IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 22 July 2024, accepted 30 July 2024, date of publication 2 August 2024, date of current version 13 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3437456

RESEARCH ARTICLE

An Approach for Single-Channel Sound Source Localization

KARIM YOUSSEF, (Member, IEEE), JULIEN MOUSSA H. BARAKAT, (Senior Member, IEEE), SHERIF SAID, (Member, IEEE), SAMER AL KORK, (Member, IEEE), AND TAHA BEYROUTHY, (Member, IEEE)

College of Engineering and Technology, American University of the Middle East, Egaila 54200, Kuwait

Corresponding author: Julien Moussa H. Barakat (julien.barakat@aum.edu.kw)

ABSTRACT Sound source localization for machines has been studied in microphone array and binaural paradigms in most cases, while much less work has been done in the single-microphone or monaural paradigm. This paper addresses this task and presents a system designed to classify azimuths of a speech-emitting source with respect to a binaural receiver, however using only one of its ears. The system uses the spectrum second derivative approximation calculated on short duration frames and based on a bank of gammatone filters, in conjunction with a classifier artificial neural network. It is tested to explore its abilities and the influence of different parameters on its performances. True recognition rates and confusion matrices are reported in different evaluations studying the effects of the frame duration, filterbank size, silence elimination, generalization capabilities and source movement. Reported results show an ability to classify azimuths correctly up to a certain extent depending on the parameters used, with confusions occurring mostly with neighboring azimuths. The presented system can be built upon for more efficient localization of speech sources in both azimuth and elevation components.

INDEX TERMS Sound source localization, monaural, machine listening, machine learning, artificial neural network, sound features.

I. INTRODUCTION

In machine listening, sound source localization (SSL) allows for the determination of the position of a source emitting sound in an environment that can be constraining in terms of noise and acoustic conditions. SSL relies on the existence of a sound receiver in the environment, used in conjunction with sound signal and data processing stages that allow to extract position information from the received sound. Different paradigms exist for sound reception, and the most widely used consist of microphone arrays [1], [2], [3], [4] and binaural receivers [5], [6], [7], [8]. In the binaural context, sound reception takes place like in humans, and many works try to produce models of the human hearing mechanisms that lead to sound source localization. Indeed, in human hearing, several elements of the auditory system play different roles in sound source localization. While localization in the horizontal plane relies mainly on interaural time and level differences (ITDs and ILDs), these features are minimal in the median plane, and yet humans can localize sound sources in this plane [9]. Other features are therefore exploited by the human auditory system for the sound source localization task. Importantly, the pinna functions as a filter with a transfer function that depends on the direction and distance of the sound source. This allows it to code spatial attributes of the sound field into temporal and spectral attributes [9], [10]. Alongside the pinna, the head and torso of a listener affect the sound reaching a listener's ears and Head-Related Transfer Functions (HRTFs) describe this filtering effect and depend on the source's azimuth and elevation with respect to the listener [11]. In addition, HRTFs vary depending on the person [12].

In the literature, work done on localization with a single microphone is limited and sparse. A literature review was made in [13] where it was mentioned that typically more than one microphone is used, and actually one

The associate editor coordinating the review of this manuscript and approving it for publication was Ines Domingues.

work on single-microphone localization was reported [14]. Additionally, a survey was shown in [15], addressing sound source localization with deep learning, where it was shown that sound source localization is usually performed with multichannel microphone arrays and no single-microphone localization technique was cited. We propose in this paper to explore the usage of monaural/single microphone features in the estimation of the sound source-receiver azimuth. The proposed approach relies on a framework employing, as a sound feature, the second derivative of the spectrum that was used in [16] to estimate the elevation and exploiting it with a neural network trained to classify source azimuths. Indeed, the shape of the outer ear is not neutral from a horizontal point of view and it can filter sounds differently with different azimuths. The sound-filtering effect of the pinna, head, and torso that provides monaural features, although more important in function of the elevation, can still be seen in function of the azimuth.

The proposed approach has different originalities and merits. Indeed, although the sound feature used in this work has been previously used in [16], its usage was for elevation estimation based on comparing feature vectors with reference feature vectors. This work exploits it for azimuth estimation, which, - to the best of our knowledge -, has not been done before, with artificial neural networks that offer potential for increased efficiency and robustness to recording conditions and information extraction. ANNs have also been used previously in localization, but, - to the best of our knowledge - not in monaural or single-microphone localization and not with the feature used here. Additionally, this work provides a framework that can be used with any asymmetrical sound-reflecting surface. The evaluations reported in this paper are done in a human-ear sound reception context, but the system relies on the reflective effects of the human pinna and can be used with any other surface that encloses a microphone and reflects sound signals differently with respect to different sound source locations. Such an approach can be extended to scenarios where sound reception is performed with microphones not necessarily placed inside human-like head and ear shapes. Different applications can be seen for it, as it reduces the number of receivers required for sound source localization, thus reducing physical and computational needs. While it can be applied in other contexts, the presented approach is evaluated in this paper with speech signals in single-source scenarios and an environment that is not highly reverberant. The proposed approach introduces a novel method for single-microphone, single-sound source localization, addressing several key challenges in the field.

- Hardware simplification: by utilizing only a single microphone, the method significantly reduces hardware complexity and costs compared to traditional multi-microphone systems. This simplification makes it more feasible for integration into various systems and applications, particularly those with space and budget constraints.
- Efficient signal processing: the method leverages a bank of gammatone filters to process incoming sound signals, mimicking the human auditory system's frequency selectivity. This biologically inspired approach enhances the precision of auditory feature extraction, which is critical for accurate sound source localization.
- Feature extraction and exploitation: a key step in the proposed system is the extraction of the spectrum second derivative. This step highlights dynamic changes and fine details in the sound signal, providing a robust set of features that improve localization accuracy. This technique distinguishes our method from other single-microphone approaches that may not capture such detailed spectral information. Furthermore, the processed features are input to a neural network, which excels at learning complex, non-linear relationships in data. This integration allows the system to achieve an accuracy and adaptability across different conditions, outperforming traditional methods that rely solely on signal processing techniques.
- Modularity and adaptability: the approach provides a modular platform with tunable parameters, allowing customization for different applications. This modularity also enables the addition of further processing steps to enhance performance or adapt to various acoustic conditions. This flexibility ensures that the system can be optimized for specific use cases, such as robotics, telepresence, or wearable devices.
- Computational efficiency: designed for efficiency, the method minimizes the amount of data to be processed and computational requirements, making it suitable for real-time applications. The reduced computational load is particularly beneficial for embedded systems and portable devices, ensuring swift and reliable performance.

By combining the simplicity of single-microphone use with advanced signal processing and machine learning techniques, the proposed method provides a significant advancement in the field of sound source localization. It offers a cost-effective, accurate, and adaptable solution for a wide range of applications, setting a new benchmark for single-microphone localization systems.

The rest of the paper is organized as follows: Related work is shown in Section II. The designed approach is presented in Section III, and evaluated in Section IV. The outcomes are discussed in Section V and a conclusion ends the paper.

II. RELATED WORK

In the literature, the amount of work aiming to perform sound source localization from one sound channel has been limited and sparse. While sound source localization was made mainly with microphone arrays, single-channel sound source localization was made mainly in a monaural context. With an examination of both contemporary machine learning and deep learning approaches as well as conventional propagation models, extensive classification of sound source

localization methods was discussed in [13]. It emphasizes the use of artificial intelligence, mathematical correlations, and physical phenomena in locating sound sources. The study also looks at potential directions for acoustic detection and localization in the future, with the goal of being a useful tool for choosing the best approaches in this area [13]. However, a thorough investigation of the number of microphones used reveals one work from this review that relied on a single microphone. In [14], the incident angle of sound by using an ''artificial pinna'' and a single microphone is determined by imitating the directional sound changes seen in the human outer ear [14]. The method predicts sound distributions and the direction-dependent changes caused by the pinna by using a machine learning methodology. Based on experimental results, a variety of sounds, including speech and barking, can be accurately localized.

However, relying on a single microphone array rather than a single microphone, in [17], a smart device is used to handle the simultaneous localization of several audio sources. The suggested method, Symphony, makes use of the direction-of-arrival (DoA) to identify source locations and the geometric arrangement of microphones to identify signal connections. Symphony employs a coherence-based module to detect signals from the same source and geometry-based filtering to separate signals from distinct sources. More recently, microphone arrays were used in a variety of studies and in different contexts. The work presented in [18] consisted of a microphone array system used for far-field sound source localization based on time difference of arrival and frequency division. The cocktail party problem, which consists of tracking and localizing specific sources among multiple, has been addressed in [19]. It relied on audio-visual features used in conjunction with a mobile robot equipped with a microphone array and capable of moving to gather better sound signals from the sources of interest.

In [16] and [20], azimuth and elevation estimation methods were proposed. They relied on gammatone filterbanks to provide inputs for monaural or binaural processing methods. Monaural and binaural cues were then exploited to reach the location information. Reported results showed the ability of the mentioned cues to reflect information on the elevation, even though elevation errors were considerably higher than azimuth errors. In [21], a single microphone SSL technique was proposed. Pyramidal horns were placed around a sound receiver, creating an asymmetrical shape that induces azimuth-dependent resonance and helps to discriminate between angles. Cepstrum-based features were used after signal reception to estimate the direction of arrival.

Self-localization of monaural microphones has been studied recently, concentrating on dipole sound sources. Due to their bidirectional sound emission pattern, dipole sources offer both special difficulties and chances for localization with just one microphone. In [22], a technique for a monaural microphone's indoor self-localization was presented. This is necessary for a number of location-based services. No matter how many devices are used, localization is accomplished on each one by creating two pairs of dipole sound fields. This is done using basic procedures that can be completed with a limited amount of CPU power and orthogonal detection of signals. Sound source localization achieved using a single microphone and additional components often employs an open-source reflector, which limits the range of the direction of arrival. In order to accomplish an omni-directional estimate, a sound source localization system with a closed surface reflector was studied by [23]. This study presented the findings from a perturbation technique analysis of such a system, together with a reflector shape evaluation. Thus, by comparing the systems utilizing open surface reflectors, the potential of the system and the hint to solve the inverse problem are explored.

The deep learning technique for identifying and localizing speech sources in challenging acoustic environments could also be employed using hearing aid microphones [24]. A neural network with a novel combination of residual and dense aggregation learning, as well as peripheral preprocessing on microphone inputs, all of which are inspired by the human hearing system, has been used in [24]. The result improves the gradient flow during training, which increases convergence speed and accuracy. Using both binaural and monaural microphone arrays, the proposed model by [24] shows promising results when it comes to joint speech source detection and localization; it even outperforms alternatives when using Short-Time Fourier Transform components. Due to the lack of spatial information in audio signals, traditional self-supervised learning employing monaural audio signals and pictures has difficulty differentiating similar-looking sound source objects. In [25], the problem of robots' autonomously identifying sound source objects in visual observations was addressed. By utilizing spatial information, the suggested approach presents self-supervised training with 360° photos and multichannel audio inputs [25]. Deep neural networks (DNNs) for vision and audio are used by the system to locate sound source items. Whereas the audio DNN confirms that sound source candidates actually produce sound, the visual DNN finds potential sound sources in an input image. The two DNNs are trained together using a probabilistic spatial audio model in a self-supervised fashion.

Using reverse correlation analysis (RCA), the outer ear contributes to the location of the front back and the height of the sound source has been examined by [26]. The magnitude spectrum of head-related transfer functions (HRTFs) from 73 participants with free-field localization behavior was integrated. Localization responses from participants are gathered both before and after the introduction of HRTF-modifying outer-ear implants. Based on the main characteristics identified by the RCA that affect localization responses, two monaural localization models are assessed. For bare ears, the models largely agree with free-field localization; however, when using modified HRTFs, the models overstate errors. Remarkably, RCA feature selection

lessens the influence of distorted HRTF elements on model accuracy and enhances alignment, indicating that it discloses crucial information for precise prediction. In a separate work, the effects of multi-band frequency compression on source localization and speech perception in monaural hearing aids were examined in [27]. Due to their inability to precisely localize sound sources, monaural hearing aids—which are intended for people with unilateral hearing loss—may make it difficult for users to understand speech in surroundings with complex auditory systems.

III. SYSTEM ARCHITECTURE

The proposed system relies on a framework for learning spectral attributes related to the sound source location and embedded in the sound signal received by a microphone. The operation of the system is based on steps shown in Figure 1. The signal is divided into short term frames, then silence removal is done with the sound activity detection shown in § III-B. Energy normalization is then carried out as shown in § III-C. Gammatone filtering is followed by feature extraction as shown in § III-D. This produces the feature vector which is input to an artificial neural network trained to output the azimuth, explained in § III-E.

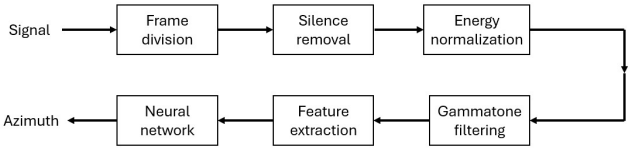

FIGURE 1. Consecutive steps involved in the workflow of the system, from signal reception until azimuth output.

A. DATABASE COLLECTION

A database consisting of sound sources emitted from multiple emitter-receiver azimuths is established. The database is designed to contain sound recordings from a single microphone/ear. Databases of binaural recordings can be used but recordings of ears can be processed separately, i.e. the system can rely on only one ear's recordings.

B. SOUND ACTIVITY DETECTION

The sound signals are first decomposed into frames of short duration. The energy of each frame is then estimated as:

$$E_i = \sum_{j=1}^{N} s_{i,j}^2 \tag{1}$$

where i is the frame number, N is the number of samples in the frame and $s_{i,j}$ is the sample j of the signal in this frame. The frame energy calculation serves in the process of eliminating frames with low activity compared to others, which can correspond to silences in the signal and do not provide information about the sound source position. This activity detection is performed based on the following thresholding process: the frame i is maintained if its energy E_i is higher that the threshold E_{Lth} calculated as follows:

$$E_{Lth} = E_{min} + L_{param} \times (E_{max} - E_{min}) \tag{2}$$

where E_{min} and E_{max} are respectively the lowest and highest frame energies over all the frames of the signal, and L_{param} is a parameter that can be modified to adjust E_{Lth}. This approach allows for the detection of silence segments with recording-dependent thresholding as thresholds that can be used in certain recordings do not apply to others.

C. ENERGY NORMALIZATION

After silence elimination, the energies of the signal frames are normalized to have a value of 1 each, by dividing each sample in the frame by the square root of the frame energy:

$$s_{i,j} = \frac{s_{i,j}}{\sqrt{E_i}} \tag{3}$$

This step maintains the dynamics within each frame's samples and its purpose is to reduce the sound loudness reduction caused by the head shadowing effect, which can bias the learning process for a machine learning system being trained on source locations from both sides of a head-like receiver.

D. FEATURE EXTRACTION

The sound signals are filtered with bank of G gammatone filters with center frequencies regularly spaced on the equivalent rectangular bandwidth scale as implemented in [28]. At the output of each gammatone filter, the energy E_i^g of the signal is calculated in dB as follows:

$$E_{i,g} = 20 \times log_{10} \sum_{j=1}^{N} s_{g,i,j}^2 \tag{4}$$

where $s_{g,i,j}$ is the sample j at of the frame i at the output of the gammatone filter g.

For the frame i, after calculation of all the gammatone filter energies, the second derivative of the spectrum is approximated as in [16] for each frequency band and time frame:

$$M_{i,g} = E_{i,g-2} - 2E_{i,g} + E_{i,g+2} \tag{5}$$

these channel-dependent second derivatives are concatenated in one vector to produce the vector V_i representing the frame i as:

$$V_i = [M_{i,1}, M_{i,2}, M_{i,3}, \ldots, M_{i,G}] \tag{6}$$

E. MACHINE LEARNING FRAMEWORK

Neural networks and Support Vector Machines (SVM) were compared as classifiers in [29] with different datasets to study the effect of dataset size and class number. Advantages were found for the neural network over SVMs. Also, in another classification task, MLPs were compared to other deep neural network architectures, ensemble algorithms, and SVM and

were found to outperform them [30].In another work, ANNs outperformed SVMs and random forests [31]. A last example is shown in [32] where ANNs were found to outperform other classifiers. Neural networks have been widely used in the field of sound source localization, where the task has been formulated as classification or regression [15]. Indeed, they can adapt to different acoustic conditions and extract relevant information from data [13]. Although different classifiers exist, and even within the same classifier family, different architectures and training methods exist, and although different classification tasks and the nature and amount of data used can give advantages to classifiers over others, for the purpose of source location classification based on received sound signals, it was decided to pursue with MLPs in this work as a means for proving the possibility of extracting azimuth information from sound. Although other classifiers may be better at this task, MLPs remain among the best candidates as classifiers.

The feature vectors calculated in the previous step are exploited by a multi-layer perceptron (MLP) with H hidden layers and h_1 cells in the first hidden layer, h_2 cells in the second, etc. The MLP exploits the feature vectors after decomposing the database into training and testing parts. The process is elaborated as a classification task where each source azimuth represents a class and the MLP is required to provide outputs, allowing to classify the input vectors provided to it. It is important to note that in the evaluations reported in this study, $H = 2$, and $h_1 = h_2 = 50$. The training and testing of the MLP are performed as follows:

1) TRAINING

The training data is used to train the MLP with the TensorFlow platform[1] for a number of epochs *NBepochs*. The feature vectors for all frames from all azimuths are concatenated into one matrix. And in parallel to the feature vectors provided as inputs, the MLP receives a vector of outputs containing for each vector, its corresponding class. The training process further decomposes the training data into training and validation, with a percentage of Val_s used for the validation data split.

2) TESTING

Testing data consists of input feature vectors and real output classes generated in parallel with the input vectors calculation. That is, to each input feature vector corresponds a ground truth output from the database. The MLP receives all the testing data in one array and for each feature vector, generates a class estimation. And the confusion matrix M is filled at each cell location in row r and column c with the number of tests that are in reality from the class r and estimated as being from the class c. Additionally, the true recognition rate is calculated as the ratio of the number of tests correctly estimated by the total number of tests. The true recognition rate will be used as a metric for the evaluation of the system performances presented in the next section.

[1] Available: https://www.tensorflow.org/

IV. SYSTEM EVALUATION

As shown in Section III, the proposed approach has several parameters that can affect its performances, as follows:

- Frame duration: the proposed approach allows for changing the duration of each frame along which the feature vector V_i is extracted. Frame durations of a few tens of milliseconds are usually considered in speech processing. However, longer duration can be investigated in this study as it aims to localize sources, not recognize the speech or the speakers. It is important to note that longer durations result in fewer frames per recording and, thus, fewer feature vectors and training data for the neural network.
- Gammatone filter number: it is another parameter that can affect the filter's selectivity, feature vector dimensions, and thus the system's training.
- Silence removal energy thresholding: the silence removal is an important step in the system that can affect its performances. Indeed, the lower the threshold E_{Lth} of the frame energy for silence removal, the more likely frames with low energy, i.e., without speech or partially containing speech, will be used in the training. Such frames assign silence parts to specific azimuth classes without containing any real information about the azimuth, thus reducing the effectiveness of the neural network training.
- Neural network architecture: the used neural network has a number of inputs equal to the number of gammatone filters used, and a number of outputs equal to the number of azimuths to classify. Other components of its architecture are the number H of layers and the numbers h_1, h_2, ... of cells in the respective layers. These parameters can greatly affect the performances of the system.

The proposed approach has been evaluated with datasets of speech recordings, for its ability to classify sound source azimuths as explained in Section III.

A. USED DATABASE

The used database features speech recordings made with still sources at different source-receiver azimuth values. The receiver is the binaural human-like head of the humanoid robot SIG2 [33], [34]. The recorded speech signals originate from the TSP speech database [35]. This speech database contains utterances recorded in anechoic conditions, corresponding to sentences provided by the list of Harvard sentences [36]. The TSP speech database contains recordings with lengths varying between 1.34 s and 4.79 s, with an average length of 2.37 s. To each speaker, a total of 60 different sentences was assigned. 10 male speakers were selected for our recordings, and the entire set of sentences of each speaker among them was used. A loudspeaker was used to emit sounds, placed at each of the 13 positions

corresponding to the theoretical constant distance of 1.5 m and the theoretical 13 azimuth angles ranging between −60° and 60° with a step of 10°. Figure 2 shows the different positions from which sound was emitted and recorded. For each of the still positions, two sentences from each speaker were used. This takes a total of 26 sentences per speaker. The database also includes recordings with movement during speech utterance. Movements were made in a way to maintain an approximately constant speaker-receiver distance at 1.5 m and with azimuth angles changing between −60° and 60°, in both left-to-right and right-to-left directions. Thus, two movement recordings with the same speech are made for each speaker, taking his remaining 10 sentences. The robot head was placed in a large room, with the ear microphones at a height of 141.5 cm. A Roland UA-101 audio interface[2] was used, and Audacity 2.0.5[3] was used to export recordings as wav files, with a sampling rate of 48 kHz and 16 bits per sample. Speaker-receiver positions were measured and adjusted to comply with the theoretical positions using a NaturalPoint OptiTrack Motion Capture system,[4] that determines real-time ground truth information about the positions of visual markers according to a specific landmark. Thus, markers placed on the source and the receiver allow us to measure their relative placements and track them, whether still or moving. The actual positions were not exactly match the theoretical ones, in terms of speaker-receiver azimuth and distance. Slight errors took place, with approximately 0.12° in azimuth and 0.56 cm in distance in average, and maxima of 0.27° and 2 cm respectively. Table 1 summarizes the main information related to the database.

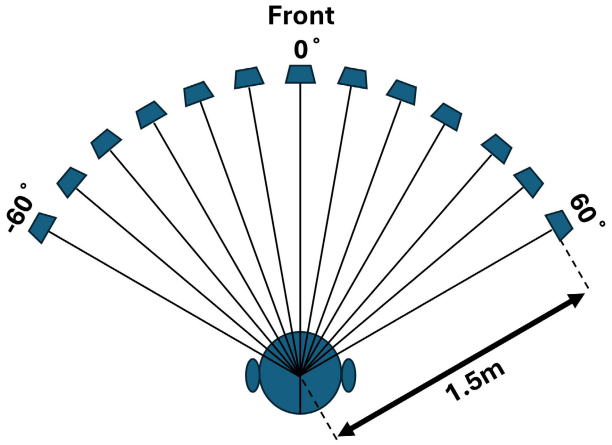

FIGURE 2. The 13 positions used to record signals.

B. RESULTS

From each recording, the first 40 seconds are used for the training, with $Val_s = 10\%$ of them used for cross-validation, and the next 15 seconds are used for the testing. A series of tests was performed to assess the ability of the system to classify azimuth angles and to study the effect of each parameter on its performance. Indeed, the system has several parameters that can affect its precision in feature extraction and its ability to use these features. In each evaluation, all system parameters were fixed to constant values except the one being studied, which was changed across a set of values.

[2]Available: http://www.roland.com/products/ua-101/
[3]Available: https://sourceforge.net/projects/audacity/
[4]Available: http://www.naturalpoint.com/optitrack/

TABLE 1. Summary of the used database.

Number of directions	13
Number of speakers	10
Number of sentences per speaker	26
Number of sentences per speaker and direction	2
Number of sentences per direction	20
TSP database Minimum sentence duration	1.34 s
TSP database Average sentence duration	2.37 s
TSP database Maximum sentence duration	4.79 s
Training duration from each direction	40 s
Testing duration from each direction	15 s

1) FRAME DURATION EFFECT

To study the effect of frame duration on the performances of the system, the training was done with the following configuration: $L_{param} = 0.01$, $G = 100$ and $NBepochs = 1000$, and with frame durations varying between 20 and 160ms. As Figure 3 shows, training does not get significantly improved for frame durations beyond 50ms. Also, no significant improvement in testing recognition rate is observed for frame durations above 100ms with around 47% of correct classifications.

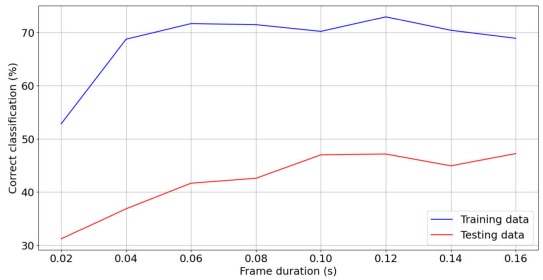

FIGURE 3. Frame duration effect on the neural network performances with the network being tested with the training data (including validation data) and testing data separately.

Figure 4 shows the evolution of the training in function of time. At each epoch, the neural network is tested with the training data, the validation data and the testing data separately. The figure shows a continuous improvement in the neural network training while no significant improvement is observed in its performances on validation and testing data after the 600^{th} epoch. Validation performances are slightly better than testing performances. This can be explained by the fact that validation data are extracted from training data after being shuffled. By this process, a validation feature vector can be associated with a time frame surrounded by time frames used for training and the sound content of this frame does

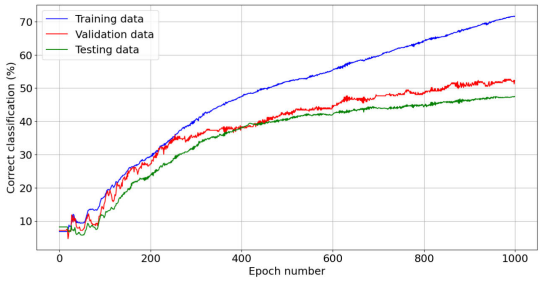

FIGURE 4. Evolution of the neural network training in function of time.

not vary significantly from its surroundings. Thus, the fact that validation data are better classified by the network than testing data, indicates that the network is probably learning, aside the sound source azimuth, information about the used signal itself. This is possible since the feature vectors used are based on a filtering of the sound signals.

Additionally, a confusion matrix obtained with the testing data with the same parameters used to generate the results in Figure 4 is shown in Figure 5. As the matrix shows, most classes have confusions with their nearest classes, with outliers in some cases. Also, for all the classes except the $-10°$ class, the class having the highest number of classifications is the real one.

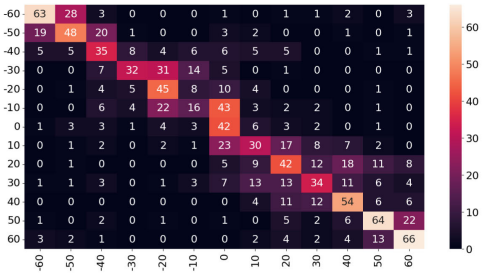

FIGURE 5. Confusion matrix visualization. Horizontal axis: classification results (degrees). Vertical axis: real angles (degrees).

2) GAMMATONE FILTERS NUMBER

To study the effect of the gammatone filterbank size on the system performances, the number of filters G was varied with the values of 20, 60, 100, 140 and 180 while keeping a frame duration of 100ms, $L_{param} = 0.01$, and $NBepochs = 1000$. Results of correct classification with training and testing data are reported in Figure 6.

Although the figure does not show a significant improvement in performance with testing data for G beyond 100 filters, it is important to note that the architecture of the neural network plays an important role in the results. Increasing G leads to increasing the dimension of the feature vectors provided to the neural network as inputs, while the hidden layers and cells of the network remain the same. Also, an increased dimension of the input layer may require a bigger database for improved training of the network. This is not taking place in the reported results as the number of

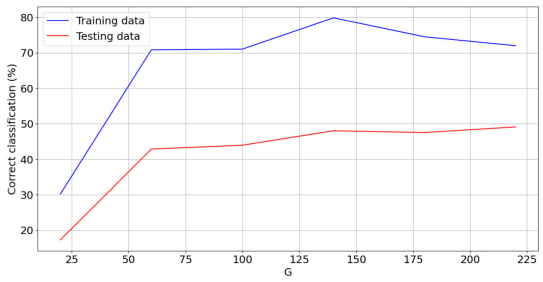

FIGURE 6. Gammatone filter number effect on the neural network performances with the network being tested with the training data and testing data separately.

samples provided as input is not changing, with the frame duration and the silence elimination threshold remaining the same.

3) SILENCE ELIMINATION THRESHOLD

The effect of silence elimination severity has been studied while keeping the frame duration of 100ms, $G = 100$ and $NBepochs = 1000$. The values of L_{param} used are 0, 0.01, 0.05 and 0.1. Figure 7 shows the correct classification results with tests performed on the training data and the testing data. It can be seen that as the silence elimination becomes more severe, correct classifications on training data improve while they deteriorate on testing data, indicating a loss of generalization capability of the neural network. This can also be associated with the reduction of the training dataset size available for the network, as can be seen in Figure 8. While 550 frames of 100ms are obtained from 55 seconds used in each direction with $L_{param} = 0$, this number becomes on average 81.3 with $L_{param} = 0.1$. Thus, a steep decrease in the database size is witnessed as the severity of silence elimination severity increases.

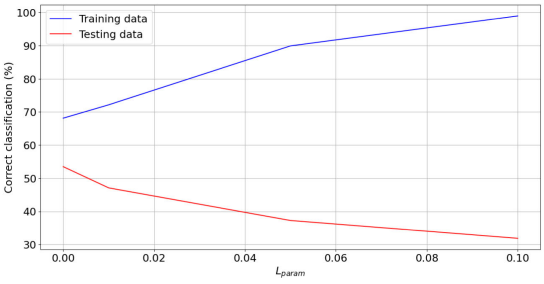

FIGURE 7. L_{param} effect on the neural network performances with the network being tested with the training data and testing data separately.

4) TRAINING AND TESTING WITH DIFFERENT AZIMUTHS

Another evaluation of the system was performed with training on data from the azimuths $-60°$, $-40°$, $-20°$, $0°$, $20°$, $40°$, and $60°$. Testing was done with the azimuths $-50°$, $-30°$, $-10°$, $10°$, $30°$ and $50°$. The training parameters were as follows: $L_{param} = 0.01$, $G = 100$, $NBepochs = 1000$. Figure 9 shows the confusion matrix obtained from this

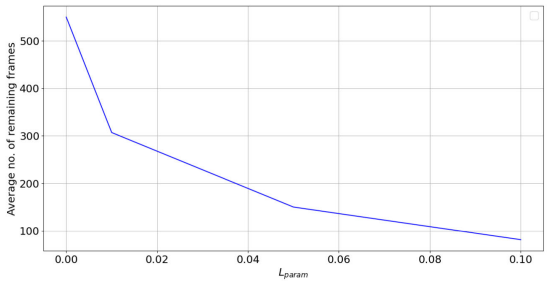

FIGURE 8. Average number of frames over all directions, remaining after silence elimination, for different values of L_{param}.

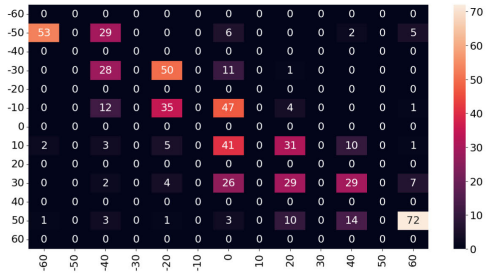

FIGURE 9. Confusion matrix visualization when testing with azimuths not included in the training. Horizontal axis: classification results (degrees). Vertical axis: real angles (degrees).

test. The matrix shows that for each of the testing angles, the highest confusion was with the surrounding azimuths included in the training. For instance, out of 97 tests done with the angle −50°, 53 were classified as the angle −60°, 29 from −40°, 6 from 0°, 2 from 40° and 5 from 60°. This is an expected results with sound recordings corresponding to neighboring azimuths generating patterns in their spectrum second derivative approximations with more resemblances than they do with other azimuths.

5) LOCALIZATION OF MOVING SOURCES

An additional test was made to investigate the ability of the designed system to track a person moving while speaking. The used database contains sound recordings with speech sources moving from −60° to 60° or in the opposite direction, while keeping a constant distance to the receiver and staying oriented towards it. Each recording corresponds to one person speaking continuously. Results are reported in Figure 10 and Figure 11 with two different speakers sounds moving from 60° to −60° and from −60° to 60° respectively. Each figure contains two parts. The upper part shows the raw frame-by-frame outputs of the neural network, and the lower part shows the raw outputs filtered with a median filter with a window length of 21. In both cases, the raw outputs show fluctuations and confusions in the classification, which are well smoothed by the median filtering. Although the database does not contain ground truth information about the exact speaker locations at each instant of time, the outputs clearly show a change of azimuth in function of time and allow to

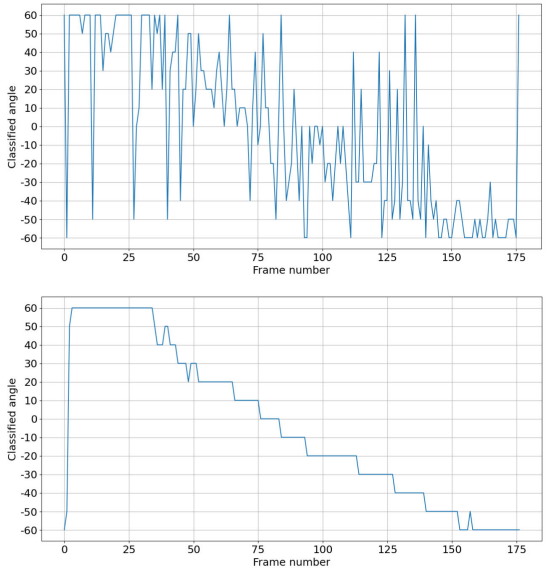

FIGURE 10. Outputs with recordings from Speaker 1 moving from 60° to −60°. Up: raw outputs. Down: outputs after median filtering.

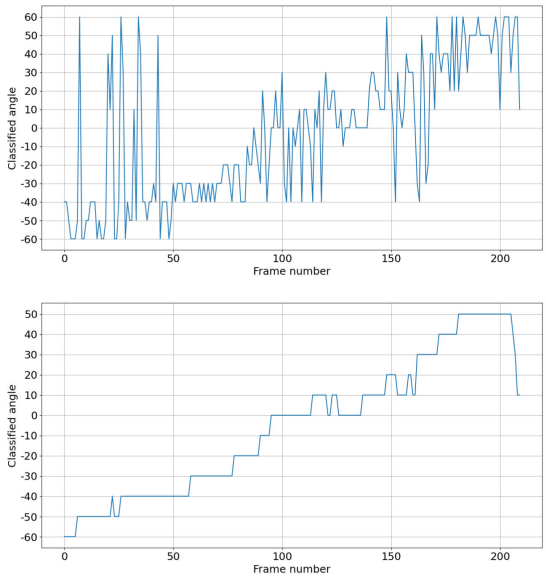

FIGURE 11. Outputs with recordings from Speaker 5 moving from −60° to 60°. Up: raw outputs. Down: outputs after median filtering.

track the movement efficiently with some confusions that can be reduced more by more filtering or other processing techniques.

V. DISCUSSION

As shown in the evaluations performed, different assumptions were made on the behavior and performances of the system. Indeed, like any other system of sound source localization, its parameters need to be addressed one by one to study their effects separately, then as a whole. It was important to vary each parameter in a wide range of possible values that it can take, and observe how the system reacts to each

value. Naturally, in all evaluations, and working with a classifier, the true recognition rate is a highly relevant metric to use, along with the confusion matrix. The results reported in Section IV serve to demonstrate the ability of a neural network to learn the azimuth of a speech source using signals captured by one microphone placed in an ear of a human-like head. From these results, it can be stated that the head and pinna filtering of arriving sound waves, and the head shadow effect, induce features in the waves reaching the eardrums that reflect the source's azimuth. This goes along with the results reported in [37], showing that near normal localization in both azimuth and elevation was achieved by listeners with unilateral deafness. As stated in [9], these listeners are capable of extracting and processing monaural spectral cues. The obtained results proved the following:

- a correct classification rate not far from 50% with data unseen in the training but from directions seen in the training, and with confusions occurring mostly with the classes neighboring the correct class. The system exhibits confusions with neighboring azimuths due to the fact that for a given azimuth, the ones surrounding it are the most likely to produce similar pinna filtering effects on the sound wave and thus induce features close to the features produced at it. For further azimuths, the pinna filtering has more differences and the system is less likely to confuse them with the real azimuth. Also, the fact that the database has been recorded in reverberant environment with noises inside affects the performances of the system and allows such confusions to arise more easily. This effect can be mitigated by introducing de-reverberation and de-noising steps in the signal processing prior to feature extraction. Such conclusions have been reached in previous studies where it was shown that signals recorded in less noisy and less reverberant conditions are better localized by a neural network [38]. This study shows that in the present conditions, one microphone allows to localize the sound source even without de-reverberation and de-noising, while showing the mentioned confusions in its estimations. Another way of mitigating the confusions is to allow the system to output one azimuth estimate for different consecutive time frames, by taking the mode of its frame-by-frame azimuth outputs. When the real azimuth is the dominant one among the outputs, taking the mode allows to better reach a correct estimate with fewer confusions, at the cost of taking longer to produce an output.
- the possibility of positively or negatively affecting the performances of the system by adjusting any of its parameters. For example, it was seen that with the shortest frame duration, performances were the weakest. However, in practical usages of the system, such as in human-machine interaction, the user is more likely to utter speech for long durations which improves the performances either by giving the possibility to use longer frames or by using short frames and exploiting their respective azimuth estimates in a majority voting for the correct azimuth for example.
- the ability of the system to classify sound signals which are in reality in classes not known in the training, with classes that are the closest to the classes used in the training.

While this system under-performs localization systems based on two or more microphones, and while better performances can still be achieved, the aim of the presented study, which is to explore the possibility of azimuth estimation with monaural features, was accomplished. One of the system parameters that were not tackled in the study, is the neural network architecture. Indeed, in all the reported tests, the number of hidden layers and cells remained the same. These parameters were reached after a series of trials and were seen to provide better results than others. Optimizing this architecture can be done for purposes of performance and computation load, but it was not in the objectives of this study. Some tests were done with a network with more hidden layers but they showed no significant advantage in terms of performances. However, other architectures and types of cells can be investigated in future work.

A comparison with microphone-array-based or binaural-based sound source localization systems favors those over the proposed approach in terms of accuracy. Indeed, the monaural context provides less information to process while in binaural and microphone-array paradigms, there is more freedom to conceive the localization with more possible features to extract and more freedom in configuring microphones that gives a higher chance of localizing sources accurately. An example of this is that in a binaural context, the monaural system proposed in this paper can be used twice, once with each ear, and their outputs can be combined into a single, more robust output. In binaural context, biologically-inspired time difference of arrival and time level of arrival exhibit a higher dependence on azimuth than the monaural features and thus can lead to better results. Also, in microphone-array contexts, techniques like beamforming and MUSIC allow to reach localization performances adjustable by the number of microphones and their locations in space. However, a comparison of a single-microphone approach like the one presented in this paper with binaural or microphone array-based approaches from other points of view exhibits some advantages for it.

- In terms of cost-effectiveness, the proposed approach uses one microphone, thus reducing the hardware demands.
- The proposed approach is also simpler to setup and maintain, and requires less space, making it advantageous in situations where space for sensors is limited.
- The proposed approach is more energy-efficient in terms of microphones to power and subsequent signal processing demands.
- The proposed approach is less prone to synchronization issues and potential phase mismatches that can occur in

TABLE 2. Comparison between different single-microphone sound source localization systems.

Approach	Reported Accuracy	Sound acquisition device	Sound feature	Localization system
Proposed approach	slightly less than 50% with the used settings, can be improved	one ear of a binaural receiver, could be any single microphone surrounded by a pinna-like shape	spectrum second derivative approximation	artificial neural network
Approach proposed in [21]	52% in average	a structure made of multiple pyramidal horns around a microphone	fundamental frequency	relation between horn radial length and fundamental frequency
Approach proposed in [14]	7.7° average error with speech (1-D)	a microphone partially enclosed with a vertical wall	moments of the signal power	Hidden Markov Model

systems where signals from more than one microphone are acquired simultaneously.

- In systems using more than one microphone, mathematical models in most cases consider that the different microphones used have a uniform frequency response characteristic. However in real implementation situations, this may not be the case and slight differences in frequency response characteristics can potentially affect the results. Thus calibration of the microphones can be needed, which can be time-consuming.
- Also in systems with more than one microphone, models often rely on a hypothetical placement and orientation of the different microphones. Misalignment can degrade the performances. A single-microphone localization system is less prone to this issue as there is one microphone to place instead of more.

These aspects thus make the proposed approach more advantageous in situations where considerations of space, cost-effectiveness, energy-efficiency and robustness to deployment problems weigh more than localization precision. Moreover, in terms of computational load, and for the specific single-microphone sound source localization approach shown in this paper, a comparison can be made with some recently proposed approaches and show that the proposed approach is computationally less demanding. For instance, a binaural sound source localization system was presented in [39] and was shown to be advantageous in comparison with other work and to be able to localize several sources. This system used gammatone filtering for both ear signals, cross-correlation between the two signals and a neural network framework. The system proposed in this paper uses gammatone filtering for the single signal available, and a neural network architecture less complex than the one proposed in [39], making it less complex computationally.

As for other single-microphone sound source localization approaches, a comparison of the proposed approach with two previously proposed approaches that were obtained and accessed is shown in Table 2. This comparison is made from different points of view, as a comparison of accuracy alone cannot be made, either for an approach being designed for a specific context that is not applicable in the context of the proposed approach in this paper or for lack of enough information to reproduce the whole system. Other criteria in which the proposed approach is more or less advantageous could have been used, but of less importance. For example, consider the size of the receiver, the computational complexity and power consumption, and the cost of training the system. From these comparisons, the approach proposed in this system shows several merits:

- its ability to be applied to other sound capturing contexts. The system has been evaluated in this study with recordings made in a human-like ear but the processing steps can be done with any other context and can give accurate results as long as the sound receiver has direction-dependent filtering. This not the case for the system shown in [21] as it was designed to work with a structure of pyramidal horns specifically.
- its ability to track moving sources, which was not reported in the other studies.
- its flexibility with all the parameters used, like the number of frequency channels considered, and time frame duration. This can allow to adjust the dimensionality of the data used and the complexity of the neural network depending on the processing capabilities of the platform where it can be implemented, with the awareness that this may affect performances. This has not been seen in [14] where even though the dimensions of the input data were not given, the feature calculation starts with the Short Time Fourier Transform which is potentially of a high dimension.
- its ability to localize speech sources with short durations, which was not studied in the other work.

VI. CONCLUSION

The study presented in this paper relied on recordings made with one ear of a binaural receiver to estimate the azimuth of a sound source emitting speech. Unlike most of the previous work relying on microphone arrays or binaural receivers and features for this task, monaural features allowed to provide information that was efficiently exploited by a MLP.

Future work will tackle different aspects. For instance, the effects of noises, room acoustics, receiver position in the room, speaker distance, and the presence of more than one source in the environment. Also, contexts of sound reception

other than human-like ears can be explored. Naturally, focus will be put on single-microphone systems with no symmetry in the receiver's design with respect to azimuth. Such geometrical configurations lead to uneven filtering of sound signals reaching the ears from different directions and thus can contribute in creating direction-dependent features. On the other hand, symmetrical receivers can cause similar sound filtering effects on symmetrical sound locations, which would lead to similar features and high confusions between these azimuths. Also, such systems can be used not only to estimate the azimuth, but also the elevation of the sound source, thus requiring databases covering more directions for their training. Additionally, the performances of these systems depend on their parameters and an optimization of parameters should be addressed for each environment and sound reception context.

REFERENCES

[1] J.-M. Valin, F. Michaud, and J. Rouat, "Robust 3D localization and tracking of sound sources using beamforming and particle filtering," in *Proc. IEEE Int. Conf. Acoust. Speed Signal Process. (ICASSP)*, Jun. 2006, pp. IV-841–IV-844.

[2] J.-M. Valin, F. Michaud, J. Rouat, and D. Letourneau, "Robust sound source localization using a microphone array on a mobile robot," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2003, pp. 1228–1233.

[3] D. Pavlidi, A. Griffin, M. Puigt, and A. Mouchtaris, "Real-time multiple sound source localization and counting using a circular microphone array," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 10, pp. 2193–2206, Oct. 2013.

[4] F. Grondin and F. Michaud, "Lightweight and optimized sound source localization and tracking methods for open and closed microphone array configurations," *Robot. Auto. Syst.*, vol. 113, pp. 63–80, Mar. 2019.

[5] T. May, S. van de Par, and A. Kohlrausch, "A probabilistic model for robust localization based on a binaural auditory front-end," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 1, pp. 1–13, Jan. 2011.

[6] T. May, S. van de Par, and A. Kohlrausch, "A binaural scene analyzer for joint localization and recognition of speakers in the presence of interfering noise sources and reverberation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 7, pp. 2016–2030, Sep. 2012.

[7] K. Youssef, S. Argentieri, and J.-L. Zarader, "A binaural sound source localization method using auditive cues and vision," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2012, pp. 217–220.

[8] K. Youssef, S. Argentieri, and J.-L. Zarader, "Towards a systematic study of binaural cues," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct. 2012, pp. 1004–1009.

[9] F. L. Wightman and D. J. Kistler, "Monaural sound localization revisited," *J. Acoust. Soc. Amer.*, vol. 101, no. 2, pp. 1050–1063, Feb. 1997.

[10] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*. Cambridge, MA, USA: MIT Press, 1996.

[11] H. Gamper, "Head-related transfer function interpolation in azimuth, elevation, and distance," *J. Acoust. Soc. Amer.*, vol. 134, no. 6, pp. EL547–EL553, Dec. 2013.

[12] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, Oct. 2001, pp. 99–102.

[13] G. Jekateryńczuk and Z. Piotrowski, "A survey of sound source localization and detection methods and their applications," *Sensors*, vol. 24, no. 1, p. 68, Dec. 2023.

[14] A. Saxena and A. Y. Ng, "Learning sound location from a single microphone," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2009, pp. 1737–1742.

[15] P.-A. Grumiaux, S. Kitić, L. Girin, and A. Guérin, "A survey of sound source localization with deep learning methods," *J. Acoust. Soc. Amer.*, vol. 152, no. 1, pp. 107–151, Jul. 2022.

[16] W. Chau and R. O. Duda, "Combined monaural and binaural localization of sound sources," in *Proc. Conf. Rec. 29th Asilomar Conf. Signals, Syst. Comput.*, vol. 2, Oct. 1995, pp. 1281–1285.

[17] W. Wang, J. Li, Y. He, and Y. Liu, "Symphony: Localizing multiple acoustic sources with a single microphone array," in *Proc. 18th Conf. Embedded Networked Sensor Syst.*, Nov. 2020, pp. 82–94.

[18] J. Zhao, G. Zhang, J. Qu, J. Chen, S. Liang, K. Wei, and G. Wang, "A sound source localization method based on frequency divider and time difference of arrival," *Appl. Sci.*, vol. 13, no. 10, p. 6183, May 2023.

[19] Z. Shi, L. Zhang, and D. Wang, "Audio–visual sound source localization and tracking based on mobile robot for the cocktail party problem," *Appl. Sci.*, vol. 13, no. 10, p. 6056, May 2023.

[20] C. Lim and R. O. Duda, "Estimating the azimuth and elevation of a sound source from the output of a cochlear model," in *Proc. 28th Asilomar Conf. Signals, Syst. Comput.*, Oct. 1994, pp. 399–403.

[21] K. Kim and Y. Kim, "Monaural sound localization based on structure-induced acoustic resonance," *Sensors*, vol. 15, no. 2, pp. 3872–3895, Feb. 2015.

[22] K. Arikawa, K. Hasegawa, and T. Nara, "Self-localization of monaural microphone using dipole sound sources," *J. Acoust. Soc. Amer.*, vol. 153, no. 1, pp. 105–118, Jan. 2023.

[23] J. Mori and S. Honda, "Monaural sound source localization using a closed surface reflector," in *Proc. 52nd Annu. Conf. Soc. Instrum. Control Eng. Japan (SICE)*, Jan. 2013, pp. 1094–1096.

[24] P. Goli and S. van de Par, "Deep learning-based speech specific source localization by using binaural and monaural microphone arrays in hearing aids," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 1652–1666, 2023.

[25] Y. Masuyama, Y. Bando, K. Yatabe, Y. Sasaki, M. Onishi, and Y. Oikawa, "Self-supervised neural audio-visual sound source localization via probabilistic spatial modeling," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2020, pp. 4848–4854.

[26] K. Balachandar and S. Carlile, "The monaural spectral cues identified by a reverse correlation analysis of free-field auditory localization data," *J. Acoust. Soc. Amer.*, vol. 146, no. 1, pp. 29–40, Jul. 2019.

[27] J. M. Katagi and P. N. Kulkarni, "Effect of multi-band frequency compression for improving speech perception in monaural hearing aids on source localization," in *Proc. IEEE North Karnataka Subsection Flagship Int. Conf. (NKCon)*, Nov. 2022, pp. 1–6.

[28] M. Slaney, "An efficient implementation of the Patterson–Holdsworth auditory filter bank," Perception Group—Adv. Technol. Group, Apple Comput., Inc., Apple Comput. Tech. Rep. 35, 1993.

[29] P. P. Conde and I. S. Carrillo, "Comparison of classifiers based on neural networks and support vector machines," in *Proc. 5th Int. Conf. Softw. Eng. Res. Innov. (CONISOFT)*, Oct. 2017, pp. 107–115.

[30] S. E. Jozdani, B. A. Johnson, and D. Chen, "Comparing deep neural networks, ensemble classifiers, and support vector machine algorithms for object-based urban land use/land cover classification," *Remote Sens.*, vol. 11, no. 14, p. 1713, Jul. 2019.

[31] E. Raczko and B. Zagajewski, "Comparison of support vector machine, random forest and neural network classifiers for tree species classification on airborne hyperspectral APEX images," *Eur. J. Remote Sens.*, vol. 50, no. 1, pp. 144–154, Jan. 2017.

[32] S. Heo and J. H. Lee, "Fault detection and classification using artificial neural networks," *IFAC-PapersOnLine*, vol. 51, no. 18, pp. 470–475, 2018.

[33] U.-H. Kim, K. Nakadai, and H. G. Okuno, "Improved sound source localization in horizontal plane for binaural robot audition," *Int. J. Speech Technol.*, vol. 42, no. 1, pp. 63–74, Jan. 2015.

[34] K. Youssef, K. Itoyama, and K. Yoshii, "Simultaneous identification and localization of still and mobile speakers based on binaural robot audition," *J. Robot. Mechatronics*, vol. 29, no. 1, pp. 59–71, Feb. 2017.

[35] P. Kabal, "TSP speech database," Dept. Elect., McGill Univ., Montreal, QC, Canada, Sep. 2002.

[36] IEEE Subcommittee on Subjective Measurements, "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoustics*, vol. TAU-17, no. 3, pp. 225–246, Sep. 1969.

[37] W. H. Slattery and J. C. Middlebrooks, "Monaural sound localization: Acute versus chronic unilateral impairment," *Hearing Res.*, vol. 75, nos. 1–2, pp. 38–46, May 1994.

[38] K. Youssef, S. Argentieri, and J.-L. Zarader, "A learning-based approach to robust binaural sound localization," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Nov. 2013, pp. 2927–2932.

[39] Q. Yang and Y. Zheng, "DeepEar: Sound localization with binaural microphones," *IEEE Trans. Mobile Comput.*, vol. 23, no. 1, pp. 359–375, Jan. 2022.

KARIM YOUSSEF (Member, IEEE) received the bachelor's degree in electrical engineering from Lebanese University, in 2010, the M.Sc. degree in intelligent systems and robotics from Sorbonne University, in 2010, and the Ph.D. degree from the Intelligent Systems and Robotics Institute, Sorbonne University, in 2013. He is currently an Assistant Professor with the Electrical Engineering Department, American University of the Middle East, Kuwait. His research interests include machine learning, robotics, sound specialization, sound signal processing, and other areas of artificial intelligence. He has a relevant track record of publications in peer-reviewed international conferences and journals in these areas.

JULIEN MOUSSA H. BARAKAT (Senior Member, IEEE) received the bachelor's degree in electrical engineering from Beirut Arab University, in 2003, the M.Sc. degree in microwave and optical communication from the University of Paul Sabatier, France, in 2004, and the Ph.D. degree from the University of Joseph Fourier France, in 2008. His work was done at the LETI, French Atomic Energy Commission (CEA), and the IMEP/Minatec, Grenoble, France, with a focus on integrated antennas functioning at 60 GHz and integrated over silicon on insulator for smart dust applications. Currently, he is an Assistant Professor with the College of Engineering and Technology, American University of the Middle East, Kuwait. He has authored/co-authored over 18 peer-reviewed publications in the areas of antenna design, optical fiber communications, and photonics. His research interests include high-impedance surfaces, SOI-integrated millimeter antennas, millimeter antenna measurement techniques, optical transmission, machine learning, the Internet of Things, and the development of optical elements. He is a member of OEA.

SHERIF SAID (Member, IEEE) received the B.Sc. degree in mechatronics engineering from October 6 University, Egypt, in 2009, the M.Sc. degree in mechatronics from Ain Shams University, Egypt, in 2015, and the Ph.D. degree in artificial intelligence from University Paris EST Creteil, France, in 2020. He is an Associate Professor with the Mechanical Engineering Department, American University of the Middle East, Kuwait, where he also serves as the Team Leader of the AUM Robotics Research Center. With a passion for interdisciplinary research, he has been immersed in academia since graduating. Through his research and teaching, he continues to contribute to the advancement of knowledge and innovation in the field of robotics and engineering. His research interest include machine learning, robotics, biometrics systems, biomedical, automatic control, and artificial intelligence. He has several publications in the field. He has also participated in international, regional, and national engineering events, conferences, and competitions.

SAMER AL KORK (Member, IEEE) received the Ph.D. degree in bioengineering from the University of Illinois at Chicago (UIC), USA, in 2009. Since then, he has held numerous academic positions, such as an Adjunct Assistant Professor with the Department of Biomedical Engineering and Information Systems, UIC; the ITT-Tech Institute, USA; and the ESIGELEC School of Engineering, Pierre Marie Curie University, France, from 2009 to 2015. In 2015, he joined the American University of the Middle East (AUM), Kuwait, where he is currently an Associate Professor and the Electrical Engineering Department Chair. He is also a Founding Member of the Robotics Center, AUM. He is a Leading Expert in areas of biomedical engineering, mobile and aerial robotics, robotic vision, artificial intelligence, and humanoid robotics supervising; and co-supervising multiple master's and Ph.D. students. He was the Lead Principal Investigator of two European projects (i-treasures and noba) with over a six-million-euro budget. He has authored and co-authored over 50 peer-reviewed publications and one book chapter on wearable technologies in biomedical and biometric applications. Since inception, he has been serving on the Steering Committee and the Technical Chair for the International Conference on Bioengineering for Smart Technologies (BioSMART). He is a regular reviewer of top-tier academic journals.

TAHA BEYROUTHY (Member, IEEE) received the M.Eng. degree in microelectronics from IMT Atlantique (Télécom-Bretagne), in 2006, and the Ph.D. degree in micro and nanoelectronics from the Grenoble Institute of Technology, in 2009. He is currently an Associate Professor of electrical engineering and the Dean of the College of Engineering and Technology, American University of the Middle East (AUM). Before joining AUM, he was an Assistant Professor and a Lead Member of the CIS Research Team, TIMA Laboratory, Grenoble. Later, he joined EASii-IC as a Senior ASIC Designer, where he focused on improving power efficiency and security in digital systems, through novel signal sampling schemes, and designing asynchronous cryptographic circuits. He has established the AI and Robotics Research Center, AUM, and has published numerous publications in micro and nanoelectronics, robotics, artificial intelligence (AI), and applied physics. His current research interests include the application of AI in STEM education, brain–computer interfacing, and machine learning in cybernetics.