Sound Localization: Feature Extraction and Machine Learning Approach

Michael Issa

San Diego State University
missa0773@sdsu.edu

May 2024

Abstract

Sound localization is a fundamental aspect of auditory perception, crucial for humans and animals to navigate their environment. This paper explores sound localization techniques, emphasizing feature extraction and machine learning for predicting the azimuthal direction of sound sources. Classical methods such as Interaural Time Difference (ITD) and Interaural Level Difference (ILD) provide foundational principles, while machine learning methods supplement it. We implemented various feature extraction algorithms, including ITD, ILD, spectral ITD, spectral ILD, and Mel-frequency cepstral coefficients (MFCCs), and trained a Random Forest Classifier on these features. Using a dataset of high-frequency sound recordings, our model achieved high accuracy in predicting sound source locations. These findings underscore the effectiveness of classical feature extraction coupled with modern machine learning for sound localization tasks. Further research may explore applications in noisy environments and real-time scenarios.

Sound localization is a fundamental aspect of auditory perception, enabling humans and animals to locate the source of sounds in their surroundings. Understanding how sound localization works in humans and other organisms provides insights into developing artificial systems capable of similar feats. In this paper, I investigate sound localization techniques, with a focus on feature extraction and machine learning for predicting the azimuthal direction of sound sources. I use relatively simple sound sound source localization features with machine learning techniques to classify high frequency sound sources.

Sound localization refers to the ability to determine the direction from which a sound originates. Humans and animals achieve sound localization through various mechanisms, including interaural time difference (ITD), interaural level difference (ILD), and spectral cues. Humans rely on the differences in arrival time and intensity between the two ears to localize sound sources accurately. We do this exceptionally well and with a high degree of robustness. When we are trying to translate this capability to artificial machines we want to individuate the relevant contributions the sound actually makes to our ability. We do not pick up the raw sound source but some modified version. These modified features are what we begin with.

Feature extraction plays a crucial role in sound localization, as it involves capturing relevant information from audio signals to facilitate spatial localization. In our project, we implemented several feature extraction algorithms, including ITD, ILD, spectral ITD, spectral ILD, and Mel-frequency cepstral coefficients (MFCCs). These algorithms analyze binaural audio signals to extract temporal and spectral cues that are indicative of sound source direction.

Interaural Time Difference (ITD) measures the time delay between the arrival of a sound wave at each ear. It can be calculated as follows:

Let t_L and t_R represent the arrival times of a sound wave at the left and right ears, respectively. Then, the ITD δt is given by the difference in arrival times:

$$\delta t = t_R - t_L$$

By comparing the phase differences between left and right ear signals, ITD provides valuable information about the azimuthal direction of sound sources. The relationship between ITD and azimuthal angle θ can be expressed as:

$$\delta t = \frac{D}{c} \cdot \sin(\theta)$$

where:

$$\begin{aligned} D &: \text{Distance between the ears} \\ c &: \text{Speed of sound in the medium} \\ \theta &: \text{Azimuthal direction of the sound source} \end{aligned}$$

Interaural Level Difference (ILD) quantifies the difference in sound intensity between the two ears. It reflects variations in sound pressure level caused by the spatial separation of the ears relative to the sound source. ILD can be calculated as follows:

Let P_L and P_R represent the sound pressure levels at the left and right ears, respectively. Then, the ILD ΔP is given by the difference in sound pressure levels:

$$\Delta P = P_R - P_L$$

ILD provides information about the relative loudness of the sound at each ear, which can be used to infer the azimuthal direction of the sound source. However, it is important to note that ILD alone may not be sufficient for accurate localization, as factors such as frequency-dependent attenuation and head-related transfer functions also play a role in sound perception.

Spectral Interaural Time Difference (ITD) analyzes frequency-specific cues in the spectral domain, allowing for more detailed localization information across different frequency bands. It can be calculated as follows:

Let $t_L(f)$ and $t_R(f)$ represent the arrival times of a sound wave at the left and right ears, respectively, for a specific frequency f. Then, the Spectral ITD $\delta t(f)$ is given by the difference in arrival times:

$$\delta t(f) = t_R(f) - t_L(f)$$

Spectral ITD provides valuable information about the azimuthal direction of sound sources across different frequency bands. The relationship between Spectral ITD $\delta t(f)$ and azimuthal angle θ for a specific frequency f can be expressed as:

$$\delta t(f) = \frac{D}{c} \cdot \sin(\theta)$$

where:

$$\begin{aligned} D &: \text{Distance between the ears} \\ c &: \text{Speed of sound in the medium} \\ \theta &: \text{Azimuthal direction of the sound source} \end{aligned}$$

Spectral Interaural Level Difference (ILD) analyzes frequency-specific cues in the spectral domain, allowing for more detailed localization information across different frequency bands. It can be calculated as follows:

Let $P_L(f)$ and $P_R(f)$ represent the sound pressure levels at the left and right ears, respectively, for a specific frequency f. Then, the Spectral ILD $\Delta P(f)$ is given by the difference in sound pressure levels:

$$\Delta P(f) = P_R(f) - P_L(f)$$

Spectral ILD provides information about the relative loudness of the sound at each ear across different frequency bands, which can be used to infer the azimuthal direction of the sound source. However, it is important to note that like ILD, Spectral ILD alone may not be sufficient for accurate localization, as factors such as frequency-dependent attenuation and head-related transfer functions also play a role in sound perception.

Mel-frequency Cepstral Coefficients (MFCCs) are widely used in speech and audio processing for feature extraction. They capture the spectral characteristics of audio signals by representing the short-term power spectrum of sound in a perceptually meaningful way. The computation of MFCCs involves several steps: Before calculating MFCCs, the input audio signal is pre-emphasized to emphasize high-frequency components. This can be achieved by applying a first-order high-pass filter:

$$s_{\text{pre-emphasis}}(n) = s(n) - \alpha \cdot s(n-1)$$

where $s(n)$ is the input audio signal at time n, and α is the pre-emphasis coefficient. The pre-emphasized signal is then divided into short frames of typically 20-40 milliseconds with overlap. Each frame is windowed using a window function such as Hamming or Hanning to minimize spectral leakage. I used a Hamming window. The power spectrum of each frame is computed using the Fast Fourier Transform (FFT), which converts the signal from the time domain to the frequency domain. The power spectrum is then passed through a bank of Mel filters, which are triangular filters spaced evenly in the Mel frequency scale. The Mel scale is a perceptually linear scale of pitches, designed to mimic the human auditory system's response to different frequencies. The output of each filter represents the energy in each frequency band. The logarithm of the energy in each filter output is taken to approximate the logarithm of the human auditory system's response to different frequencies. Finally, the Discrete Cosine Transform (DCT) is applied to the logarithmically transformed filterbank energies to decorrelate the features and extract the MFCCs. Typically, only the lower-order coefficients are retained as they contain most of the relevant information.

The extracted features provide essential cues for predicting the azimuthal direction of sound sources. ITD and ILD encode temporal and intensity differences, respectively, which are crucial for estimating the horizontal angle of sound arrival. Spectral cues, such as spectral ITD and ILD, offer additional information about the spatial distribution of sound energy across different frequency bands. MFCCs capture higher-level spectral features that complement traditional localization cues, enhancing the robustness of sound source localization algorithms.

To predict the azimuthal direction of sound sources, we employ a machine learning model trained on the extracted features. In our project, we utilized a Random Forest Classifier—a versatile ensemble learning algorithm capable of handling complex datasets with high-dimensional feature spaces. The Random Forest Classifier learns the relationship between the extracted features and the corresponding azimuthal direction labels from the training data, enabling it to make accurate predictions on unseen test data.

The Random Forest Classifier is well-suited for sound source localization tasks due to its ability to handle non-linear relationships and high-dimensional feature spaces effectively. By constructing multiple decision trees and aggregating their predictions, the Random Forest Classifier provides robust and accurate predictions of azimuthal direction based on the extracted features. Through hyperparameter tuning and cross-validation, we optimize the performance of the classifier to achieve high accuracy in predicting sound source locations.

The history of sound source localiztion is long and advancing at a rapid pace witht the advent of AI tools. A survey study by Jekateryńczuk and Piotrowski (2024) reviews classical and emerging methods in sound source localization. Classic methods for sound source detection and acoustic localization, such as Time Difference of Arrival (TDOA), Steered Response Power with PHAse Transform (SRP-PHAT), MUltiple SIgnal Classification (MUSIC), and Generalized Cross-Correlation (GCC), have been fundamental in this field. TDOA relies on measuring time delays between sound signals arriving at different microphones to triangulate the direction of the source. SRP-PHAT maximizes spatial power distribution to estimate direction, while MUSIC exploits spectral diversity for localization. GCC estimates time delays between microphone pairs in the time domain. These methods, though effective, face challenges like sensitivity to noise and reverberation or computational intensity. They form the basis for advanced AI-driven techniques, which aim to overcome these limitations while building on their principles for improved accuracy and robustness in real-world scenarios.

Contemporary methods in sound source detection and acoustic localization leverage advanced artificial intelligence (AI) techniques, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Convolutional Recurrent Neural Networks (CRNNs), Residual Neural Networks (ResNets), transformers, encoder-decoder networks, and hybrid neural network architectures. These methods address the limitations of classic techniques by harnessing deep learning capabilities to extract intricate features from audio data, enabling more accurate localization in diverse environments. For instance, CNNs analyze spectrograms to classify sound sources, while CRNNs combine CNNs and RNNs for sequential data analysis. Transformers excel in processing temporal sequences, enabling precise localization in both anechoic and reverberant environments. Encoder-decoder networks capture input features and decode them into desired output information, while hybrid architectures integrate both audio and visual modalities for comprehensive detection and localization. These AI-driven approaches offer significant advancements in performance, adaptability, and robustness, promising enhanced accuracy across military and civilian applications.

Usually the more advanced methods are needed in real world applications where noisy environments and high-stakes decisions place a high premium on success. However, I use a mixture of feature extraction techniques used in more traditional localization tasks and train a machine learning model with them. For the most part, the extracted features from the wav file make the biggest contribution to the success of the work. Given the features we extracted, I trained a random forest model, which performed exceptionally well on my tuning and test data. We had around 138,542 rows of data after extracting the features from a open dataset by Gardner and Martin (1994). The dataset includes various 1 second, high frequency sounds conducted in a controlled environment with a binaural microphone recording. The raw wav files were transformed into our training features. After training our model we got a tuning accuracy score of .98, and we had even better performance on our test dataset with a .99 accuracy. The accuracy metric is with respect to correctly classified azimuthal directions where this ranges form 0-360 degrees.

The incredibly good results speak to the fact that, in controlled environments with little noise, classical feature extraction and algorithmic techniques work exceptionally well in figuring out azimuthal source. Next steps for the future might include testing on noisy data or on uninterrupted chunks of sound.

In conclusion, sound localization is a multifaceted process that relies on the extraction of relevant features from audio signals and the application of machine learning techniques for accurate spatial localization. By leveraging algorithms such as ITD, ILD, spectral cues, and MFCCs, coupled with machine learning models like the Random Forest Classifier, we can develop sophisticated systems capable of accurately predicting the azimuthal direction of sound sources. These advancements have broad applications in fields such as robotics, virtual reality, and human-computer interaction, enhancing our understanding of auditory perception and enabling the development of intelligent audio-based systems.

MLA Bibliography Entries

1. Jekateryńczuk, Gabriel, and Zbigniew Piotrowski. ”A Survey of Sound Source Localization and Detection Methods and Their Applications.” *International Journal of Audiology*, vol. 39, no. 7, 2000, pp. 383–396. *PubMed Central*, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10781166/.

2. Gardner, Bill, and Keith Martin. ”HRTF Measurements of a KEMAR Dummy-Head Microphone.” *MIT Media Lab*, https://sound.media.mit.edu/resources/KEMAR.html.