基于卷积神经网络的室内麦克风阵列声源定位算法

焦琛*，张涛，孙建红
天津大学电气自动化与信息工程学院，天津 300072

摘要 在室内麦克风阵列声源定位算法的研究中，混响和噪声对定位精度影响很大，传统的声源定位算法无法在高混响和低信噪比的环境中保持较高的定位精度。为了解决这一问题，提出一种基于卷积神经网络的室内声源定位算法，该算法提取麦克风阵列接收信号的相位加权广义互相关函数作为训练特征，获取目标声源三维位置信息。基于 NOIZEUS 数据库的实验结果表明，该方法能够通过训练适应不同的声学环境，与其他基于学习的室内声源定位算法相比，其在高混响与低信噪比环境下仍具有较好的定位性能与鲁棒性，具有较大的研究和应用价值。

关键词 图像处理；室内声源定位；卷积神经网络；广义互相关函数
中图分类号 TB52+9；TN98 **文献标志码** A **doi**：10.3788/LOP57.081021

Convolutional Neural Network Based Indoor Microphone Array Sound Source Localization

Jiao Chen*, Zhang Tao, Sun Jianhong
School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, *China*

Abstract For indoor sound source localization algorithm based on microphone arrays, its accuracy is greatly influenced by the reverberation and noise. Traditional sound source localization approaches cannot keep high localization accuracy in strong reverberation and low signal-to-noise ratio environments. To tackle this problem, a novel indoor sound source localization algorithm based on convolutional neural network is proposed. By extracting the phase weighted generalized cross correlation function of the received signals from microphone arrays as training feature, the three-dimensional localization information of target sound source can be obtained. Experiments based on NOIZEUS database demonstrate that the proposed algorithm can be adapted to different acoustic conditions via training. Compared with other learning based indoor sound source localization algorithms, the proposed algorithm has good localization performance and robustness in strong reverberation and low signal-to-noise ratio environment, suggesting high research and application value.

Key words image processing; indoor sound source localization; convolutional neural network; generalized cross correlation function
OCIS codes 100.3005; 100.4550; 070.5010; 070.1060

1 引 言

声源定位(SSL)技术最初主要应用在军事领域，近年来随着社会的进步、生活质量的提高，其被广泛应用于视频会议[1]、智能家居[2]、车载通话设备[3]等场景中。其中，室内环境是声源定位技术最常用的应用场景，区别于室外远场环境，室内环境下混响严重，对定位性能有很大影响。对于室内声源定位算法而言，如何提高抗噪声和抗混响能力是长久以来的研究重点和难点问题。传统的声源定位算法，如基于时延估计的声源定位算法[4]、基于高分辨率谱估计的声源定位算法[5]以及基于可控波束形成的声源定位算法[6]，在高混响和低信噪比(SNR)环境下定位性能很差。

随着机器学习和深度学习的快速发展，许多传统的问题都可以使用该方法解决，室内声源定

收稿日期：2019-08-29；**修回日期**：2019-09-07；**录用日期**：2019-09-19
基金项目：国家自然科学基金(61350009,61179045)
* **E-mail**：jiaochen@tju.edu.cn

位问题也不例外。文献[7]中使用 LS-SVM 算法解决声源定位算法，该算法对于解决在噪声和混响条件下的二维定向问题有一定效果。文献[8]中使用后向传播(BP)神经网络进行声源定位，该方法对于在混响和噪声环境下的二维定向比基于 LS-SVM 的方法具有更小的均方根误差(RMSE)，但二者都无法获取声源的三维位置信息。文献[9]中使用概率神经网络(PNN)算法解决声源定位问题，其最重要的优点是不需要任何迭代训练，定位速度快而且具有一定的抗混响和抗噪声的性能，但在高混响与低信噪比环境下仍无法保持较高的定位精度。

在恶劣的声学环境下，为了进一步提高定位性能，需要解决两个问题。一方面，在室内环境中，多径传播造成的室内混响导致谱峰偏移或多重谱峰，影响声源定位性能[10]；另一方面，噪声在一定程度上减弱声源信号强度，低信噪比环境中很难达到高定位性能[11]。因此，本文提出使用基于卷积神经网络(CNN)[12-14]的算法来解决声源定位问题。声源信号来自 NOIZEUS 数据库[15]，使用麦克风阵列接收房间内的声音信号，将其相位加权广义互相关函数(GCC-PHAT)作为输入特征，声源的三维位置作为网络输出。结果表明，相比于传统声源定位算法和其他机器学习声源定位算法，本文算法在不同程度的噪声和混响环境下，尤其是高混响与低信噪比环境下具有更高的定位分类精度和更强的鲁棒性。该研究对于室内声源定位算法的抗噪声与抗混响能力具有一定的借鉴意义。

2 基本原理分析

2.1 房间分类定位模型

室内声源定位问题研究的目标声源是在一个三维(3D)矩形封闭室内的静止单源，声源信号可以出现在房间内任意位置，在这里使用极坐标系(r,φ,θ)来描述房间中声源的位置，其中 r 表示声源 S 到麦克风阵列中心的距离，φ 表示声源 S 与麦克风阵列中心的方向角，θ 表示声源 S 与麦克风阵列中心的俯仰角。房间被划分为 K 个独立等大的子空间，假设子空间的体积足够小，每个子空间都可以表示为一个空间内独一无二的三维坐标，由此可以将三维空间内的声源定位由线性回归问题转化为基于概率分布的非线性分类问题来处理，这极大地减少计算量。基于从阵列接收的声源信号中提取出的位置特征，可以选择不同的分类器以确定声源属于哪一个子空间。

如图 1 所示，$s(t)$表示目标声源，$x_i(t)(i=1,2,\cdots,m)$表示麦克风阵列中的阵元，m 为阵列中麦克风数量。在分类问题上，每个子空间都是一个类别，共有 K 个类别，表示为 $C=\{c_1,c_2,\cdots,c_K\}$，$c_i\in\mathbf{R}^3, i\in\{1,2,\cdots,K\}$。子空间的体积越小，$K$ 值越大，分类的复杂度越高，同时定位分类的分辨率越高，定位精度也会相应提升。所有 K 类都是可能的声源位置，每个可能的类 c_i 都有一组决定声源位置属于哪一类的特征。c_i 包含声源信号的三维位置信息，基于传声器接收信号 X 和训练特征 f_i，可以确定 c_i，进而获得目标声源的位置信息。

分类问题可以表示为

$$c_s=classify\left(X,\sum_{\forall i}f_i\right), \tag{1}$$

式中：$classify(\cdot)$表示分类器函数；c_s 为目标声源的预测位置。

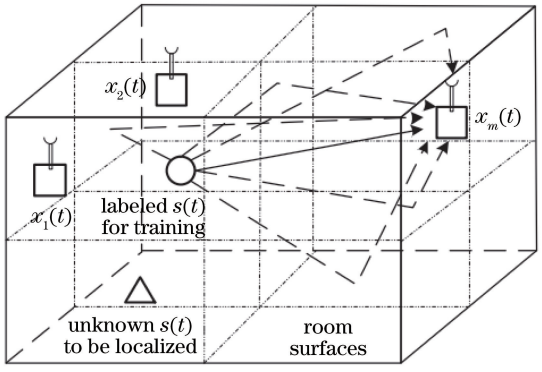

图 1 声源定位问题的空间集群分类

Fig. 1 Space cluster classification for SSL

2.2 定位方法

基于卷积神经网络的室内声源定位方法分为两个阶段，第一个阶段是训练模型，第二个阶段是定位测试。

在训练模型阶段：1)在噪声环境中，利用麦克风阵列接收位于训练位置 $r_k(k=1,2,\cdots,K)$的声源发出的混响信号；2)对接收到的混响信号进行预处理后，计算相位变换加权广义互相关函数；3)由计算出的广义互相关函数生成特征数据 Y；4)利用特征数据 Y 训练卷积神经网络模型。

在定位测试阶段：1)在噪声环境中，利用麦克风阵列接收某一位置的声源发出的混响信号；2)对接收到的混响信号进行预处理后，计算相位变换加权广义互相关函数；3)由计算出的广义互相关函数生成特征数据 Y'；4)利用训练阶段训练的卷积神经网络模型估计声源的位置。

整个流程如图 2 所示。

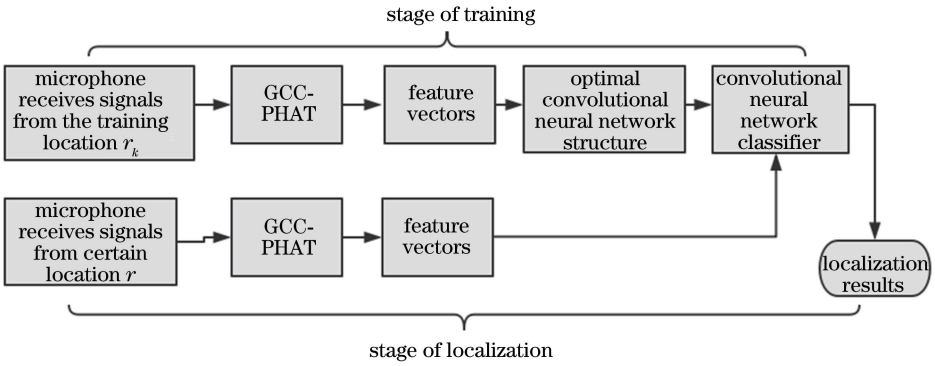

图 2 定位方法流程图

Fig. 2 Flow chart of localization method

2.3 仿真

室内混响环境由 RoomSim 仿真工具[16]进行模拟，噪声为与声源信号无关的高斯白噪声，为了使结果可信，所有涉及到的训练方法所用的数据都来自同一个样本数据库，训练数据和测试数据按照 4∶1 的比例随机选取。房间大小设置为 4 m×4 m×4 m，声速为 340 m/s，实验中分别设混响时间为 0，300，600 ms 三种情况。使用 48 kHz 采样频率的纯净语音作为声源，语音长度为 2.7 s，语音频率为 220 Hz～3.4 kHz，房间中的信号分 SNR 为 0 dB 和 10 dB 的信号以及 clean(纯净语音信号)三种情况。对传声器接收到的语音信号进行分帧处理，每帧的持续时间选择为 0.064 s，两帧之间的重叠率设定为 62.5%。

参照文献[9]中声源空间位置的划分方式，将整个房间划分为 540 份。其中半径值 r 分别取(0～1 m、1～1.5 m、>1.5 m)；方向角 φ 的取值范围为 −180°～180°，步长为 18°；俯仰角 θ 的取值范围为 −90°～90°，步长为 20°。在每 0.25 m×0.25 m 的空间范围内随机生成 1 个声源信号，房间内共生成 4096 个位置样本，其中声源距离传声器的最远距离为 6.84 m，声源到达传声器的位置最多需要 0.02 s。阵列接收的有效语音信号长度为 2.72 s，其余为空白语音或混响引起的重复语音。有效语音长度应为 112 帧，因此选取传声器接收到的语音信号的前 112 帧，这样可以在一定程度上减少计算量，减小混响对定位精度的影响。

考虑到文献[17]中七元麦克风阵列模型良好的定位性能，选择其作为声源定位实验的传声器阵列模型。麦克风阵列尺寸对声源定位性能有很大影响，传声器阵列间距越大，俯仰角、方向角和距离的测量标准差越小，即定位性能越好，但传声器阵列结构过大会增加经济成本且增加设备实现难度，因此综合考量定位性能与房间大小，设置阵列中相邻传声器之间的距离为 0.2 m。由于阵列的任何两个传声器之间的最大距离为 0.4 m，则可能的最大时间延迟为 1.17 ms，由于采样频率为 48 kHz，故样本中的最大延迟数为 56，考虑到混响造成的相关函数谱峰偏移，本文选取数据中心分别向前向后共 128 个点的互相关作为特征。用 $M(M\geqslant 2)$ 个传声器进行声源定位，可以获得要使用的 21 组 GCC-PHAT 特征。综上，对接收到的语音信号进行分帧处理后得到 112 帧语音信号，每一帧数据提取出的特征维度为 2688，得到每个样本数据的维度为 112×2688。

2.4 特征提取

对于基于到达时间差(TDOA)的声源定位算法来说，时间差的获取需要依靠互相关函数，而其有效性受混响和噪声的影响比较大，因此衍生出加权算法[18]，以突出相关函数峰值，在一定程度上可以削弱噪声和混响的干扰。最常用的加权函数是相位变换(PHAT)和最大似然(ML)。在高信噪比环境下，GCC-PHAT 具有良好的抗混响性能，GCC-ML 对噪声具有良好的鲁棒性，但混响对其影响较大。在室内环境中，混响较为严重，因此选择 GCC-PHAT 作为训练特征。当混响时间为 0 ms 且无噪声干扰时，提取的特征峰最清晰、最明显，混响会导致谱峰偏移或虚假谱峰，且混响时间越长，这种现象越明显。噪声会造成谱峰模糊或局部峰值减小，且随着室内噪声的增加，谱峰模糊现象会加重。以上问题，都会对谱峰搜索造成干扰，进而影响时延估计的精度，最终导致无法准确判断目标声源位置。这也是传统声源定位算法在高混响和低信噪比环境下无法达到较高定位精度的原因。

2.5 CNN 网络结构

CNN 网络由输入层、卷积层、池化层、全连接层和输出层组成。其中卷积层主要用于对不同的局部矩阵和输入图像的卷积核矩阵进行卷积运算，以提取图像特征。池化层主要用于压缩卷积层提取的特征，减小特征维数，从而减少计算量，防止过拟合，提高计算速度。本章卷积神经网络模型为改进的 LeNet -5 结构，其由 4 个卷积层和 4 个池化层组成。为了尽可能保证特征的完整性，选择 max pooling 方法。模型结构参数如表 1 所示，其中 kernel_size 表示卷积核大小，stride 表示步长，pad 表示边缘扩充参数，pooling 表示池化方式，dropout 表示神经元失效率，iterations 表示迭代次数，batch_size 表示批尺寸。

表 1 CNN 结构

Table 1 CNN structure

Network layer	Network parameter
Input layer	Dimension: 112×2688
Convolution layer	Number of convolution kernel: 8; kernel_size: 5×5; stride: 1; pad: 0
Pooling layer	Pooling: max pooling; kernel_size: 3×3; stride: 1; pad: 0; dropout: 50%
Convolution layer	Number of convolution kernel: 16; kernel_size: 5×5; stride: 1; pad 0
Pooling layer	Pooling: max pooling; kernel_size: 3×3; stride: 1; pad 0; dropout: 50%
Convolution layer	Number of convolution kernel: 32; kernel_size: 5×5; stride: 1; pad 0
Pooling layer	Pooling: max pooling; kernel_size: 3×3; stride: 1; pad 0; dropout: 50%
Convolution layer	Number of convolution kernel: 64; kernel_size: 5×5; stride: 1; pad 0
Pooling layer	Pooling: max pooling; kernel_size: 3×3; stride: 1; pad 0; dropout: 50%
Connection layer	Number of neurons: 1024; activation function: ReLU; dropout: 50%
Output layer	Activation function: softmax; learning rate : 0.001; iterations: 1000; batch_size: 64

3 结果分析

本文将基于 CNN 与 SVM、PNN、BP 神经网络的室内 SSL 算法，与传统的 TDOA 进行了对比实验，并对它们的位置分类结果进行了比较。其中，参照文献[7]，设置 SVM 卷积核为三层高斯核函数，标签编码方式使用最小输出编码方法；参照文献[8]，设置 BP 神经网络为单隐藏层，包含 1024 个神经元；参照文献[9]，设置 PNN 训练速度为 0.5。分类结果及算法复杂度如表 2、3 所示。

表 2 各类算法在不同环境的分类准确率

Table 2 Classification accuracy of different algorithms in different environments

Signal	Reverberation time /ms	Accuracy of TDOA /%	Accuracy of SVM /%	Accuracy of PNN /%	Accuracy of BP /%	Accuracy of CNN /%
Clean voice	0	62.61	78.58	96.87	96.32	96.67
Clean voice	300	61.64	75.28	96.30	95.53	95.03
Clean voice	600	60.02	74.86	93.25	92.84	93.80
SNR: 10 dB	0	42.92	44.93	90.16	94.08	95.83
SNR: 10 dB	300	46.31	49.82	89.44	90.36	93.87
SNR: 10 dB	600	36.53	46.35	88.73	87.69	92.79
SNR: 0 dB	0	38.61	45.57	89.89	88.81	94.32
SNR: 0 dB	300	34.61	44.03	89.09	86.37	93.49
SNR: 0 dB	300	24.87	44.60	88.61	85.20	90.78

由表 2 可知，无论使用哪种方法，随着室内混响和噪声情况的加剧，算法定位性能都有所下将。其中使用 SVM 的方法训练得到的定位分类准确率无论在什么样的室内条件下都远低于使用 PNN、BP 神经网络、CNN 网络方法训练获取的定位分类准确率，并且该方法在恶劣的环境下定位性能下降更为严重。由图 3 可知，本文提出的基于 CNN 的室内传声器阵列声源定位算法的定位精度高于使用 PNN、BP 神经网络方法训练的定位精度，尤其是在高混响和低信噪比环境下，仍然能保持较高的定位性能。表 3 显示了不同声源定位算法的离线训练时长与在线定位时长，结果表明 SVM 离线训练时间最短，CNN 的离线训练时间最长。但是，这 4 种算法的在线定位时间相差并不大。考虑到在室内环境已知的情况下，可以提前训练算法的网络模型，因此，由在线定位时长度量的计算复杂度并不是算

表 3 各算法的实时定位时间

Table 3 Real-time localization time of different algorithms

Signal	Reverberation time /ms	Iterations	Localization time /s	Iterations	Localization time of PNN /s	Iterations	Localization time of BP /s	Iterations	Localization time of CNN /s
Clean voice	0	4496	7.4	4671	6.8	10504	10.3	18723	11.2
Clean voice	300	4215	7.8	5387	7.1	13661	9.8	19025	10.8
Clean voice	600	4615	8.1	6062	6.1	22572	10.3	29781	10.9
SNR: 10 dB	0	4853	9.3	5283	7.4	21973	9.7	33671	12.4
SNR: 10 dB	300	49.82	8.2	6231	7.3	34603	9.7	42101	11.8
SNR: 10 dB	600	5430	9.1	5735	7.1	38524	10.9	48776	11.2
SNR: 0 dB	0	4551	9.2	5089	6.9	35871	10.6	54786	11.2
SNR: 0 dB	300	4876	9.4	6086	8.3	41654	10.8	57623	10.5
SNR: 0 dB	300	5576	9.3	6487	8.1	47726	10.8	58341	11.9

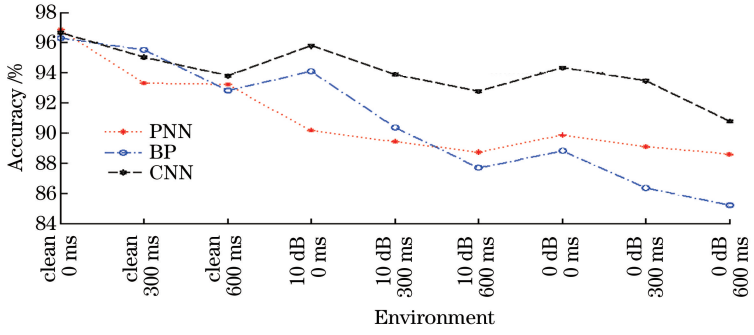

图 3 各类算法在不同环境的分类准确率对比

Fig. 3 Comparison of classification accuracy of different algorithms in different environments

法选择的主要考虑因素。

4 结 论

为解决室内声源定位算法在高混响和低 SNR 环境下定位性能差的问题，提出一种基于 CNN 网络的室内传声器阵列声源定位算法。通过从阵列接收到的信号中提取与声源位置密切相关的 GCC-PHAT 作为训练特征，声源信号的三维位置坐标作为输出，将室内声源定位问题转换成分类问题，该问题融合了室内声源定位与深度学习这两个当前研究的热点问题。结果表明，基于 CNN 网络的室内传声器阵列声源定位算法在恶劣的室内环境下仍然能保持较好的定位性能，例如在混响时间为 600 ms、SNR 为 0 dB 时定位分类准确率仍然能够达到 90% 以上，该方法在糟糕的室内环境下表现出优于基于 PNN 和 BP 神经网络的算法以及传统算法的性能，且并没有明显增加在线定位时间，具有较大的应用和研究价值。

参 考 文 献

[1] Zhao Z, Chen W H, Semprun K A, et al. Design and evaluation of a prototype system for real-time monitoring of vehicle honking[J]. IEEE Transactions on Vehicular Technology, 2019, 68(4): 3257-3267.

[2] Shivappa S, Trivedi M, Rao B. Audiovisual information fusion in human-computer interfaces and intelligent environments: a survey[J]. Proceedings of the IEEE, 2010, 98(10): 1692-1715.

[3] Popper A N, Fay R R. Sound source localization [M]. New York: Springer-Verlag, 2005.

[4] Li X, Deng Z D, Rauchenstein L T, et al. Contributed review: source-localization algorithms and applications using time of arrival and time difference of arrival measurements [J]. Review of Scientific Instruments, 2016, 87(4): 041502.

[5] Alameda-Pineda X, Horaud R. A geometric approach to sound source localization from time-delay estimates [J]. ACM Transactions on Audio, Speech, and Language Processing, 2014, 22(6): 1082-1095.

[6] Yook D, Lee T, Cho Y. Fast sound source localization using two-level search space clustering [J]. IEEE Transactions on Cybernetics, 2016, 46 (1): 20-26.

[7] Chen H W, Ser W. Acoustic source localization using

LS-SVMs without calibration of microphone arrays[C]// 2009 IEEE International Symposium on Circuits and Systems, May 24-27, 2009, Taipei, Taiwan, China. New York: IEEE, 2009: 1863-1866.

[8] Xiao X, Zhao S, Zhong X, et al. A learning-based approach to direction of arrival estimation in noisy and reverberant environments [C] // 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), April 19-24, 2015, Brisbane, QLD, Australia. New York: IEEE, 2015: 2814-2818.

[9] Sun Y X, Chen J J, Yuen C, et al. Indoor sound source localization with probabilistic neural network [J]. IEEE Transactions on Industrial Electronics, 2018, 65(8): 6403-6413.

[10] Hurst P J, Norrell A. DAC quantization-noise cancellation in an echo-canceling transceiver [J]. IEEE Transactions on Circuits and Systems II: Express Briefs, 2008, 55(2): 111-115.

[11] Prodeus A M. Performance measures of noise reduction algorithms in voice control channels of UAVs [C] // 2015 IEEE International Conference Actual Problems of Unmanned Aerial Vehicles Developments (APUAVD), October 13-15, 2015, Kiev, Ukraine. New York: IEEE, 2015: 189-192.

[12] Zhang Y S, Yang G W, Wang Q Q, et al. Weld feature extraction based on fully convolutional networks[J]. Chinese Journal of Lasers, 2019, 46(3): 0302002.
张永帅, 杨国威, 王琦琦, 等. 基于全卷积神经网络的焊缝特征提取[J]. 中国激光, 2019, 46(3): 0302002.

[13] Wang W X, Fu Y T, Dong F, et al. Infrared ship target detection method based on deep convolution neural network [J]. Acta Optica Sinica, 2018, 38(7): 0712006.
王文秀, 傅雨田, 董峰, 等. 基于深度卷积神经网络的红外船只目标检测方法[J]. 光学学报, 2018, 38(7): 0712006.

[14] Wang S Y, Tao S X, Yang F, et al. Laser range-gated imaging target recognition based on convolutional neural network [J]. Laser & Optoelectronics Progress, 2019, 56(2): 021001.
王书宇, 陶声祥, 杨钒, 等. 基于卷积神经网络的激光距离选通式成像目标识别[J]. 激光与光电子学进展, 2019, 56(2): 021001.

[15] Patil U G, Shirbahadurkar S D. Performance analysis of SS based speech enhancement algorithms for ASR with non-stationary noisy database-NOIZEUS [C] // 2018 2nd International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC) I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2018 2nd International Conference on, August 30-31, 2018, Palladam, India. New York: IEEE, 2018: 636-641.

[16] Alien J B, Berkley D A. Image method for efficiently simulating small - room acoustics [J]. The Journal of the Acoustical Society of America, 1976, 60(S1): S9.

[17] Yang X, Xing H Y, Zhang J, et al. Performance analysis of sound source localization algorithm based on seven-element microphone array [J]. Chinese Journal of Sensors and Actuators, 2019, 32(7): 1034-1039.
杨旭, 行鸿彦, 张军, 等. 基于七元传声器阵列的声源定位算法及性能分析[J]. 传感技术学报, 2019, 32(7): 1034-1039.

[18] Knapp C, Carter G. The generalized correlation method for estimation of time delay [J]. IEEE transactions on acoustics, speech, and signal processing, 1976, 24(4): 320-327.