Archives of Computational Methods in Engineering (2022) 29:4631–4642
https://doi.org/10.1007/s11831-022-09747-2

REVIEW ARTICLE

A Review on Sound Source Localization Systems

Dhwani Desai[1] · Ninad Mehendale[1]

Received: 5 October 2021 / Accepted: 11 April 2022 / Published online: 5 May 2022

Abstract
Sound Source Localization (SSL) systems focus on finding the direction of a sound source. Sound source localization is an essential feature in robots and humanoids. Research is being done for two decades to optimize SSL techniques and enhance their accuracy. Presented in this review we have categorized various proposed SSL techniques into four main types. Out of which one type is SSL that is based on conventional algorithms like Generalized Cross Correlation (GCC), Multiple Signal Classification (MUSIC), Time Difference of Arrival (TDOA), etc. using multiple microphones array configurations. The second type involves techniques based on binaural signal processing using conventional algorithms (GCC, MUSIC, TDOA). SSL techniques that fall under the third and fourth types of categories are developed recently in the last decade with the rise of Convolutional Neural Network (CNN) algorithms. The third type of SSL technique makes use of multiple microphone array configurations using CNN and the fourth type involves the most recently evolved technique based on binaural signal processing using CNN. The different SSL techniques based on multiaural and binaural signals using the conventional algorithm as well as CNN are presented in this review. The review paper provides an overview of SSL systems in terms of the number of microphones used, layouts of microphonic arrays, algorithms to perform SSL and localization in 3D space (azimuth, elevation and distance). From the review we found that out of all SSL techniques, CNN is the emerging and optimized one. By using CNN in SSL systems, the least error rate of 0.1 % was achieved.

1 Introduction

Sound source localization is one of the main functions of the human ear. Human beings can hear sound waves in the frequency range of 20 Hz to 20 Khz. The human auditory system not only detects sound signals but also can cancel out noise and reverberations and extract useful information from multiple sound signals. Research has been done to generate models and systems for the artificial simulation of a human ear. An artificial human auditory system can be developed by designing an electronic robotic ear. The robotic ear should perform all or at least one of the three main functions of the human ear. These three main functions are Sound Event Detection (SED), sound type classification and Sound Source Localization (SSL). SED systems detect relevant sound events by efficient cancellation of noise and echo. Sound type classification systems have to classify between voiced and unvoiced signals, speech detection, vehicle or emergency signal detection, etc. SSL systems detect the direction of source of sound. SSL systems analyze variations in components of sound signals (amplitude, frequency, phase) received at each audio sensor (microphone).

We present a review of different design techniques of the SSL system proposed till date. Various parameters should be taken into consideration before designing SSL systems. Figure 1 shows four main different parameters which need to be considered while designing an SSL system. The first parameter is the number of microphones used to record audio signals from a single or multiple sound sources. SSL systems can be designed using different number of microphones that is from 2,3,4,6,8,16 or even more. SSL systems using two microphones are termed binaural SSL systems. Likewise, SSL systems that use 3,4 and multiple microphones are termed tri-aural, tetra-aural and multiaural SSL systems respectively. These microphones can be arranged in different geometrical combinations. The second parameter which has to be chosen before designing SSL system is 3-dimensional space (azimuth, elevation and distance) that SSL system will locate. SSL systems can be designed to detect only angular (azimuth) estimation in a single plane, or can also incorporate height and depth (elevation) estimation to perform 2D sound source localization or can even be designed to

✉ Ninad Mehendale
ninad@somaiya.edu

[1] K. J. Somaiya College of Engineering, Mumbai, India

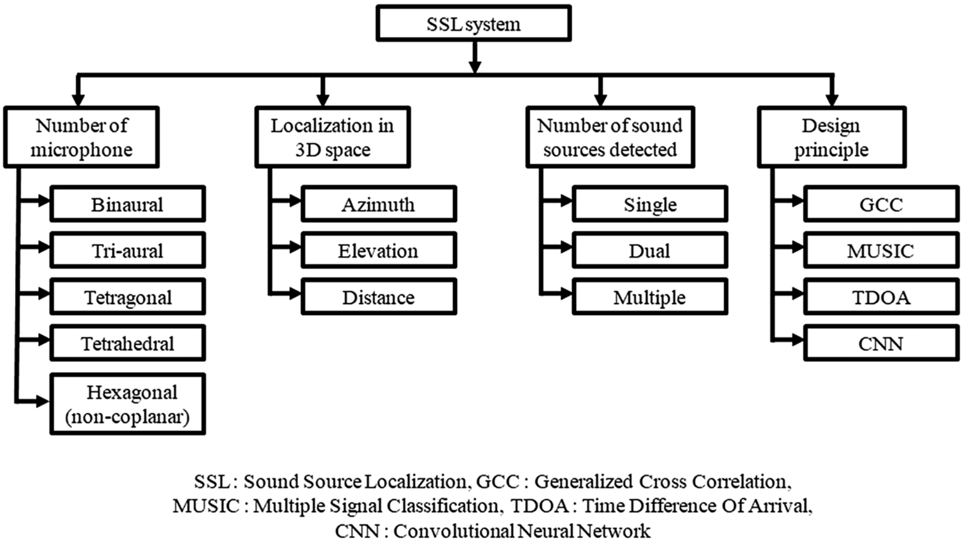
SSL system
Number of microphone
Binaural
Tri-aural
Tetragonal
Tetrahedral
Hexagonal (non-coplanar)
Localization in 3D space
Azimuth
Elevation
Distance
Number of sound sources detected
Single
Dual
Multiple
Design principle
GCC
MUSIC
TDOA
CNN
SSL : Sound Source Localization, GCC : Generalized Cross Correlation,
MUSIC : Multiple Signal Classification, TDOA : Time Difference Of Arrival,
CNN : Convolutional Neural Network

Fig. 1 SSL techniques can be classified based on four main parameters such as the number of microphones, localization in 3D space, number of sound sources detected and principle technique used. SSL system can be designed using binaural, tri-aural, tetragonal, tetrahedral, Hexagonal (non-co-planar) microphone configurations. SSL systems can locate the sound source in 3D space by azimuth, elevation and distance estimation, It can detect single, dual or multiple sound sources. Generalized Cross Correlation (GCC), Multiple Signal Classification (MUSIC), Time Difference of Arrival (TDOA) and Convolutional Neural Network (CNN) are various techniques used to design SSL systems

estimate the distance to have a complete 3D localization of source of sound. The third parameter which has to be defined is the number of sound sources it can detect. SSL systems can be designed to detect only a single sound source at a given time or dual sound sources simultaneously or even multiple sources of sound together. The fourth and main parameter is the design technique used to design an SSL system. Variations in components of sound signals (amplitude, frequency, phase) received at each microphone of the microphonic array are analyzed. For this purpose, various algorithms like GCC, MUSIC, TDOA, CNN are used. A detailed description of these algorithms is mentioned in Sect. 4.

SSL systems can have multiple microphones and they can be arranged in various geometrical combinations known as microphone array configurations. A researcher can choose to have any type of microphone array configurations such as linear, circular, non-coplanar, tetrahedral, etc as per their requirement. Figure 2 shows different microphone array configurations. Figure 2a shows binaural microphone setup wherein there are only two microphones to perform SSL like in the case of human hearing. Figure 2b shows tri-aural microphone setup where three microphones are placed at a distance apart and it surrounds the source of sound whose location is to be detected. Figure 2c uses tetra-aural arrangement of microphone which also are placed such that it surrounds the sound source for its localization. Instead of microphones surrounding the sound source, clustered microphone array having two, three, six or even more microphone array configurations can also be used as shown in Fig. 2d to detect sound sources situated at some distance apart from the microphone array configuration. Microphone array configurations can also have non-co-planar arrangements as shown in Fig. 2e which has non-co-planar hexagonal microphone array configuration.

Figure 3 shows a flowchart of an SSL system design. The first step is to extract components of sound signal like frequency, phase, amplitude or power. These signal components are recorded using different sets of microphones such as binaural, tri-aural,tetra-aural, multi-aural and different microphonic array configurations like linear, circular, spherical, etc. Variations in components of sound signals received at each microphone of the microphone array are then analyzed using either conventional method containing mathematical algorithms like Approximate Maximum Likelihood (AML), GCC, Multichannel signal Classification (MCCC), TDOA, MUSIC or recently developed new CNN based method. SED and source localization either in 1D, 2D or 3D plane (azimuth, elevation and distance) is then achieved using any of these algorithms.

2 Human Auditory system

Figure 4 shows human ear anatomy which is divided into the outer, middle and inner ear. The outer ear is made up of a fold of cartilages which cancels out noise and also attenuates sound signals that enter into the auditory canal. The auditory

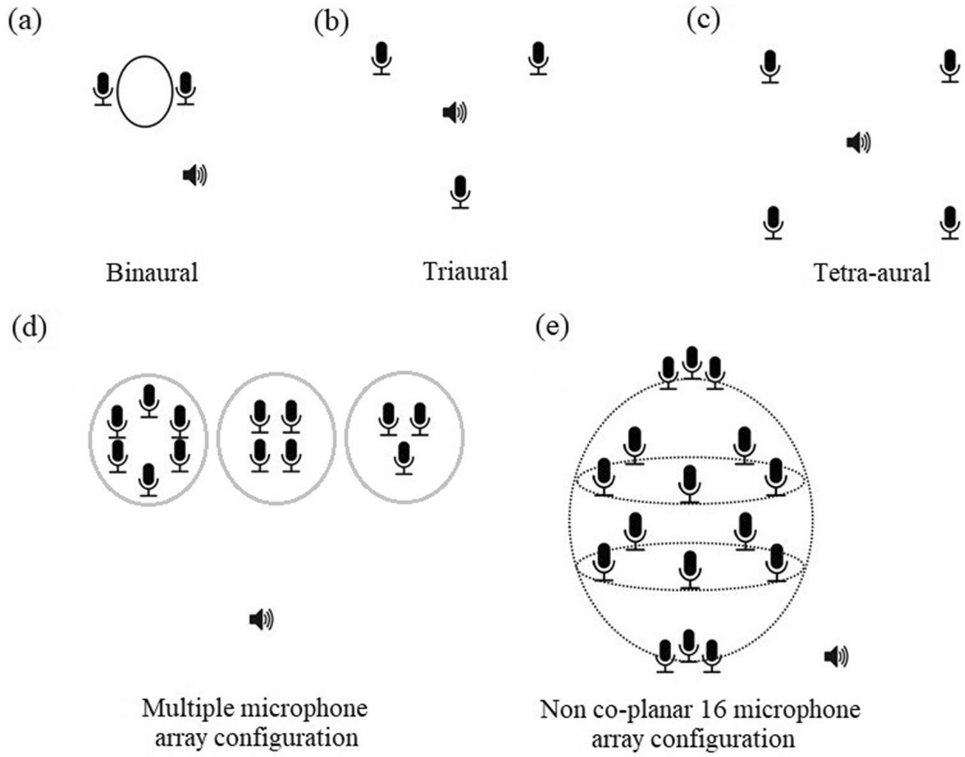
(a)
Binaural
(b)
Triaural
(c)
Tetra-aural
(d)
Multiple microphone array configuration
(e)
Non co-planar 16 microphone array configuration

Fig. 2 Different microphone array configurations. **a** Binaural setup, **b** tri-aural setup with sound source between the three microphones. **c** Tetra-aural setup surrounding the sound source. **d** Clustered microphone array configurations that have six, four or three microphones. **e**16 non-co-planar hexagonal microphone array structures. In (**a**), (**d**) and (**e**) sound source is in the vicinity of the microphone array and in (b) and (c) microphone array surrounds the sound source

canal is a tube-like structure that connects the outer and middle ear and forms a pathway for sound signals to reach the middle ear. Sound waves via the auditory canal reach the eardrum (tympanic membrane). The middle ear consists of three small bones known as malleus, incus and stapes. It converts low-pressure sound signals to high-pressure waves. The high-pressure waves then pass through the oval window or vestibular window and enter into the inner ear. The cochlea is a fluid-filled structure of the inner ear. The cochlea has three sections scala media, scala tympani and scala vestibule. Scala vestibule and scala tympani contain perilymph and are situated between labyrinth bones. Scala media contains endolymph liquid. The basilar membrane receives sound waves from scala media and below the basilar membrane is an organ of Corti which converts mechanical waves into electric signals in neurons. The organ of Corti consists of hair cells that vibrate as per the frequency of sound waves and transfer electric signals to nerve fibers. The tectorial membrane rests above hair cells and moves back and forth with each cycle of the sound wave and thus tilting the hair cells as per sound wave frequency. The nerve signal is then transmitted to the cochlear nucleus and then to the trapezoid body which helps in the localization of sound. The superior olivary complex in the brain receives a signal from the cochlear nucleus and finds out the interaural level difference for binaural computations. Brain stem cells and mid-brain finally integrate all the information and then interpret the sound signal.

Functions of the human auditory system include sound detection, sound type classification, sound source localization, speech signal detection, interpretation and classification, noise cancellation and echo cancellation, multiple sound source detection, classification and localization, detection and analysis of pitch, tone roughness, modulation of sound signals, extraction and interpretation of useful information from incoming sound signals [1].

3 Human Auditory system model

Sound source localization in the human auditory model is based on auditory cues which are important features associated with incoming audio signals which help in the localization of sound sources. Auditory cues are divided into monoaural and binaural cues. Monoaural cues refer to audio signal modifications with respect to a single ear. Monoaural cues are based on the distance of the sound source from one of the ears, frequency of an audio signal, effect of echoes and reverberations to an incoming audio signal, etc. Binaural Cues also known as inter-aural differences are differences in audio signals reaching the two ears or audio sensors which is very useful information or feature for binaural sound source localization. Common binaural cues are Inter-aural Level Difference (ILD), Inter-aural Time Difference (ITD) and Inter-aural Phase Difference (IPD). ILD is the difference in intensity levels of sound signals reaching both left and right

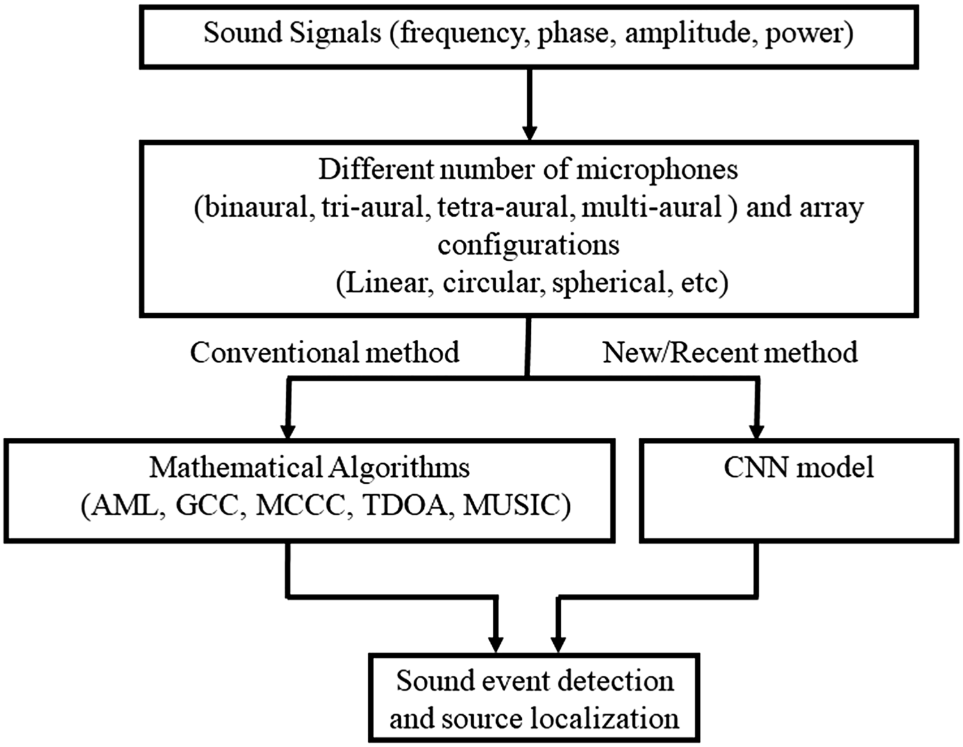
Sound Signals (frequency, phase, amplitude, power)
Different number of microphones (binaural, tri-aural, tetra-aural, multi-aural) and array configurations (Linear, circular, spherical, etc)
Conventional method
New/Recent method
Mathematical Algorithms (AML, GCC, MCCC, TDOA, MUSIC)
CNN model
Sound event detection and source localization

Fig. 3 The sound signal is sensed by a microphone array. Frequency, phase, amplitude and power spectrum are primary features of the detected sound signal. The microphone array can have different configurations like linear, circular, spherical, cuboidal, etc. and the number of microphones can differ from system to system. In the conventional method, sound signal features from all microphone channels are then analyzed using various mathematical algorithms like Approximate Maximum Likelihood (AML), Generalized Cross Correlation (GCC), Multichannel Cross Correlation (MCCC), Time Difference of Arrival (TDOA), Multiple Signal Classification (MUSIC). Using these algorithms, we can detect the sound event and localize the source of the sound. New methods suggest the use of CNN models which works like the human brain and automatically extracts features from raw sound signal and gives sound event detection and location of source as output

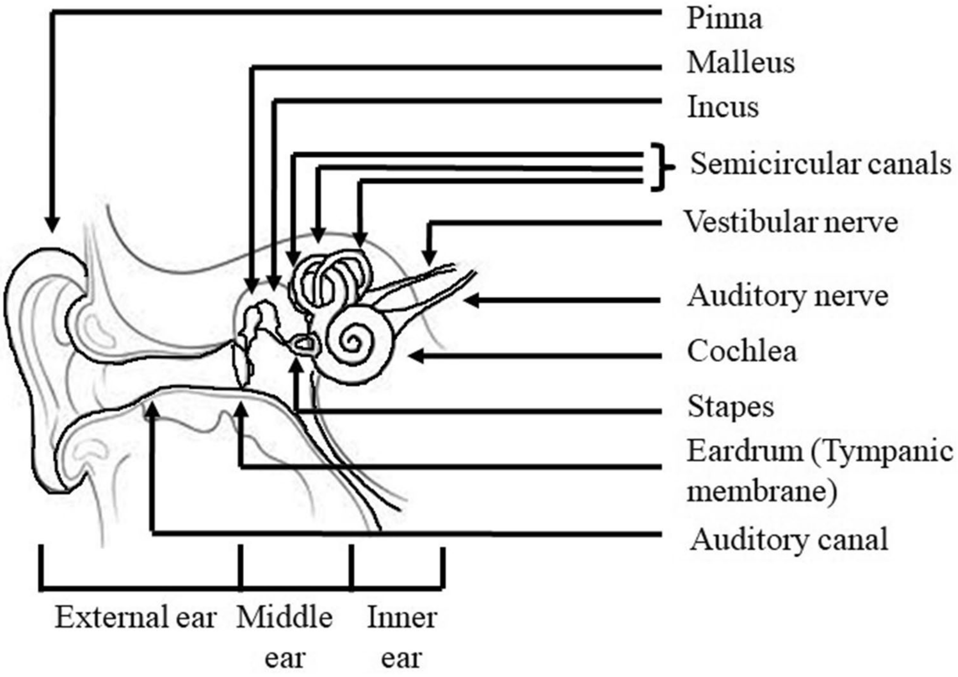

Fig. 4 The anatomy of the human ear is divided into three parts outer, middle and inner ear. Pinna and auditory canal form the outer ear. The eardrum, malleus, incus, stapes, and oval window forms the middle ear. The cochlea, auditory nerve and vestibular nerve are parts of inner ear

ears from the same sound source. It may happen that the sound source is closer to one of the two ears, so that the ear may receive a higher intensity signal than the other because of the head acting as a shadow for the other ear. ITD refers to the time difference of arrival of the audio signal at the left and right ear from the same sound source. It is known that the average distance between the human ear is about 20-22 centimeters approximately. In case if the sound source is coming from the left side of your head, then it reaches the right ear after 0.6ms as compared to the left ear. IPD refers to the difference in phase of the audio signal reaching each ear. IPD is derived from ITD and the frequency of the sound wave. A mathematical model of the human auditory system can be formulated based on time-frequency analysis. Figure 5 is a block diagram explaining the mathematical model of the human auditory system in which the first element is the microphone which converts the audio pressure wave into an analog electrical signal. Microphones provide an omnidirectional response and are sensitive to all frequencies arriving from all angles Smith et al. [2]. The combined effect of the outer and middle ear in the human auditory system performs bandpass filtering with a frequency not more than 80Hz. This effect is responsible for preprocessing the sound signal. Middle ear filtering mimics the middle ear which transforms eardrum vibrations into stapes motion using bandpass filtering. There are different types of models to mimic inner ear function which mainly involves either the use of Gammatone filters or Dual response Non Linear filter (DRNL filter). Gammatone filters divide the time-domain audio signal based on different frequency bands for frequency spectrum analysis. It simulates frequency-selective properties of the human basilar membrane. As per Jepsen et al. [3], the basilar membrane of the inner ear has been represented as a DRNL filter or a set of Gammatone filters. Inner hair cell converts mechanical energy to nerve signals. Inner hair cell is modeled by extracting envelope of the output of individual Gammatone filters. It is modeled as half-wave rectification followed by a first-order low pass filter with a cutoff frequency of 1KHz. The signal is then transformed as intensity representation by squaring expansion. The adaptive properties of the auditory system are modeled using five nonlinear feedback loops with different time constants. The feedback loops consist of low pass filters and division operations. Therefore, adaptation loops perform as fast temporal variations. Modulation processing of the human auditory system is represented as a first-order low pass filter with a cutoff frequency at 150Hz. This is followed by a modulation filter bank. The autocorrelation function is used for predicting the pitch of the audio signal. Rate map simulates auditory firing rate map. The Medial Olivo Cochlear processor (MOC) functions as auditory neurons. Input to MOC processor is time frame frequency representation signal from ratemap. ITD, ILD and IPD are known as auditory cues details of which are covered in coming sections. Cross-Correlation Function (CCF), ITD, ILD are computed for sound source localization.

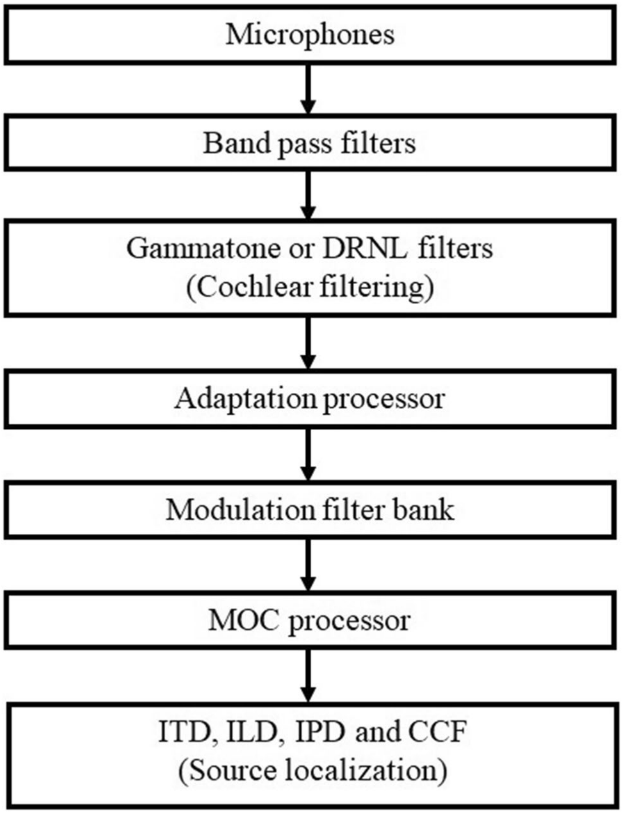
Microphones
Band pass filters
Gammatone or DRNL filters (Cochlear filtering)
Adaptation processor
Modulation filter bank
MOC processor
ITD, ILD, IPD and CCF (Source localization)

Fig. 5 Human auditory system model in which microphone acts as outer ear, bandpass filters, Gammatone filter and Dual response Non Linear filter (DRNL) filters function as an auditory canal for cochlear filtering, adaption processor works as feedback loops and performs auditive adaption as done by the human brain, modulation filter bank identifies various modulation frequencies, Medial Olivo Cochlear (MOC) processor depicts human brain neurons which finds out Interaural Time Difference (ITD), Interaural Level Difference (ILD), Interaural Phase Difference (IPD) and Cross-Correlation Function (CCF) for sound source localization

4 Some Conventional SSL Algoritms

4.1 Music

MUltiple SIgnal Classification algorithm is used for SSL. The number of sound sources can be one or more. There are let's say M sensors that receive time-delayed signals from sound sources with reference to a particular sensor. Matrix X is a function of incident signal, center frequency, incident signal angle and number of sensors M. Matrix X represents the received sound signals. The auto-correlation matrix R_{xx} of X finds the correlation between rows of X. Number of eigenvectors that are associated with signal subspace is equal to the number of sources. If the number of sensors and sound sources is known, matrix $U_n \epsilon C^{MXM-D}$ can be formed. U_n consists of a set of eigenvectors associated with noise subspace. Matrix U$_n$ consists of eigenvectors whose eigenvalues λ_{min} are variance of noise. λ_{min} occurs in clusters which decreases when more data is processed. Steering vectors corresponding to Difference of Arrival (DOA) are present in signal subspace and are orthogonal to noise subspace. Hence, we have $a^H(\hat{\theta})U_n U_n^H a(\hat{\theta})$ for $\hat{\theta}$ corresponding to DOA of multi-path component. DOA's can be known by locating peaks of MUSIC spatial spectrum

$$P_{MUSIC}(\hat{\theta}) = \frac{1}{a^H(\hat{\theta})U_n U_n^H a(\hat{\theta})} \tag{1}$$

4.2 TDOA

Time for sound signal to reach the microphone is calculated and speed of sound signal is measured. The difference in arrival of sound signal reaching both the microphones is used to calculate the distance of sound source from microphones.

$$\Delta d = c^* \Delta t \tag{2}$$

where c is speed of light, δt is difference in arrival times at microphones.

$$\Delta d = \sqrt{(x_2 - x)^2 - (y_2 - y)^2} - \sqrt{(x_1 - x)^2 - (y_1 - y)^2} \tag{3}$$

where (x_1, y_1) and (x_2, y_2) are known positions of beacons. By using non-linear regression, equation is converted to form hyperbola. After calculating many hyperbolas, SSL can be done by finding the intersection.

4.3 GCC-PHAT

To estimate TDOA, the delay between the cross-correlation between two signals should be maximum. Phase transform GCC increases its robustness.

Let x_i and x_J be two signals

$$\hat{G}_{PHAT}(f) = \frac{X_i(f)[X_j(f)]^*}{|X_i(f)[X_j(f)]^*|} \tag{4}$$

where $X_i(f)$ and $X_j(f)$ are Fourier transforms of two signals and $[]^*$ is complex conjugate. The TDOA for two microphones is given as

$$\hat{d}_{PHAT}(i,j) =_{d}^{argmax} (\hat{R}_{PHAT}(d)) \tag{5}$$

where $\hat{R}_{PHAT}(d)$ is the inverse Fourier transform

4.4 AML

Maximum likelihood method measures distance R_i between anchor node P_i and unknown node P. R_i is a random variable subject to Gaussian normal distribution $N(\mu_i, \sigma^2)$, where μ_i is true distance between P_i and P and σ^2 is fixed constant. Distances between different anchor nodes and unknown nodes are independent of one another. Let sample value of measured distance R_i be $r_i, i = 1, 2, ..., N$

The probability density functions of R_i are

$$f_i(t) = \left(\frac{1}{\sqrt{2\pi\sigma}} \right) e^{-(t-\mu_i)^2/2\sigma^2} \tag{6}$$

Then the likelihood function is as follows:

$$\begin{aligned} L(x,y) &= \Pi_{i=1}^n f_i(r_i) \\ &= \Pi_{i=1}^n \left(\frac{1}{\sqrt{2\pi\sigma}} \right) e^{-(t-\mu_i)^2/2\sigma^2} \\ &= (\frac{1}{\sqrt{2\pi\sigma}})^n e^{-(t-\mu_i)^2} \end{aligned} \tag{7}$$

where $\mu_i = \sqrt{(x_i - x)^2 + (y_i - y)^2}$, which is equivalent to

$$\sum_{i=1}^{n} =(x_i - x)\frac{\mu_i - r_i}{\mu_i} = 0 \tag{8}$$

$$\sum_{i=1}^{n} =(y_i - y)\frac{\mu_i - r_i}{\mu_i} = 0 \tag{9}$$

5 Sound Source Localization Using Multiple Microphones

The robotic ear should perform three main functions of the human ear that include Sound Event Detection (SED), Sound Source Localization (SSL) and sound type classification. Three dimensions azimuth, elevation and range are considered for SSL in 3D space. Three principle features that can be used from an audio signal are frequency, phase (time delay) and amplitude. Various mathematical algorithms like Beamforming, MUSIC, TDOA, GCC, MCCC, Maximum likelihood method, etc. are applied for manipulating and analyzing the sound signal for SSL. These algorithms alone or in combination do not work well for binaural signals (as in the human ear) that is using only two microphones as sensors. Therefore, more than two microphones and different microphone array configurations like linear, circular, spherical, non-coplanar, etc. are studied and used for optimization in source detection. Conventional robotic ear design techniques focused on permutations and combinations of various mathematical algorithm selection and microphone array configurations to detect sound events and their exact location. Chen et al. [4] proposes 3 and 4 microphone array configurations using beamforming algorithms and localization algorithms based on time delay estimation. It uses AML which gives better results as compared to the time delay estimation method. The experimental results conclude that beamforming and time synchronization algorithms together yield good results. Fazenda et al. [5] proposes siren detection using acoustic source localization using two methods by time delay estimation done by generalized cross correlation and another method by sound intensity method using a 2D orthogonal acoustic probe. Two pairs of microphone arrays are used which concluded that time delay estimation along with GCC gives accurate results as compared to the sound intensity method. Currently, microphones use pressure sensor technology for the detection of sound but Zhou et al. [6] proposes a novel microphone that detects sound by sensing acoustic particle velocity using nanofibres which can give bidirectional sound detection. Hoshiba et al. [7] proposes design of Spherical Microphone Array System (SMAS) for sound source detection. The proposed model is designed for Unmanned Aerial Vehicle (UAV). Song et al. [8] proposes sound direction detection using condenser microphone array. Fazenda et al. [9] uses coincident microphone technique for SSL. Hoshiba et al. [7], Song et al. [8] and Fazenda et al. [9] proposes different types of microphone array configurations for sound source detection and localization. With the development of machine learning and deep learning algorithms, recent robotic ear designs focus on formulating different types of Deep Neural Networks (DNN) and CNN architectures which automatically extract features from sound signals and can detect sound events and perform source localization (SL). Using CNN, SSL is possible using binaural signals and hence, CNN-based robotic ear can work using two microphones only. CNN-based SSL can be done either using classification or by regression. Chakrabarty et al. [10] proposes DOA estimation using CNN network. Input to CNN model is phase component of short-time Fourier transform (STFT) of the input signal. In the proposed method CNN architecture was designed which consists of three convolutional layers followed by three fully connected layers. ReLU activation was applied for convolutional as well as fully connected layers and softmax activation was applied at the final layer. Unlike conventional CNN architecture, the pooling layer was not used. The drawback of Chakrabarty et al. [10] is overcome by Li et al. [11]. A method involving CNN along with Long Short Time Memory (LSTM), which gives better performance in a noisy environment and is robust to change in the microphone array has been proposed by Li et al. [11] and Luet al. [12]. Three convolutional layers followed by one LSTM layer and then a fully connected layer wew used by Li et al. [11]. The activation function used is ReLU and softmax activation is applied at the final layer. Sasaki et al. [12] uses a spherical microphone array which consists of 16 microphones for azimuth as well as elevation detection based on MUSIC algorithm. Grondin et al. [13] proposes four-microphone array configuration using Convolutional Recurrent Neural Network (CRNN) for azimuth and elevation localization. It uses two CRNN in parallel to perform SED and SSL each. For SSL, the proposed CRNN architecture consists of two feed-forward layers followed by Bidirectional Gated Recurrent Unit (Bi-GRU) layer which is then connected with fully connected layer. ReLU activation function is used and hyperbolic tangent function is used at the final layer. Adavanne et al. [14] proposes Sound Event Detection and Localization (SELD) using CRNN. The proposed model uses phase and magnitude components separately from each channel and hence, is not dependent on array configuration. It can detect multiple DOA, it is robust to unknown DOA and can be used in any type of microphone array structure. It proposes a CRNN based neural network known as SELDnet. Table 1 summarises all conventional methods of SSL.

6 Binaural Sound Source Localization Using Conventional Algorithms

Human auditory system-based binaural localization primitively used conventional algorithms implemented on binaural signal cues and monaural spectral cues. Presented below are some of the systems based on conventional methods for localizing sound sources. Raspaud et al. [15] uses joint estimation of ILD and ITD for binaural SSL. It proposes a model in which Head Related Transfer Function (HRTF) is not required as it performs combined binaural cues azimuth estimation. The proposed model uses the CIPIC database. Li et al. [16] proposes a Bayes rules-based hierarchical system for binaural SSL. It uses binaural cues ITD and IID and a monaural cue for 3D localization. Commands (Com), CNN and Authentic Sound Effects (ASE) datasets were used in simulations. The simulations were conducted using different combinations of cues listed as DIR, TRA, TOG, MFCC, MON. The best results are produced by DIR and TRA with an error rate of less than 0.1 %. MON showed the highest error rate of 38.7 %. May et al. [17] designed a probabilistic model for binaural SSL using ITD and ILD for azimuth estimation. GMM is used to evaluate the joint binaural features and perform source localization in a probabilistic way. It could detect multiple sound sources. The model does azimuth estimation at an error rate of 10 to 18 %. Zannini et al. [18] exploits ITD and ILD cues for binaural SSL. A reference dataset was built to establish a reference localization method known as a data lookup algorithm. CIPIC database

Table 1 Summary of literature review of design of robotic ear that is various techniques used for sound event detection and source localization. The table classifies all the given methods based on principle used for source localization, microphone array configuration, number of sound source detected and localization of sound in 3D space.

Methods	Principle technique	Microphone array configuration	Number of sound source	Accuracy (%)
Chen et al. [4]	AML and TDOA	3 and 4 microphone multiple array configuration	Single source and dual source detection	–
Fazenda et al. [5]	Time delay using GCC and Sound intensity estimation	Single array with 4 microphones	Single source	88-95
Chakrabarty et al. [10]	CNN with phase component as input	4 and 8 microphone array	Single source	90
Li et al. [11]	CNN with LSTM	4 and 5 microphone single array	Single source	75
Sasaki et al. [12]	MUSIC	Spherical microphon array (16 microphones)	Dual Source	86
Grondin et al. [13]	CRNN	4 microphone array	–	87
Hoshiba et al. [7]	MUSIC	Hexagonal microphone array	Single source	95
Song et al. [8]	Time delay estimation	Condensor microphone array using 3 microphones	Single source	95
Adavanne et al. [14]	CRNN using phase and magnitude component	Generic array	Multiple source	86

of head-related impulse response (HRIR) was used to simulate the signal arriving at each ear. The accuracy was poor when the reverberation is above 0.1 sec. Parisi et al. [19] proposes cepstrum prefiltering based binaural SSL in reverberant environments. Cepstral prefiltering reduces the effect of reverberation in a closed environment and thus, improves performance. Pang et al. [20] suggests using binaural SSL based on reverberation weighting and generalized parametric mapping. Two-step SSL is used for azimuth estimation based on the parametric model and template matching. CIPIC database is used for training the model. The proposed model gives an accuracy of 68.53 %. Rodemann et al. [21] uses binaural and spectral cues for azimuth and spectral localization having error rate of 2.8 % and 12.3 % respectively. The disadvantage of this approach is that it requires long calibration cycles in 2D space. The system makes use of bionic ears for the effective preprocessing of sound signals. Wu et al. [22] proposes a model for 3D speech source localization using composite feature vector along with HRTF. The model extracts ITD, ILD and IPD from speech signals and creates a composite feature vector by combining it with HRTF features. The result suggests a 20 % improvement in localization performance. Dietz et al. [23] detects concurrent speakers from binaural signals. Short time constants (5 ms) were employed in the auditory model for robust localization of concurrent sources. The DOA estimation error was less than 5 degrees for 3 speakers in presence of noise. Chan et al. [24] proposes sound source detection for a robot that uses two microphones and the signal is given to a silicon cochleae thereby based on adaptive ITD-based intrinsic time-scale decomposition algorithms cross-correlation principle direction is identified. Woodruff et al. [25] also suggests multiple source localization in reverberant and noisy environment. The model performance azimuth estimation from binaural cues and segregation is performed on the basis of binaural and mono-aural cues. The model deals with both voiced and unvoiced speech and performs source localization in adverse conditions which have limited knowledge of binaural setup and training. The model possesses an accuracy of 97 %. Table 2 summarises all conventional SSL techniques which uses binaural sound signals.

7 Binaural Sound Localization Using CNN

Pang et al. [28] proposes a binaural SSL method using Time-Frequency-CNN (TF-CNN) detecting phase and level difference from binaural signals to simultaneously localize azimuth as well as elevate in the 3D space. TF-CNN consists of four convolutional layers with ReLU as activation function and softmax layer is used as output layer. Jiang et al. [29] proposes a fusion of CNN and DNN by extracting binaural cues like ILD and CCF. CNN was used for front-back classification and DNN for azimuth estimation which was then concatenated followed by the output layer. DNN consists of three hidden layers and ReLU as activation layer. CNN consists of two convolutional layers and ReLU as activation function. Both DNN and CNN are concatenated using a fully connected layer. The model was robust to noisy and reverberant environments and results showed 83 % accuracy. Xu et al. [30] has designed a binaural cascade of asymmetric resonators with fast-acting compression (CAR-FAC) cochlear system that analyses binaural signal. It uses the regression CNN technique by extracting CCF and IPD. It proposes a deep CNN network that has two convolutional layers, two pooling layers and one fully connected layer.

Table 2 Summary of various techniques used for sound event detection and source localization based on binaural signals. The table here compares all methods of SSL and their error rates which uses conventional algorithms.

Method	Features	Principle	Angle	Error rate (%)
Li et al. [16]	ITD, ILD and spectral cue	Bayes Rule	Azimuth and elevation	0.4-38.7
Rodemann et al. [21]	IID and ITD	-	Azimuth and elevation	2.8 and 12.3
Raspaud et al. [15]	ITD and ILD	-	Azimuth	-
May et al. [17]	ITD, ILD and IPD	GMM	Azimuth	10-18
Zannini et al. [18]	ILD and ITD	Phase difference look up algorithm	Azimuth	
Dietz et al. [23]	ITD and IPD	GMM	Azimuth	–
Chan et al. [24]	–	Adaptive ITD with vision sensor	–	–
Woodruff et al. [25]	ILD and ITD	Spatial clustering method	Azimuth	–
Parisi et al. [19] }	ILD and ITD	Cepstrum analysis	–	—
Wu et al. [22]	ITD, ILD, IPD and spectral cues	GCC	Azimuth and elevation	–
Pang et al. [20]	IID and ITD	–	Azimuth	31.47

The convolutional layer has an absolute value function as activation function while the fully connected layer uses tanh function as activation function. The activation function of the output layer is set as linear. The model does azimuth estimation from -90 to 90 degrees with a 3.68 % error rate approximately. Ma et al. [32] proposes a model of robust binaural localization combining spectral source models and deep neural networks. The model uses target source and noise together using DNN and performs azimuth estimation. It makes use of ITD, ILD and CCF feature vectors. The model is robust to noise and reverberation environment and gives less than 5 % error rate. Opochinsky et al. [34] does binaural sound source localization based on a weakly supervised deep learning approach which can be modeled using only a few labeled samples and a larger set of unlabeled samples. It uses DNN architecture which consists of three fully connected layers. The network is optimized using ADAM-approximate. It does an azimuth estimation of 0 to 180 degrees with an error of less than 10 %. The training dataset is of 1920 samples. Liu et al. [31] has designed CNN network for binaural localization based on ITD. The proposed method uses two kinds of outputs: Grouped CCF and Encoded CCF. It designs two models: GCC net grouped and GCC net encoded to perform TDOA estimation and trains them in three types of environment: Anechoic room, multi conditional training and in realistic environments. The proposed CNN model has three convolutional layers followed by the ReLU activation function. The final fully connected layer is followed by the sigmoid activation function. It performs azimuth estimation between -80 to 80 with an error rate of 0.1 % to 0.25 %. Wang et al. [35] proposes binaural localization based on deep neural network and affinity propagation clustering in mismatched HRTF conditions. The training set is generated by HRTF recorded by the KEMAR dummy head using the CIPIC database. The model consists of three stages (a) Study of localization similarity between HRTF (b) Clustering analysis applied to HRTF to improve DNN model accuracy and (c) To improve the generalization ability of the DNN model. The proposed DNN model has three hidden layers and one output layer. It uses ReLU activation layer. ReLU is followed by dropout layer which then has softmax activation function. It performs azimuth estimation with 63 % accuracy. He et al. [27] performs multiple speaker detection and localization using CNN. It proposes a CNN network that has four convolutional layers with ReLU activation layer and fully connected layer along with sigmoid activation function. It takes GCC phase transform (GCC-PHAT, and GCC as input features and performs likelihood-based coding. The model estimated 3D sound-based localization i.e. azimuth, elevation and distance with 90% accuracy. Vecchiotti et al. [33] suggest an end to end binaural sound localization from raw waveform in both anechoic and reverberant environments. Two systems are proposed that differ in frequency analysis. The first one is auditory based using gamma tone filters while the second one is fully data-driven and uses CNN layers. The CNN network is called WaveLoc which consists of a gammatone filter bank followed by normalization. After normalization, two convolutional layers are used followed by the max-pooling layer and then ReLU activation function is employed. which estimates azimuth (-90 to 90 degrees) in anechoic condition with a 1.5 to 3 % error rate. Bianco et al. [36] proposes semi-supervised localization with deep generative modeling with variational autoencoders (VAE). VAE generates phase of relative transfer function parallel with DOE classifier using labeled and unlabeled HRTF samples. VAE-SSL method uses CNN in a reverberant environment. The proposed CNN model uses two convolutional layers along with max-pooling layer followed by two fully connected layers which use the ReLU activation function. VAE-SSL possesses higher accuracy as compared to CNN and steers response power phase transform (SRP-PHAT) systems. It performs a -90 to 90

Table 3 Summary of literature review of design of robotic ear that is various techniques used for sound event detection and source localization based on binaural signals. The table here compares all methods of SSL which uses CNN and their error rates

Method	Network	Features extracted	SSL detection	Error rate (%)
Youssef et al. [26]	MLP	ITD,ILD and IPD	Azimuth and elevation	1.22 and 2.06
He et al. [27]	Liklihood based NN coding	GCC-PHAT	Azimuth, elevation and distance	10
Pang et al. [28]	TF-CNN	ITD and ILD	Azimuth and elevation	10
Jiang et al. [29]	DCNN	ILD and CCF	Azimuth and back forth	16.57
Xu et al. [30]	Two digital cochleae regression CNN	IPD and spectral cues	2D (– 90 to 90 degress)	3.68
Liu et al. [31]	GCC-Net	TDOA CCF	Azimuth (– 80 to 80 degrees)	0.1 –0.25
Ma et al. [32]	DNN and spectral cues	ITD, ILD and CCF	Azimuth	<5%
Vecchiottiet al. [33]	WaveLoc-GTF and WaveLoc-Conv	–	Azimuth (– 90 to 90 degrees)	1.5 to 3
Opochinsky et al. [34]	Supervised deep learning	–	Azimuth (0 to 180 degrees)	<10 %
Wang et al. [35]	Classification DNN and clustering HRTF	ILD and CCF	Azimuth	37
Bianco et al. [36]	VAE-SSL	–	Azimuth (– 90 to 90 degrees)	2.9–4.2.

degree azimuth estimation having a 2.9 to 4.2 % error rate. Nguyen et al. [37] has designed an autonomous sensory-motor learning for sound source localization by a humanoid robot. It provides a procedure to collect and label audio and motor data for sensory-motor training and uses CNN to localize speech signal sources. The average SSL accuracy in both PAN and TILT is 5.6 degrees. The robot in their proposed model can automatically collect multi-model data and learn from the data. Youssef et al. [26] proposes binaural signal localization in a humanoid robot by extracting ITD and ILD cues based on a learning-based approach. It uses a multi-layer perceptron network with azimuth and elevation estimation having error rates from 1.2 % to 2 % and 2.6 % respectively. Choi et al. [38] uses CNN based DOA estimation using stereo microphones for drone. The model is trained using recorded speech utterances at 10 directions spanning from 0 to 180 degrees in steps of 20 degrees at every 5, 10 and 20 meters. It uses power level-based features to classify the deep learning model and has an error rate of 0.0412 0.135 MSE. Table 3 summarises SSL techniques that use CNN algorithms on binaural signals.

8 Conclusion

Simulation of the human ear is a difficult task as the human ear performs multiple functions like sound event detection, sound type classification, detection of pitch, modulation, tone and roughness of audio signal, speech signal recognition and classification and sound source localization. Presented in this review are different techniques for sound source localization which is one of the important functions of the robotic ear. Unlike humans who use binaural sound localization techniques, robotic ears initially used multiple microphones and different array configurations for SSL using both conventional localization algorithms as well as using CNN. But in order to reduce the complexity of hardware as well as computational load, human auditory system-based SSL techniques have emerged. These techniques earlier used conventional algorithms but now in the last decade with the development of machine learning and deep learning technologies binaural, SSL systems based on CNN have been developed which provides SSL from raw audio signals. Using CNN, audio signals require little or no preprocessing. In the review, it is observed that most of the systems provide only azimuth estimation and there are very few which can provide 3D localization (Distance, Azimuth and elevation). Also, the performance of systems lacks when dealing with real-life environments which include noise, echo and reverberations. Systems also lag when it comes to multiple source localization. SSL systems designed to provide accurate source localization and binaural SSL using CNN are the best-optimized techniques above all other techniques.

Acknowledgements Authors would like to thank University of Mumbai for the funding and K. J. Somaiya College of Engineering for all the facilities and infrastructure provided.

Author Contributions Conceptualization was done by Ninad Mehendale (NM). The formal analysis was performed by Dhwani Desai (DD) and NM. Manuscript writing- original draft preparation was done by DD. Review and editing was done by NM. Visualization work was carried out by DD and NM.

Funding Funding was received from University of Mumbai under Minor Research Project scheme ATD/ICD/2019-20/762, Project number:1070, Date:17 March, 2020.

Data Availibility Not applicableCode availability Not applicable.

Declarations

Conflict of interest Authors D. Desai and N. Mehendale declare that there has been no conflict of interest.

Ethical approval All authors consciously assure that the manuscript fulfills the following statements: 1) This material is the authors' own original work, which has not been previously published elsewhere. 2) The paper is not currently being considered for publication elsewhere. 3) The paper reflects the authors'own research and analysis in a truthful and complete manner. 4) The paper properly credits the meaningful contributions of co-authors and co-researchers. 5) The results are appropriately placed in the context of prior and existing research.

Consent to participate This article does not contain any studies with animals or humans performed by any of the authors. Informed consent was not required as there were no human participants. All the necessary permissions were obtained from Institute Ethical committee and concerned authorities.

Consent for publication Authors have taken all the necessary consents for publication from participants wherever required.

References

1. Council NR et al (2004) Hearing loss: determining eligibility for social security benefits. Springer, New York
2. Smith LS (2015) Toward a neuromorphic microphone. Front Neurosci 9:398
3. Jepsen ML, Ewert SD, Dau T (2008) A computational model of human auditory signal processing and perception. J Acoust Soc Am 124(1):422
4. Chen JC, Yip L, Elson J, Wang H, Maniezzo D, Hudson RE, Yao K, Estrin D (2003) Coherent acoustic array processing and localization on wireless sensor networks. Proc IEEE 91(8):1154
5. Fazenda B, Atmoko H, Gu F, Guan L, Ball A (2009) Acoustic based safety emergency vehicle detection for intelligent transport systems. In: 2009 ICCAS-SICE (IEEE), pp 4250–4255
6. Zhou J, Miles RN (2018) Directional sound detection by sensing acoustic flow. IEEE Sens Lett 2(2):1
7. Hoshiba K, Washizaki K, Wakabayashi M, Ishiki T, Kumon M, Bando Y, Gabriel D, Nakadai K, Okuno HG (2017) Design of UAV-embedded microphone array system for sound source localization in outdoor environments. Sensors 17(11):2535
8. Song KT, Chen JL (2003) Sound direction recognition using a condenser microphone array. In: Proceedings 2003 IEEE International Symposium on Computational Intelligence in Robotics and Automation. Computational Intelligence in Robotics and Automation for the New Millennium (Cat. No. 03EX694), vol 3 (IEEE), vol 3, pp 1445–1450
9. Fazenda B (2008) Localisation of sound sources using coincident microphone techniques. Proc Inst Acoust 29(7):106
10. Chakrabarty S, Habets EA (2017) Broadband DOA estimation using convolutional neural networks trained with noise signals. In: 2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) (IEEE), pp 136–140
11. Li Q, Zhang X, Li H (2018) Online direction of arrival estimation based on deep learning. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (IEEE), pp 2616–2620
12. Sasaki Y, Tanabe R, Takernura H (2018) Online spatial sound perception using microphone array on mobile robot. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (IEEE), pp 2478–2484
13. Grondin F, Glass J, Sobieraj I, Plumbley MD (2019) Sound event localization and detection using CRNN on pairs of microphones. arXiv preprint arXiv:1910.10049
14. Adavanne S, Politis A, Nikunen J, Virtanen T (2018) Sound event localization and detection of overlapping sources using convolutional recurrent neural networks. IEEE J Sel Top Signal Process 13(1):34
15. Raspaud M, Viste H, Evangelista G (2009) Binaural source localization by joint estimation of ILD and ITD. IEEE Trans Audio Speech Lang Process 18(1):68
16. Li D, Levinson SE (2003) A bayes-rule based hierarchical system for binaural sound source localization. In: 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, Proceedings.(ICASSP'03). Vol 5 (IEEE), pp V–521
17. May T, Van De Par S, Kohlrausch A (2010) A probabilistic model for robust localization based on a binaural auditory front-end. IEEE Trans Audio Speech Lang Process 19(1):1
18. Zannini CM, Parisi R, Uncini A (2011) Binaural sound source localization in the presence of reverberation. In: 2011 17th International Conference on Digital Signal Processing (DSP) (IEEE), pp 1–6
19. Parisi R, Camoes F, Scarpiniti M, Uncini A (2011) Cepstrum prefiltering for binaural source localization in reverberant environments. IEEE Signal Process Lett 19(2):99
20. Pang C, Liu H, Zhang J, Li X (2017) Binaural sound localization based on reverberation weighting and generalized parametric mapping. IEEE/ACM Trans Audio Speech Lang Process 25(8):1618
21. Rodemann T, Ince G, Joublin F, Goerick C (2008) Using binaural and spectral cues for azimuth and elevation localization. In: 2008 IEEE/RSJ International Conference on Intelligent Robots and Systems (IEEE), pp 2185–2190
22. Wu X, Talagala DS, Zhang W, Abhayapala TD (2015) Binaural localization of speech sources in 3-D using a composite feature vector of the HRTF. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (IEEE), pp 2654–2658
23. Dietz M, Ewert SD, Hohmann V (2011) Auditory model based direction estimation of concurrent speakers from binaural signals. Speech Commun 53(5):592
24. Chan VYS, Jin CT, van Schaik A (2012) Neuromorphic audio-visual sensor fusion on a sound-localising robot. Front Neurosci 6:21
25. Woodruff J, Wang D (2012) Binaural localization of multiple sources in reverberant and noisy environments. IEEE Trans Audio Speech Lang Process 20(5):1503
26. Youssef K, Argentieri S, Zarader JL (2012) A binaural sound source localization method using auditive cues and vision. In: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (IEEE), pp 217–220
27. He W, Motlicek P, Odobez JM (2018) Deep neural networks for multiple speaker detection and localization. In: 2018 IEEE International Conference on Robotics and Automation (ICRA) (IEEE), pp 74–79
28. Pang C, Liu H, Li X (2019) Multitask learning of time-frequency CNN for sound source localization. IEEE Access 7:40725
29. Jiang S, Wu L, Yuan P, Sun Y, Liu H (2020) Deep and CNN fusion method for binaural sound source localisation. J Eng 2020(13):511
30. Xu Y, Afshar S, Singh RK, Wang R, van Schaik A, Hamilton TJ (2019) A binaural sound localization system using deep convolutional neural networks. In: 2019 IEEE International Symposium on Circuits and Systems (ISCAS) (IEEE), pp 1–5
31. Liu H, Yuan P, Yang B, Wu L (2019) Robust interaural time difference estimation based on convolutional neural network. In: 2019

IEEE International Conference on Robotics and Biomimetics (ROBIO) (IEEE), pp 352–357

32. Ma N, May T, Brown GJ (2017) Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments. IEEE/ACM Trans Audio Speech Lang Process 25(12):2444
33. Vecchiotti P, Ma N, Squartini S, Brown GJ (2019) End-to-end binaural sound localisation from the raw waveform. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (IEEE), pp 451–455
34. Opochinsky R, Laufer-Goldshtein B, Gannot S, Chechik G (2019) Deep ranking-based sound source localization. In: 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) (IEEE), pp 283–287
35. Wang J, Wang J, Qian K, Xie X, Kuang J (2020) Binaural sound localization based on deep neural network and affinity propagation clustering in mismatched HRTF condition. EURASIP J Audio Speech Music Process 2020(1):4
36. Bianco MJ, Gannot S, Gerstoft P (2020) Semi-supervised source localization with deep generative modeling. In: 2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP) (IEEE), pp 1–6
37. Nguyen Q, Girin L, Bailly G, Elisei F, Nguyen DC (2018) Autonomous sensorimotor learning for sound source localization by a humanoid robot. In: Workshop on Crossmodal Learning for Intelligent Robotics in conjunction with IEEE/RSJ IROS
38. Choi J, Chang JH (2020) Convolutional Neural Network-based Direction-of-Arrival Estimation using Stereo Microphones for Drone. In: 2020 International Conference on Electronics, Information, and Communication (ICEIC) (IEEE), pp 1–5

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.