Journal Homepage: -www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)

Article DOI:10.21474/IJAR01/13377
DOI URL: http://dx.doi.org/10.21474/IJAR01/13377

RESEARCH ARTICLE

INVESTIGATION OF SIGNAL THRESHOLDING EFFECTS ON THE ACCURACY OF SOUND SOURCE LOCALIZATION

Saulius Sakavičius
Department of Electronic Systems, Faculty of Electronics, Vilnius Gediminas Technical University.

Manuscript Info

Manuscript History
Received: 05 July 2021
Final Accepted: 09 August 2021
Published: September 2021

Key words:-
Sound Source Localization, Acoustics, Microphone Array, Digital Signal Processing, Direction Of Arrival Estimation

Abstract

In this article theoretical analysis of the signal thresholding effects on the accuracy of cross-correlation based sound source direction of arrival estimation was presented. The aim of the investigation was to determine the theoretical limits and challenges of the accuracy of the localization of a speaker within an acoustic enclosure by cross-correlation of two microphone signals and to offer means to increase the accuracy of sound source direction of arrival estimation via selection of audio frames based on the time lag estimation realiability measure. For the investigation, audio material from an openly accessible database was used. Presented are the methods for obtaining various features of the microphone signal frames, signal amplitude to minimum error amplitude calculation and experimentation with threshold-based audio frame selection.

Introduction:-

Sound source localization is an important topic in communications, human-machine interfacing, robotics and security (Argentieri 2015, Kotus 2013). Localization of sound sources is often performed using compact devices such as ambient intelligence systems (home automation devices), smartphones and robots. By compact here we mean that the spatial dimensions of the entire system are comparable to the wavelengths of audio signals received by these systems. Even though the source localization is well understood and is comparatively reliable in a reflection-less environment, such as acoustically free field, the performance time difference of arrival (TDoA) based methods, such as generalized cross-correlation with phase transform (GCC-PHAT) or steered response power with phase transform (SRP-PHAT) deteriorates considerably when the multipath wave propagation due to the reflections of the sound waves within an acoustic enclosure becomes apparent (Brandstein, 1997; Datum, 1996, DiBiase, 2001). When the acoustic wave propagation is analyzed in a real-world environment, additional factors such as: acoustic properties of the enclosure, the size and geometry of the microphone array, the signal-to-noise ratio (SNR) of the microphone array and the associated acquisition system, sampling rate and quantization resolution, duration of the analysis window (in case of a digital signal processing system), must be considered (Xiao, 2016).

Experimental studies with real sound data (Lollmann, 2018) yielded dependences on the accuracy of sound source direction determination, which aimed to evaluate the influence of the following optional parameters: the sampling rate, the oversampling rate, the type of voice activity detector and its parameters. Our studies have shown that in order to increase the accuracy of audio source localization, it is necessary to separate the segments of signals received by microphones with and without a useful signal (speech), thus avoiding incorrect determination of the direction of the source.

Corresponding Author:- Saulius Sakavičius
Address:- Department of Electronic Systems, Faculty of Electronics, Vilnius Gediminas Technical University.

Audio source localization is performed by performing a cross-correlation of the signals received in the two synchronized microphones, obtaining the peak of the time lag function and theTDoA estimate, from which the angle of the source with the section connecting the pair of microphones is then calculated. To evaluate the accuracy of audio source localization, we have used an audio signals with labeled source coordinates provided in the publicly available database LOCATA (Lollmann, 2018); source coordinates were labeled with at a sampling frequency of 120 Hz. We have used only one pair of microphones from the 32 microphone array (Eigenmike). The distance between microphones was 8 cm.

In this article we investigate the factors that cause the has an impact of the accuracy of sound source direction of arrival (DoA) estimation via microphone signals cross-correlation, namely, the length and the SNR of the analysis frame. We propose a measure for cross-correlation time-lag estimate reliability, called signal-to-minimum-error-amplitude ratio, SMEAR.

Sound source direction of arrival estimation using cross-correlation of two microphone signals

The first phase of the study aimed to compare the location of the sound source obtained from the displacement of the cross-correlation peak (that is, the time lag) with the actual location of the sound source calculated from the change in the ground truth coordinates of the speaker as labeled in the dataset. The coordinates of each of the microphones and the coordinates of the sound source in three-dimensional space were used to determine the actual location of the sound source. Since the speaker (sound source) was moving within the enclosure on a horizontal plane and the microphone grid was not moving during the recording of the data set, it was rational to use only two spatial coordinates (*x*, *y*), ignoring the vertical axis (*z*). The sound wave time of arrival (ToA) between the sound source and the *i*-th microphone $\mathrm{ToA_i}$ is calculated for each microphone according to the Pythagorean theorem:

$$\mathrm{ToA_i} = \sqrt{(x_s - x_{mi})^2 - (y_s - y_{mi})^2}\,/v_s \quad , \tag{1}$$

here x_{mi}, y_{mi} are the coordinates of the *i*-th microphone, x_s, y_s are the coordinates of the sound source, v_s – speed of sound in air. After calculating the ToA time for both microphones, we can calculate the time difference of arrival TDoAij, which should coincide with the displacement of the correlation peak.

$$\mathrm{TDoA_{ij}} = \mathrm{ToA_i} - \mathrm{ToA_j} \quad . \tag{2}$$

The results of the study (Fig. 1) were obtained using signal analysis frames of different lengths. No frame selection algorithm was used in the study (eg, based on the signal Zero Crossing Rate (ZCR) or Short-Time Energy (STE)).

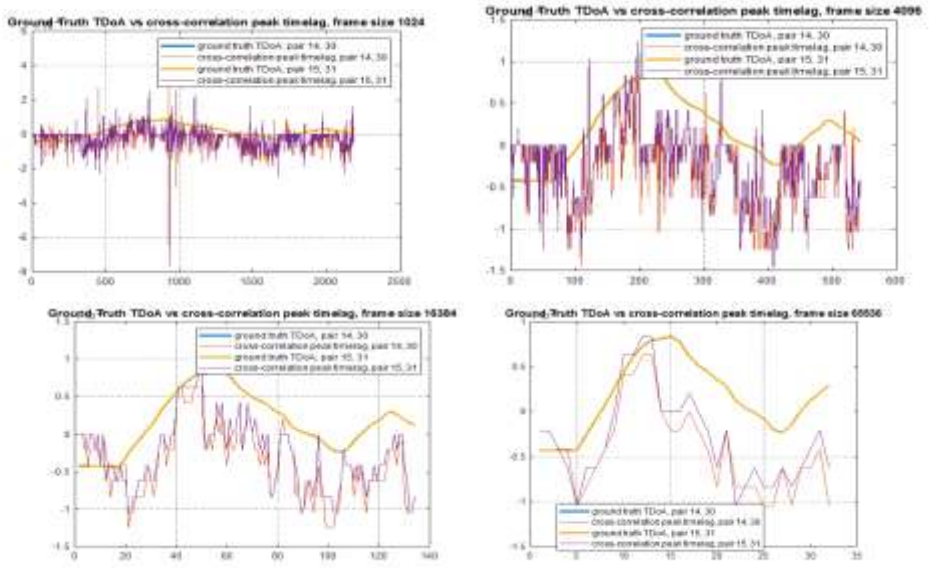

Fig. 1:- Audio source localization results using 1024, 4096, 16384, and 65536 sample frames; LOCATA dataset Eigenmike signals, speech source.

As can be seen from the presented trajectories of the sound source motion (Fig. 1), the noise of the source DoA estimation from the TdoA obtained from cross-correlation peak time lag decreases with increasing analysis frame length. Nevertheless, the error remains considerable.

Longer frames of analysis can give a more accurate estimate of the delay time difference compared to shorter frames. This can be attributed to the consideration that the speech signal contains both periodic portions (Fig. 2, left) and also has expressed transient envelopes (Fig. 3). For shorter analysis frames, a portion of a signal might not contain a transient and only contain the periodic signal. If the wavelenght of such signal within the analysis frame is shorter than the distance between the microphones, the TDoA estimation from the cross-correlation time lag becomes ambiguous, as one can not certainly determine whether the time lag was obtained for the same period of the wave or if the time lag contained more than one periods (Fig. 2, right). If the analysis window is longer, there is a higher probability that a non-periodic, transient envelope of the signal is contained within the frame, for which the cross-correlation time lag estimate is robust (Fig. 4).

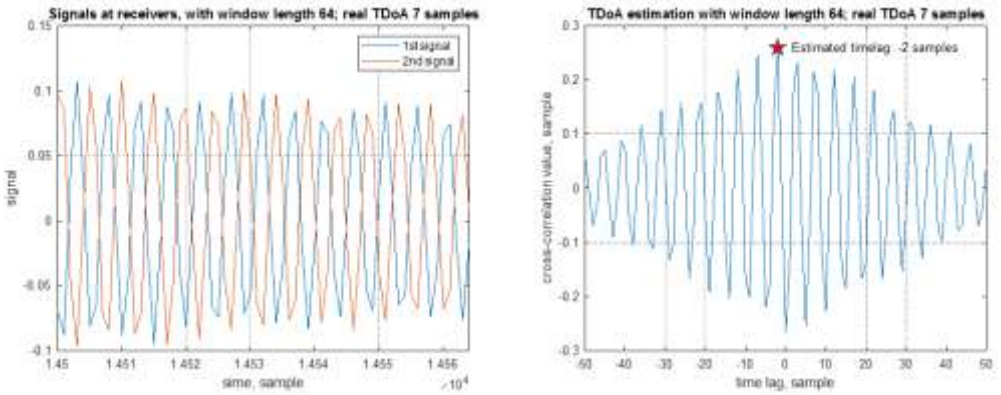

Fig. 2.:- Comparison of signals received in microphones and their correlation result using 64 sample analysis frame; time lag estimation is ambiguous (synthetic signal).

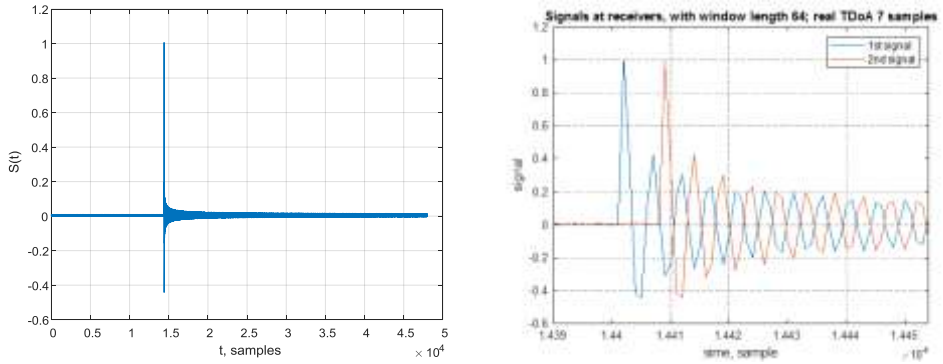

Fig. 3:- Comparison of single audio signals recorded on two microphones (entire signal, left; transient portion of the signal, right) ; synthetic signal.

By calculating the correlation for the 64 sample analysis frame, we obtain a clear correlation maximum corresponding to the signal delay (7 samples) (Fig. 4). However, such transient might not be included in a frame of the same length; in this case, an incorrect the correlation peak (incorrect time lag) (Fig. 2). The length of the analysis frame is considered too small to calculate the correct difference in delays from the envelope variation because the signal noise amplitude is larger than the envelope variation in the analysis window (Fig. 3). The signal graph shows that the time lag between the signals relative to each other is about 180 degrees, and indeed more (several periods); in this situation it is impossible to determine the correct time lag.

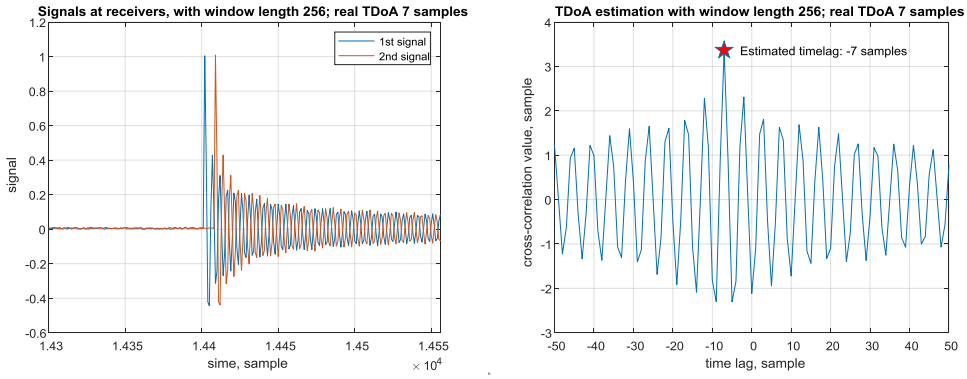

Fig. 4:- Comparison of signals received in microphones and their correlation result using 256 sample analysis frame; signal contains a transient envelope; synthetic signal

If the selected analysis frame is longer, there is an increased probability of a transient occurring in such window, and the change in the signal envelope will be greater than the amplitude of noise. The influence of the envelope of the signal on the correlation result is greater than the influence of noise.

Signal amplitude to minimum error amplitude ratio thresholding

The speech signal has exhibits time-varying properties. Some phonemes in some words are very similar to random noise while others exhibit signal periodicity. Moreover, the amplitude of speech signal is also inconsistent and contains transients. By analyzing and comparing the similarities and differences in the amplitudes of the audio signals recorded by the two microphones, it was observed that the influence of noise on the low amplitude signal can significantly affect the calculated correlation result and lead to incorrect estimation of the cross-correlation time lag. For this reason, it was decided to investigate how the accuracy of sound source localization changes for correlation by selecting only those audio signal frames in which the ratio between the signal amplitude range and the noise amplitude range exceeds a certain threshold.

We speculate that selection of frames based on such threshold would increase the accuracy of the source DoAestimation as the frames which produce unreliable TDoA estimates would be filtered out. Rationale for this would be that some audio frames would contain noisy audio signals. For such frames the cross-correlation time lag can not be reliably obtained and such frame is unusable for DoA calculation.

We speculate that the cross-correlation time lag can be considered reliable for a frame that exhibits a high coherence of signals at a single time lag. We select amplitude of the difference of the signals (the error amplitude) as the coherence measurement. Lower error amplitude of the signals within a frame indicates high coherence of the signals. Since the real TDoA of the signals is unknown, we measure the signal coherence at every time lag within limits set by the microphone array geometry: maximum time lag must be lower than the time a sound wave takes to propagate between the microphones. We select the lowest error amplitude and hereafter call it minimum error amplitude (MEA).

We also consider that a reliable cross-correlation time lag can be obtained for a frame which exhibits a large SNR, since as it was shown previously, noisy signals can lead to incorrect time lag estimates. We calculate the SNR of the frame as the ratio of the signal amplitude to the MEA. We call this ratio hereafter the SMEA ratio or SMEAR. We investigate the influence of SMEAR thresholding in the following section.

Influence of SMEAR thresholding on the source localization accuracy

We have evaluated the accuracy of the source DoA estimation using SMEAR thresholding with various threshold values: 2, 3 and 5.

The DoA estimates were compared with the ground truth DoA obtained from the dataset source position labels. The results of this investigation is presented in Fig. 5.

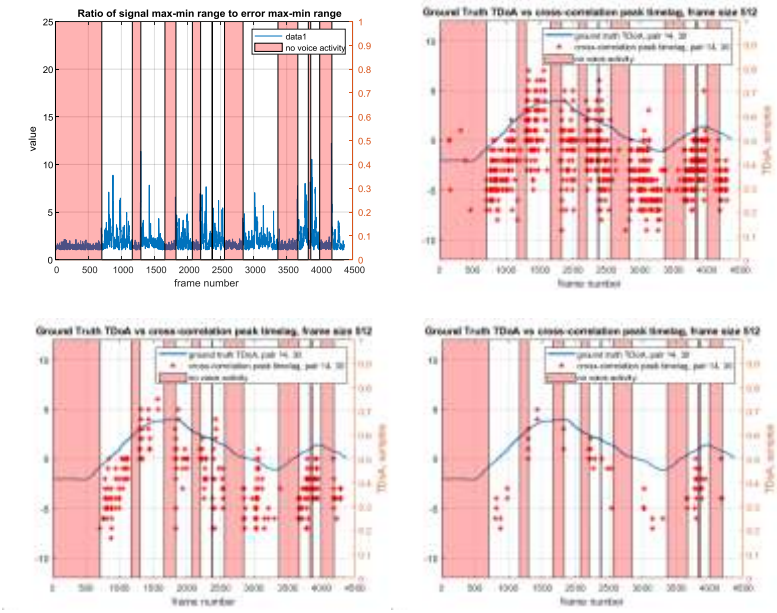

Fig 5:- Illustrations of SMEAR calculated for audio signal frames (top left) audio source positioning deviations by selecting different SMEAR threshold values (top right: 2, bottom left: 3 bottom right: 5); LOCATA dataset Eigenmike signals, speech source.

Despite the fact that the introduction of thresholding of audio signal frames has slightly improved the localization of the audio source, there is still too much uncertainty (from one to several tens of degrees). The reason for this situation is illustrated by the comparison of multiple signal analysis frames at different recording locations (Fig. 6). It can be seen from the figure that even high-amplitude signals, with low amplitude noise levels, can lead to incorrect setting of the time lag estimation, depending on the time of the signal it will be calculated.

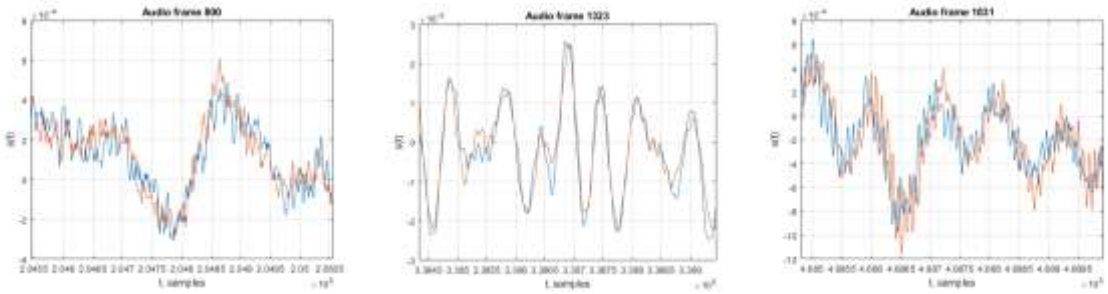

Fig 6:- Comparison of signals recorded by two microphones at different times; LOCATA dataset Eigenmike signals, speech source.

Conclusions:-

Analysis of speech signal recordings received by a pair of microphones shows that even at a small distance of 8 cm between two microphones, due to room acoustics, time-varying signal characteristics and other distortions, accurate localization of the sound source is not possible with real world signals using signal cross-correlation without additional processing. The study showed that by considering the signal amplitude to noise amplitude ratio, we can eliminate some of the erroneous results of the sound source localization, but other types of noise remain, making signal analysis ineffective on the time axis.

References:-

1. Argentieri, S.; Danès, P.; Souères, P. A survey on sound source localization in robotics: From binaural to array processing methods.Computer Speech & Language2015,34, 87–112. doi:10.1016/j.csl.2015.03.003.

2. Brandstein M. S., Silverman H. F. 1997. A robust method for speech signal time-delay estimation in reverberant rooms. IEEE International Conference on Acoustics, Speech, and Signal Processing, 1: 375–378.
3. Datum M. S., Palmieri F., Moiseff A. 1996. An Artificial Neural Network for Sound Localization Using BinauralCues. The Journal of the Acoustical Society of America, 100(1): 372–383.
4. DiBiase, J.H.; Silverman, H.F.; Brandstein, M.S. Robust localization in reverberant rooms. InMicrophone arrays; Springer, 2001; pp.157–180.
5. Kotus, J. Multiple sound sources localization in free field using acoustic vector sensor.Multimedia Tools and Applications2013,74, 4235–4251. 00009, doi:10.1007/s11042-013-1549-y.
6. Lollmann H. W., Evers C., Schmidt A., Mellmann H., Barfuss H., Naylor P. A., Kellermann W. 2018. The LOCATA Challenge Data Corpus for Acoustic Source Localization and Tracking. IEEE 10th Sensor Array and Multichannel Signal Processing Workshop (SAM), 1: 410–414.
7. Xiao X., Xu C., Zhang Z., Zhao S., Sun S., Watanabe S., Wang L., Xie L., Jones D. L., Chng E. S., Li H. 2016. A Study of Learning Based Beamforming Methods for Speech Recognition: 1–6.