Microphone Array Speech Source Localization using SRP-PHAT and MUSIC

Bartłomiej Woś, Patryk Błoński, Bartosz Kawa

I. Introduction

SOUND source localization is a technology to determine the position of the objective sound source by analyzing sound signals and has been applied in many areas [1], [2]. The purpose of the work is to implement two algorithms (SRP-PHAT [3] and MUSIC [4]) that are capable of localizing multiple speech sources. Multiple Signal Classification (MUSIC) is a high-resolution direction-finding algorithm based on the eigenvalue decomposition of the sensor covariance matrix observed at an array. MUSIC belongs to the family of subspace-based direction-finding algorithms. Steered response power-phase transform (SRP-PHAT) combines the robustness and short-time analysis characteristics of the steered response power method with the advantage of the phase transformation method in time delay estimation. Hence, SRP-PHAT features robustness against noise and reverberation. The idea of speech source localization has many potential applications in our future world where we are constantly surrounded by all kinds of noises and distracted by them in our everyday life, moreover more and more people suffer from sorts of hearing issues which can be solved (in some degree) by devices equipped with source speech localization. In terms of hearing issues mostly older people have problems with understanding what other person speaks to them in crowded places, in this situation speech source localization could help them with hearing sound from only one direction. Furthermore, if a person can not localize the source of sound correctly it is dangerous for this person to even go outside, for example, this person won't know from which direction the car is coming and this can cause an accident. Localizing the source of speech can also be useful in enhancing the quality of teleconferences, which are now becoming increasingly popular. Another thing where sound source localization can be useful is in locating noise in engines or any kind of machinery, so engineers will know where to expect a problem and fix it faster. An institution where sound source localization can be very useful is the military, where sound source localization can be used in many ways, for example, a device equipped with accurate sound source localization can provide the location of a shooter. There are many other ways to use sound source localization, but these examples should confirm the value of the idea.

II. Problem formulation

The main problem is to estimate the DOA (Direction Of Arrival). Estimation starts with calculating delay between microphones with using the equation below :

$$\tau = \frac{S * \alpha^T}{v} \tag{1}$$

S - matrix containing coordinates of microphones
α - vector of angles
v - sound velocity

Delays between microphones is used to estimate the steering vector with presented equation :

$$sv = e^{-2j*\pi*\tau} \tag{2}$$

τ - delay between microphones

The steering vector is an essential element for estimating the direction of arrival. It is essential for the application of presented DOA estimation methods.
The calculations described here are for a microphone array consisting of eight microphones.

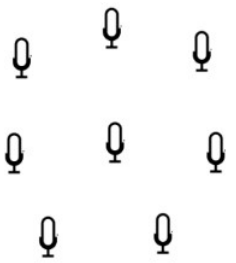

Fig. 1. simplified schematic of used microphone array

Next step is to use estimated steering vector in selected methods(MUSIC and SPR-PHAT).

bwos@student.agh.edu.pl, Bartlomiej_Wos - Bartłomiej Woś,
blonski@student.agh.edu.pl, pblonski - Patryk Błoński,
bartekkawa@student.agh.edu.pl, Bartosz_Kawa - Bartosz Kawa

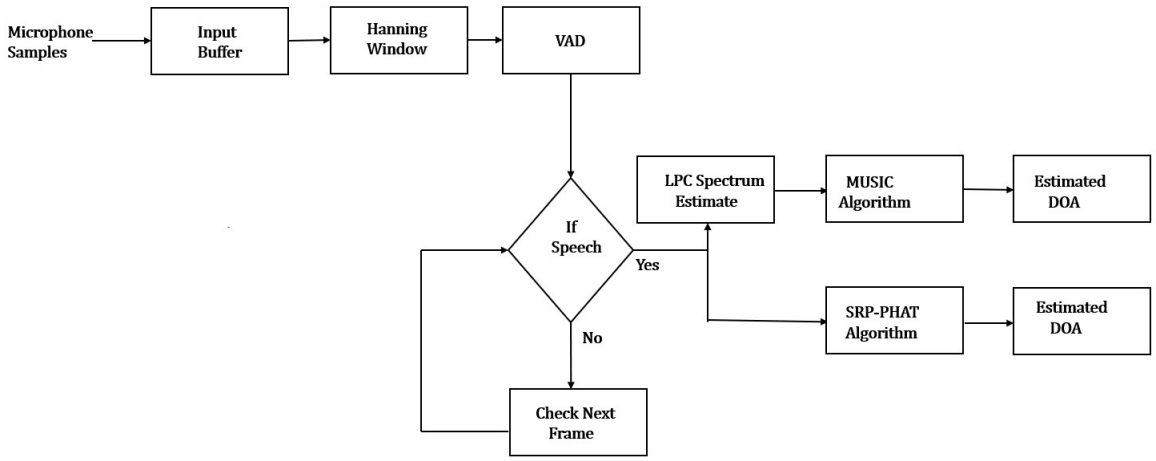

Fig. 2. Block diagram of the real-time processing of the proposed DOA estimation methods.

III. Methods, approaches solutions

A. RIR Generator

To check if implemented methods work correctly the RIR Generator was used. In python language it is possible to use module called rir-generator. This module allows to generate room impulse response and with that it is possible to carry out simulations of the implemented methods and check if these methods work correctly. With usage of RIR Generator it is easy to check how good selected methods work in multiple scenarios like different room dimensions, source location or microphone array geometry. Simulations performed with RIR Generator was essential to validate used methods before implementing these methods in real time.

B. VAD (voice activity detector)

Vad is used in our work to determine whether frames taken from microphones contain speech. This is important because by not taking into account the speech content check, we can get erroneous values, when locating the speaker. VAD is based on energy and zero-crossings measures of the speech signal.

Algorithm 1 VAD

Input: $x(0...L-1,t)$; L - window length

1. Calculate short-time energy for each incoming window.

$$\mathbf{E_n} = \sum_{k=0}^{L-1} \mathbf{x^2}[k]\mathbf{h}[n-k]$$

2. Calculate zero-crossing rate.

$$\mathbf{Z_n} = \sum_{k=0}^{L-1} |\mathbf{sgn}(\mathbf{x[k]}) - \mathbf{sgn}(\mathbf{x[k-1]})|w[n-k]$$

where:

$$\mathbf{h}[n] = w^2[n]$$

$$w[n] = \begin{cases} \frac{1}{2N}, & \text{for } 0 < n < N-1 \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

3. Set thresholds for parameters above according to your sound environment.

For this approach Hamming window was utilized.

C. LPC

LPC stands for Linear Prediction Coding.It uses the autocorrelation method of autoregressive (AR) modelling to find the filter coefficients. In terms of this project LPC improved selected methods by choosing only main frequency samples, with LPC usage only samples corresponding to main frequency of every frame are taken.For calculating LPC coefficients the Yule-Walker equation was implemented.

$$\begin{bmatrix} R_{xx}(0) & R_{xx}(1) & ... & R_{xx}(p-1) \\ R_{xx}(1) & R_{xx}(0) & ... & R_{xx}(p-2) \\ . & & & \\ . & & & \\ . & & & \\ R_{xx}(p-1) & R_{xx}(p-2) & ... & R_{xx}(0) \end{bmatrix} \cdot \begin{bmatrix} a_1 \\ a_2 \\ . \\ . \\ . \\ a_p \end{bmatrix} \equiv - \begin{bmatrix} R_{xx}(1) \\ R_{xx}(2) \\ . \\ . \\ . \\ R_{xx}(p) \end{bmatrix}$$

where:
$R_{xx}(0) \cdots R_{xx}(p)$.- autocorrelation estimate of signal
p - number that represents order of AR model

In this case an 80-th order of model was used.

D. SRP-PHAT

Beamforming techniques are applied to both source-signal capture and source localization. If the location of the source is not known, then a beamformer can be used to scan, or steer, over a predefined spatial region by adjusting its steering delays. The output of a beamformer, when used in this way, is known as the steered response. The steered response power (SRP) may peak under a variety of circumstances, but with favorable conditions, it is maximized when the steering delays match the propagation delays. By predicting the properties of the propagating waves, these steering delays can be mapped to a location, which should coincide with the location of the source.

Algorithm 2 SRP-PHAT

Input: $x(0...M-1,t)$; M - number of microphones
1. Use the VAD alogorithm to determine if the block contains a speech frame.
2. Convert signal to the frequency domain

$$\mathbf{X}(0...M-1,\omega) = \mathbf{FFT}(x(0...M-1,t))$$

3. $$\mathbf{Y}(\omega, \Delta_0...\Delta_{M-1}) \equiv \sum_{m=0}^{M-1} \mathbf{G_m}(\omega)\mathbf{X_m}(\omega)e^{-j\omega\Delta_m}$$

where:

$$\mathbf{G_m}(\omega) \equiv \frac{1}{|\mathbf{X_m}(\omega)|}$$

$$\Delta_m = \frac{-\vec{\zeta_0}(\theta)\cdot\vec{d_m}}{c}$$

$$-\vec{\zeta_0} \equiv \begin{bmatrix}\sin\theta\\ \cos\theta\\ 0\end{bmatrix}$$

$$\vec{d_m} \equiv \begin{bmatrix}d_x\\ d_y\\ d_z\end{bmatrix}$$

4. $$\mathbf{P}(\theta) \equiv \int_{-\infty}^{\infty} \mathbf{Y}(\omega, \Delta_0...\Delta_{M-1})\mathbf{Y}'(\omega, \Delta_0...\Delta_{M-1})d\omega$$

return $\mathbf{P}(\theta)$ for $\theta \in [0, 360)$

The algorithm returns a vector $\mathbf{P}$ for angles from $0°$ to $360°$ in increments of $1°$. The largest value of the vector corresponds to the angle for which it was calculated. Vector $\vec{d_m}$ corresponds to the position of the given microphone in the coordinate system. The speed of sound c is taken as $343\frac{m}{s}$

E. MUSIC

MUSIC stands for MUlitple SIgnal Classificitation, is a well known direction-finding algorithm. It is based on eigenvalue decomposition of the speech signal covariance matrix observed at array.

Algorithm 3 MUSIC

1. Estimate the correlation matrix.

$$\mathbf{R} = \frac{1}{K}\sum_{k=1}^{K-1} x_k x_k^H$$

2.Find eigendecomposition of R.

$$\mathbf{R} = Q\Lambda Q^H$$

3.Partition Q to obtain Q_n - smallest eigenvalues corresponding to the noise subspace.
4.Plot MUSIC pseudospectrum as a function of ϕ.

$$\mathbf{P_{MUSIC}} = \frac{1}{s^H(\phi)Q_nQ_n^Hs(\phi)}$$

Estimated signal directions are the largest peaks of P_{MUSIC} function

IV. Evaluation

A. Simulated Data

A simulation to check the performance of the implemented methods was carried out using the RIR Generator. The data taken for the simulation were :

$$microphone\ array = \begin{bmatrix}0.00 & 0.00 & 0.00\\ -38.13 & 3.58 & 0.00\\ -20.98 & 32.04 & 0.00\\ 11.97 & 36.38 & 0.00\\ 35.91 & 13.32 & 0.00\\ 32.81 & -19.77 & 0.00\\ 5.00 & -37.97 & 0.00\\ -26.57 & -27.58 & 0.00\end{bmatrix}$$

$$velocity = 343.0$$

$$room\ dimensions = \begin{bmatrix}4.0 & 6.0 & 3.0\end{bmatrix}$$

where:
$microphone\ array$ contains in each row the coordinates of one of the microphones of the microphone array , the coordinates are given in mm
$velocity$ was set to 343.0 m/s and it is the speed of sound.
$room\ dimensions$ contains example room dimensions for simulation purposes.

For simulation with usage of RIR Generator it was important

to provide reverberation time. In our work we decided to use the following value:

$$reverberation\,time = 0.4$$

The microphone array was moved to the center of the room. The sampling frequency was assumed to be 16kHz, and the frame length was 20ms. The sound source was set at 225 degrees for the first 1.25s, 270 degrees for the next 1.25s and then 90 degrees.
The following graphs present the results of the simulations obtained through the implemented algorithms.

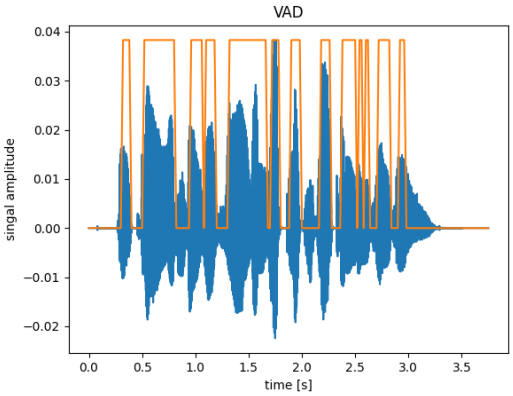

Fig. 3. A graph showing the performance of the VAD function (a value of 0 means that the signal fragment had noise alone)

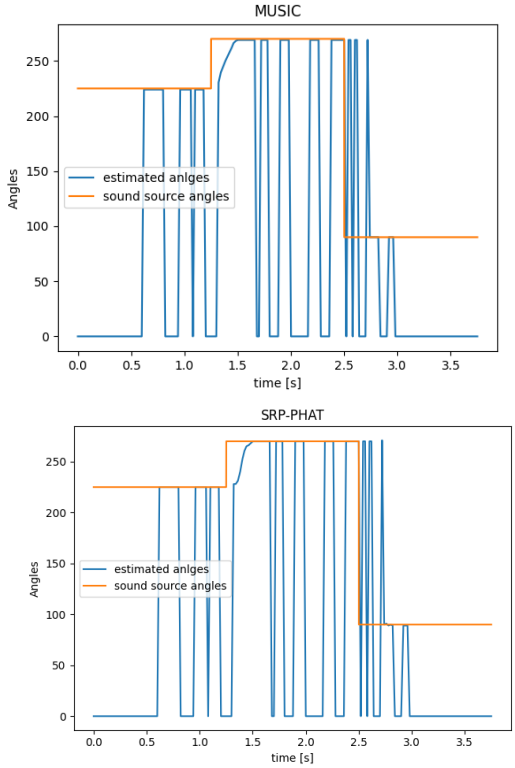

Fig. 4. Graphs showing the simulated angles over time and the performance of the implemented methods (MUSIC and SRP-PHAT). A value of 0 indicates that the frame contains noise itself, so the algorithms were not used.

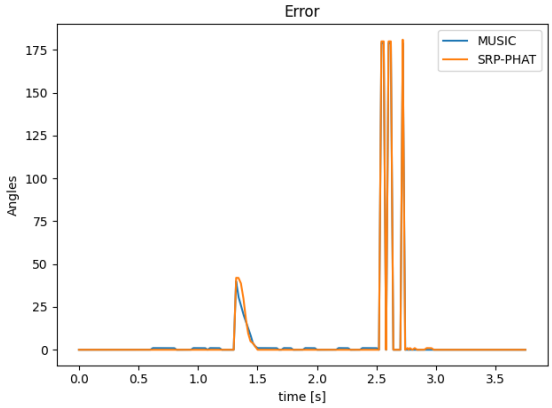

Fig. 5. Absolute error of measurement for both methods

	RMSE
SRP-PHAT	5.7°
MUSIC	4.6°

TABLE I
RMSE CALCULATED FOR MUSIC AND SRP-PHAT FOR ONE ANGLE CONSTANT IN TIME (FROM 1.25S - 2.5S)

B. Real-time performance

As a sensor device a Raspberry Pi with a microphone array was selected. Samples were transfered using ethernet to a laptop computer, where the program was running.

The algorithms were tested in a room with relatively high reverberation which was affecting the results. It was not possible to carry out accurate tests as in the case of simulations, because it was hard to tell where exactly the sound source was located. However, based on the experiments, we were able to confirm that the algorithms show reasonably consistent direction. Yet, there were slight delays when displaying the direction, which may be due to the use of matplotlib library, which is not optimized for real-time operation, or the length of the calculations performed by the proposed methods. Both SRP-PHAT and MUSIC are capable of detecting multiple speakers at once but their locations are not as accurate as with a single speaker.

It was also necessary to replace the vad function. The function used in the summations did not work properly during real-time testing, so a ready-made vad function from the webrtcvad [6] library was used.

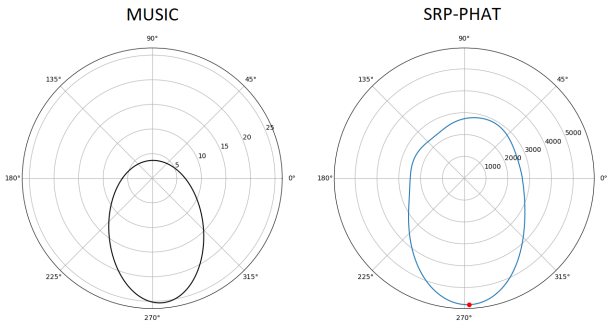

Fig. 6. Real-time display

V. Conclusions

This work shows a comparison of two methods, SRP-PHAT and MUSIC. The implemented methods work well in the simulated environment, but analysis with microphone-recorded data shows that real-world conditions are more challenging due to the mix of signal components in real environments. The algorithms presented in this paper are suitable for real-time operation, however it was computionally challenging and for this reason our implementation does not allow for effective estimation of direction due to the delays involved.

VI. Changes

Compared to proposed solution of our problem we did not use method presented in ”Robust Three-Microphone Speech Source Localization Using Randomized Singular Value Decomposition”. due to the low quality of the mentioned publication, we followed it to a small extent but eventually MUSIC and SPR-PHAT methods were used to localize the source of speech.

References

[1] *Hearing aid system with 3D sound localization* Wen-Chih Wu, Cheng-Hsun Hsieh, Hsin-Chieh Huang, Oscal T.-C. Chen, Yu-Jen Fang, TENCON 2007 - 2007 IEEE Region 10 Conference.

[2] Zhang, C.Q., Gao, Z.Y., Chen, Y.Y., Dai, Y.J., Wang, J.W., Zhang, L.R., Ma, J.L. *Locating and tracking sound sources on a horizontal axis wind turbine using a compact microphone array based on beamforming.* Appl. Acoust. 2019, 146, 295–309.

[3] *A High-Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments Using Microphone Arrays*, Joseph Hector DiBiase

[4] *Multiple Emitter Location and Signal Parameter Estimation*, RALPH 0. SCHMIDT

[5] *Robust Three-Microphone Speech Source Localization Using Randomized Singular Value Decomposition*, Serkan Tokgoz, Issa M. S. Panahi

[6] *webrtcvad libarary*, available: https://pypi.org/project/webrtcvad-wheels/