Sound Localization & Source Separation for Robotics Applications

MingYang Lee, Yueyue Li

The Cooper Union for the Advancement of Science and Art, 41 Cooper Sq., New York, NY, 10003

Correspondence should be addressed to lee11@cooper.edu, yuekeira@gmail.com.

ABSTRACT
This project aims at improving current sound processing methods for robotics applications. The improved methodology involves the usage of a microphone array and allows the robot to differentiate and localize simultaneous sound sources to perform better in a complex sound environment. First, independent source signals from the recordings were extracted; then, the incidence angles of those sound sources were found. Several digital signal processing techniques are incorporated in the proposed method. Frequency-Domain Independent Component Analysis(FDICA) is used to extract independent sound sources from sound mixtures, and Time Delay of Arrival(TDOA) method is used to perform source localization. This paper will detail the implementation of signal processing technique using MATLAB and also the design of the octahedron microphone array. [add a sentence for result]

1 Introduction

This project aims at improving current sound processing methods for robotics applications. The improved methodology involves the usage of a microphone array that allows the robot to differentiate and to localize simultaneous sound sources. First, independent source signals from the recordings were extracted; then, the incidence angles of those sound sources were found. Several digital signal processing techniques are incorporated in the proposed method: the Frequency-Domain Independent Component Analysis(FDICA) is used to extract independent sound sources from sound mixtures, and the Time Delay of Arrival(TDOA) method is used to perform the source localization. This paper will detail the implementation of signal-processing techniques using MATLAB and the design of the octahedron microphone array.

Attempts have been made by various scholars and organizations to solve this "Cocktail Party Problem", which consists of extracting one sound from a mixture of sounds. This extraction, often referred to as the "blind source separation", is usually solved by using the Independent Component Analysis(ICA). The theory of the ICA algorithm was detailed in the book named "Independent Component Analysis" by Hyv¨arinen, Karhunen, and Oja[3]. Other attempts have been made to localize acoustic source, for example, the LABORIUS, Research Laboratory on Mobile Robotics and Intelligent System at the University of Sherbrooke, Canada, has published a project on Robust Sound Source Localization Using a Microphone Array on a Mobile Robot. [4]

A previous project named SLIMA: Sound Localization and Isolation with a Microphone Array [1], done by a group of undergraduate students at the Cooper Union for the Advancement of Science and Art, established a basic framework for performing source localization and isolation. Their work shows the promise of this technology and their simulation demonstrated that a microphone array could accurately isolate and localize multiple, simultaneous, audio sources. However, their approach involves separate microphone arrays for the two tasks, thus potentially require more space for the physical construct, and lacks the coherence of combining the source separation and localization into one integrated system.

Sound source separation is used to distinguish independent sound from a sound mixture, and it can be done using Independent Component Analysis (ICA). ICA allows robots to extract and recover the different sound content of a particular source from the mixtures of signals captured from microphones. After separating each sound source, the location of those sources can be determined. Sound localization aims at giving robots spatial instructions, such as what direction to turn its head or what position to walk towards. This can be done using the Time Delay of Arrival (TDOA) method: since there are multiple microphones in an array, the sound intensities and phase information captured by an individual microphone is different from the others. The time difference of arrival on microphone pairs can be used to approximate corresponding angles of incidence, and the angles calculated can be used to approximate the possible source location.
This paper details the design of the apparatus, (including the microphone selection, the microphone array geometry, and the data acquisition device), the system flow of acquiring signals, processing signals to perform source separation and localization, and the discussion of the experimental result and future work.

2. Apparatus Design

The main consideration in the apparatus design is the capability of capturing and transferring multi-channel audio signals to a computer. Thus, the microphone selected should have corresponding operating range to the human audible range, and the data acquisition unit should be aMble to sample all channels simultaneously.

2.1 Microphone

The microphone used is the Polsen OLM-19 model, which consists of an omnidirectional microphone, an external 5V power supply, and a 3.5mm audio jack. The audio jack is converted to a balanced XLR connection before connected the data acquisition device for noise reduction. The diameter of the microphone is only 50 mm, allowing the array to be easily portable.

2.2 Microphone Array

The microphones are arranged in an octahedron shape as shown in figure 1. Mic 1 through 4 are arranged in the mid-horizontal plane, and Mic 5& 6 are located on the vertical axis through the center of the mid-plane.

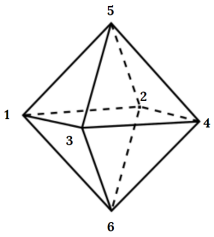


Figure 1. (left) Diagram of the microphone array's geometry, each number on the octahedron vertices represents a microphone.
Figure 2. (right) Constructed microphone array using 3D printed frame. The microphones are clipped onto the frame and can be easily rearranged.

2.3 Data Acquisition

The data acquisition device used for this project is the TASCAM US1608 audio interface. It can support a simultaneous sampling frequency of 44.1kHz across all 16 input channels, including 8 XLR channels.

Signal recording is performed using MATLAB Data Acquisition app, and individual microphone captured signal are stored unmixed and independent from the other signal for ten seconds session.

3. Digital Signal Processing

After recording with the said microphone array and audio interface, signals are processed using MATLAB.
The first step for signal processing is using the Frequency-Domain Independent Component Analysis(FDICA) to extract each of the individual sound from mixtures. And the second step is to find the incidence angle for each of the sound source through the TDOA method that uses the correlation between the extracted signal and the original signal. For this project, the implemented MATLAB code for Short-Time Fourier Transform and Inverse Short-Time Fourier transform is open source on MathWorks written by Hristo Zhivomirov. For complex signal source separation, one of the function called jade incorporated from JF Cardoso.

3.1 Source Separation Using FDICA

❖ 3.1.1 ICA Algorithm

Independent Component Analysis(ICA) is used to extract individual signal components in several mixtures of those signals and noise. For this project, the individual signal components are the different sound content from multiples speakers and background noise, denoted as $\boldsymbol{s}$. Signals captured by microphones are denoted as $\boldsymbol{x}$.

Suppose there are **N** different sound sources (speakers and noise), and **M** identical microphones in a microphone array, then individual sound signals are $\boldsymbol{s_1, s_2, \ldots s_N}$, and microphone signals are $\boldsymbol{x_1, x_2 \ldots x_M}$. Since each microphone-captured signal is a mixture of all individual sound signals, we have

$$x_i(t) = a_{i1}s_1(t) + a_{i2}s_2(t) + \ldots + a_{iN}s_N(t) \quad (1)$$

All individual sound signals $\boldsymbol{s_1, s_2, \ldots s_N}$ can be placed in a column vector of length N, collectively called Vectors $\boldsymbol{s}$. Similarly, the signals received by the microphones, $\boldsymbol{x_1, x_2 \ldots x_M}$, can be placed in a column vector of length **M** called Vector $\boldsymbol{x}$. All of the mixture coefficients a_{ij} can be placed into an **M*N** matrix, often referred to as the *mixing matrix* **A**. Decomposing this system of equations (1) into a matrix format, we have

$$\mathrm{x} = \mathbf{A}\,\mathrm{s} \quad (2)$$

ICA essentially estimates the mixing matrix A based on the statistical property of $\boldsymbol{x}$, and then computes the inverse matrix $\mathbf{A}^{-1}$. By doing so, entries of vector

s can be computed if signals s_i are statistically independent and non-gaussian by

$$s = \mathbf{A}^{-1}\mathbf{A}s = \mathbf{A}^{-1}x \qquad (3)$$

An assumption that in a realistic environment, each sound signal from the individual sound source is not correlated with the other sound source is made for this project [1]. For example, the sound content of speaker 1 does not correlate with the sound content of speaker 2. Thus, ICA could be employed to extract the sound content of various speakers.

- 3.1.2 Blind Signal Separation Using FDICA

There are several proposed ICA algorithms used for blind source separation(BSS), including FDICA and time-domain ICA(TDICA).
Conventional TDICA fails to separate source signals under heavily reverberant conditions due to the low convergence in the iterative learning of the inverse mixing matrix. Thus, we choose FDICA as the main algorithm for performing the BSS [2].
The FDICA is conducted in the following sequence: first, the signals are transformed from time-domain to frequency-domain using STFT, Short-Time Fourier Transform. Then the signals are divided into narrow sub-bands, and the inverse of the mixing matrix A is optimized in each sub band. Finally, the results are reconstructed back from the smaller sub bands. [2]
The method for FDICA for this project requires that all sources are mutually independent of each other. In this case, the mixing matrix **A** will be a complex-valued matrix, and it is obtained by the iterative process using the following equation:

$$A^{-1}{}_{i+1} = \eta(diag((< \Phi(s)s^H >) - < \Phi(s)s^H >)(A^{-1}{}_i)^{-1} + A^{-1}{}_i \qquad (4)$$

Where < > denotes the averaging operator, and *i* is the step of iteration and $\eta$is the step size parameter. Nonlinear vector function $\Phi(\cdot)$ is define as:

$$\Phi(s) = 1/\{1 + exp(-s^{(R)})\} + j \cdot 1/\{1 + exp(-s^{(I)})\} \quad (5)$$

Where $s^{(R)}$and $s^{(i)}$are the real and imaginary part of *s* respectively.

3 .2 Source Localization Using TDOA

Sound localization using TDOA is one of the most conventional ways to localize sound a source. Since the microphone array configures microphones such that the distance between any two is not zero, a specific sound signal will arrive at each microphone at a slightly different time. The incidence angle of a sound source could be estimated using this time difference, under the assumption that sound travels at a constant speed in air.
As shown in figure xx, x1 and x2 are two microphones, and d stands for the distance between them. Using cross-correlation of the two received signal from microphone x1 and x2, we could determine the time delay of these two signals. Then, from the diagram, the angle of incidence could be calculated by

$$\alpha = arccos(\frac{\Delta T * c}{d}) \qquad (6)$$

where c is the speed of sound in air.

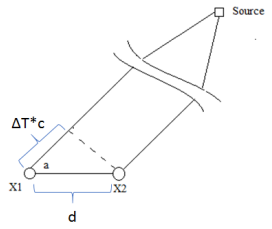


Figure 2. Diagram for TDOA calculation

3.3.1 Cross-Correlation of Signal

The accuracy of the time delay between signals from a pair of microphones is a key parameter to implement the above localization method. This time delay is found by performing a cross-correlation function in MATLAB.
If two signals have a time delay ΔT between them, the cross-correlation function will present a local maximum at the value of delayed time. Noted that in MATLAB, the parameter used for cross-correlation function is the number of samples rather than time. Hence, the time delay between two signals should be calculated as

$$argmax(f * g)/fs = \Delta T \qquad (7)$$

Where f and g are the two microphone signals, and fs is the sampling frequency.

3.3 Multi-Source Localization

Traditional TDOA method can only yield one time-delay when the cross-correlation is performed on two signals. Thus, it is necessary to find the time delay of an extracted sound signal relative to a microphone pair. Multi-source localization is achieved by implementing TDOA localization method to both extracted source signals and microphone signals.

After ICA, each extracted source signal is first compared with the four signals captured from the microphones on the mid-plane, i.e., mic 1 through 4. Whichever microphone signal that has the highest correlation with the extracted source means that this microphone is the closest to the source location. Then we use cross-correlation again to find the time delay between the extracted source and the adjacent microphones. Noted that for multi-source localization, the time delay of arrival is not simply the time delay between two microphone signals, but time delay between two microphone signals relative to the extracted source signal.

Using TDOA, the azimuth angle on the horizontal plane can be found. Similarly, for the vertical axis, an elevation angle for each sound source can be obtained.

The following diagram is an example of how to find the azimuth angle for one extracted source signal. In figure 3, S is the one of the extracted source signals. After performing cross-correlation with microphone signal x_1 through x_4, the result shows that the maximum correlation value occurs with, for example, microphone 1. Time delay is then calculated between S and microphone 3, say ΔT_{s3}, and between S and microphone 4, say ΔT_{s4}. Hence the actual time delay relative to source signal S is

$$\Delta T = \Delta T_{s3} - \Delta T_{s4} \qquad (8)$$

From previous discussion, we could determine the azimuth angle α for source S.

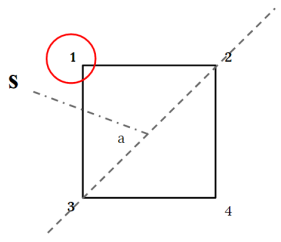


Figure3. Example diagram for multi-source localization

Then similarly, the time delay between the top and bottom microphone relative to the source signal can be found to acquire elevation angle for the sound source using TDOA.

4. Experiment

Experiments are set up to test the validity of our methods. First, we experimented with only one sound source. The microphone array was set on the table (about 0.7 m above the ground), along with the audio interface and the laptop doing the signal processing (figure 4). A speaker was standing at various locations around the microphone array with a distance ranging from 0.5 meters to 3 meters. The horizontal angle of the speaker location relative to the microphone array central axis was recorded to compare to the estimated results. Two handheld condenser microphones are also placed close to each participant's mouth to record the reference true signal. Multi-source experiments were performed with two speakers speaking stationary at various locations, and a background audio playing (music, speech, TV show, etc.) All experiments are conducted in a realistic environment, i.e., not inside of an anechoic chamber or a quiet room, with a background noise level of around 45 dBA.

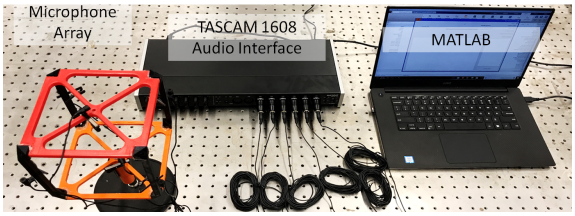


Figure 4. Experimental setup

5. Results and Discussion

5.1 Single-Source Localization Results

The accuracy of the single-source localization method was obtained through experiments in which a speaker at a known location was recorded and compared to the calculated results. The table below shows the comparison between the actual and calculated angle of incidence given a sound signal.

Trail	Actual (degree)	Calculated (degree)	Difference (degree)
1	270	273.7	-3.7
2	45	45	0
3	160	154.3	5.7
4	330	331	-1
5	90	95.8	-5.8

Table 1. Single-source localization results

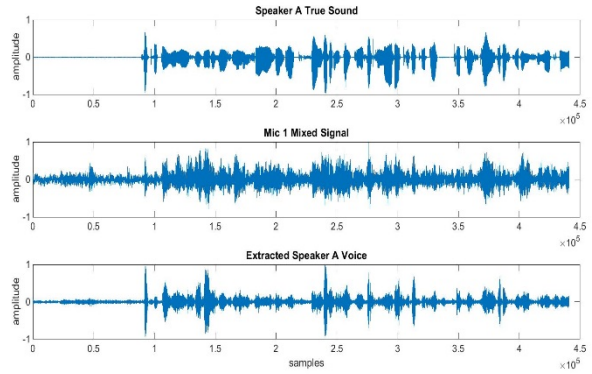
The calculated results are all within +/- 6 degrees from the actual values. Though the statistical model cannot be deducted from this few measurements, these results promised that the TDOA method for calculating the sound incidence angle is viable.

5.3 Multi-Source Localization Results

For source extractions, a total of 5 sources were estimated. For each extracted source signal, the experimenters listened and subjectively identified which speaker the sound best represents. For the trail shown below, Speaker A is a female participant, and speaker B is a male participant.

The following graphs show the comparison between the participants' voices and extracted source signals in time-domain for three similar sessions that each span for10 second.

(a)



(b)

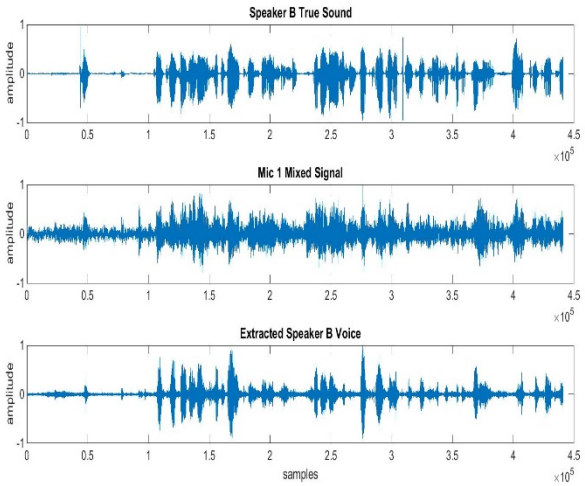


Figure 5. Waveform of the reference true signal, mixed signal captured by the microphone, and the extracted signal.

As presented in figure 5, the extracted source signals resemble the actual signal in time-domain. This resemblance indicates that the ICA algorithm can extract sound source successfully. However, since the number of sources was set to be 5, (one less than the microphone number), which is more than the number of speakers, some of the extracted source signals still have some mixtures of other sources. But for every trail, at least one extracted signal can be identified as a good representation of each sound source.

However, the sound quality of the extracted signals is unstable, especially for the male speaker. The reason for this instability is suspected to be due to the low-frequency content of the male voice, which might be mixed up with the low-frequency background noise from the rooms' HVAC system.

The table below shows the localization results of one of the trails, during which two speakers are talking at the same time while the background music is playing. Since six microphones are used, we can extract up to five sources. Out of the five extracted sources, three of them contain useful audio content from the two-original speech.

Extracted Signal	Planar Angle			Elevation Angle		
	Actual	Est.	Error	Actual	Est.	Error
Speaker A	225	230	-5	95	92	3
Speaker B	0	56	-56	110	99	11
Speaker A	225	228	-3	105	180	-75

Table 2. Multi-Source Localization Results (All data presented in unit of degree)

Two of the extracted sources sound like Speaker A, while one sounds like Speaker B. The accuracy of the multi-source localization is highly dependent on the quality of the ICA extracted signal. When the extracted signals can be identified subjectively as one of the known sound sources, the error between the actual and the estimated angle are relatively small (less than 10 degrees). However, for example, the above trail doesn't have a clear extracted signal for Speaker B. Thus, the error is relatively large. The elevation angle is calculated using the similar methodology. However, its error doesn't correspond to the quality of the ICA algorithm.

6. Conclusions

Our experiment shows promising results for both ICA source separation and single-source localization. For source separation, although some extracted signals are still mixed to various degrees, At least one signal can be extracted and located successfully during each trail. For single-source localization, it is

accurate and stable. Each trail only results in an error of less than 6 degree.

For most of the trials conducted for the multi-source localization, the results are still unstable, depending on the quality of the ICA extracted result. However, the single-source localization results show promising accuracy, promoting motivation to refine the ICA algorithm used to extract source signals.

The combination of FDICA and TDOA methods can be considered as successful, however the ICA source separation can be improved upon for better results.

7. Future Work

Some advancement to our methodology could help our project to achieve better results. For source separation, FDICA is implemented for this paper since it avoids the problem caused by reverberation in the time domain and it is much easier and faster to execute in MATLAB. However, other ICA method is worth investigating, including the combination of Time Domain ICA(TDICA) and FDICA and some ICA using machine learning algorithm instead of iterative process to find the mixing matrix. Better ICA result can eventually lead to more accurate localization result.

For now, method for determine sound event for each extracted signal is achieved by subjective decision. And for some extracted signals we encounter from our trail, they are still a mixture of several sources and should be discarded during localization. A machine learning audio event classifier could be implemented for faster and more accurate selection of extracted signals as this technique is aiming for robot or machine to perform better interaction with a human in a multi-source sound environment. Similar audio event classifier had already been proposed by Google's LabRosa team and could be implemented for our use. However, since our algorithm altered the frequency content of the signal, we might need additional data that's been processed using our algorithm to train the audio event classifier.

Finally, we are using MATLAB for digital signal processing. However, MATLAB has limitations on real-time analysis for large and continuous data. Scripting language with more powerful computation ability, such as Python, could be used in the future for a fast, real-time version of our processing method.

7 Acknowledgements

This work is supported by the Mechanical Engineering Department at the Cooper Union for the Advancement of Science and Art. We appreciate the department chair Professor Melody Baglione's advisory throughout the time we spent. We would also like to thank all those involved with this project, particularly Professor Tim Hoerning, Professor Stuart Kirtman, and Professor Carl Sable from Cooper Union Electrical Engineering Department on their help on the digital signal processing and electrical hardware. Laboratory technician Doug Thornhill for his help on some of the hardware and lab equipment, and Mike Giglia from the Cooper Union Makerspace for the help on 3D printed frame.

8 References

1.Bolbrock, Derderian, Gibbons, Maragos. "SLIMA: Source Localization and Isolation with a Microphone Array". IEEE, 2007.

2. Nishikawa, Saruwatari, Shikano. Comparison of Time-Domain ICA, Frequency-Domain ICA and Multistage ICA for Blind Source Seperation.

3.Aapo Hyv¨arinen, Juha Karhunen, and Erkki Oja. *Independent Component Analysis*. John Wiley & Sons, INC. 2001.

4.Jean-Marc Valin, Franc¸ois Michaud, Jean Rouat, Dominic L´etourneau. "Robust Sound Source Localization Using a Microphone Array on a Mobile Robot". IEEE Intelligent robots and Systems, 2003.