source models and deep neural networks[J]. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2018, 26(11): 2122–2131. doi: 10.1109/TASLP.2018.2855960.

[6] WANG Yuxuan, HAN Kun, and WANG Deliang. Exploring monaural features for classification-based speech segregation[J]. *IEEE Transactions on Audio, Speech, and Language Processing*, 2013, 21(2): 270–279. doi: 10.1109/TASL.2012.2221459.

[7] JIANG Yi, WANG Deliang, LIU Runsheng, *et al*. Binaural classification for reverberant speech segregation using deep neural networks[J]. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2014, 22(12): 2112–2121. doi: 10.1109/TASLP.2014.2361023.

[8] ZHANG Xueliang and WANG Deliang. Deep learning based binaural speech separation in reverberant environments[J]. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2017, 25(5): 1075–1084. doi: 10.1109/TASLP.2017.2687104.

[9] WANG Deliang and 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. doi: 10.1109/TASLP.2018.2842159.

[10] WIERSTORF H, GEIER M, and SPORS S. A free database of head related impulse response measurements in the horizontal plane with multiple distances[C]. The Audio Engineering Society Convention 130, Berlin, Germany, 2011.

[11] HUMMERSONE C, MASON R, and BROOKES T. Dynamic precedence effect modeling for source separation in reverberant environments[J]. *IEEE Transactions on Audio, Speech, and Language Processing*, 2010, 18(7): 1867–1871. doi: 10.1109/TASL.2010.2051354.

[12] MA Ning, BROWN G J, and GONZALEZ J A. Exploiting top-down source models to improve binaural localisation of multiple sources in reverberant environments[C]. The 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, 2015: 160–164.

[13] COOKE M, BARKER J, CUNNINGHAM S, *et al*. An audio-visual corpus for speech perception and automatic speech recognition[J]. *The Journal of the Acoustical Society of America*, 2006, 120(5): 2421–2424. doi: 10.1121/1.2229005.

[14] MAY T, MA Ning, and BROWN G J. Robust localisation of multiple speakers exploiting head movements and multi-conditional training of binaural cues[C]. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing, Brisbane, Australia, 2015: 2679–2683. doi: 10.1109/ICASSP.2015.7178457.

[15] MAY T. Robust speech dereverberation with a neural network-based post-filter that exploits multi-conditional training of binaural cues[J]. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2018, 26(2): 406–414. doi: 10.1109/TASLP.2017.2765819.

李如玮：女，1972年生，博士，副教授，硕士生导师，研究方向为语音信号处理.

李 涛：男，1994年生，硕士生，研究方向为语音信号处理.

孙晓月：女，1995年生，硕士生，研究方向为语音信号处理.

杨登才：男，1978年生，博士，副研究员，信号与光信号处理.

王 琪：女，1991年生，在站博士后，研究方向为语音信号处理.