Generalized Cross-Correlation based Noise Robust Abnormal Acoustic Event Localization utilizing Non-negative Matrix Factorization

Sungkyu Moon
Department of Visual Information Processing
Korea University,
Seoul, Korea,
skmoon@ispl.korea.ac.kr

Suwon Shon
School of Electrical Engineering
Korea University,
Seoul Korea
swshon@ispl.korea.ac.kr

Wooil Kim
School of Computer Science & Engineering
Incheon National University,
Incheon, Korea
wikim@incheon.ac.kr

David K. Han
Office of Naval Research
Arlington,
VA, Korea
ctmkhan@gmail.com

Abstract

In this paper, robust sound source localization for surveillance system is presented. In particular, we propose an algorithm for abnormal acoustic event localization using non-negative matrix factorization based frequency bin weighting. Based on the abnormal acoustic event localization experiments in real acoustic environment, the proposed algorithm's excellent strength is validated in terms of representative performance measures compared to the conventional method.

1. Introduction

Finding abnormal acoustic event localization has been an active research area with applications in surveillance systems [1-3] . In recent years, it became an important topic in surveillance systems related to acoustic processing such as sudden noise detection and localization in car, gunshot or scream detection [1-2]. Among the various algorithms, Generalized Cross Correlation (GCC) based algorithms are popularly used because of their simple implementation [4-6].

For a real environment, the abnormal acoustic event localization algorithm has to be robust to various types of noise source and low SNR. In GCC based algorithm, using only desired signal-dominant frequency bins while avoiding noise-dominant bins is required to achieve reliable estimation performance.

Previous investigations use speech characteristic and noise estimating methods for noise robust weighting. Denda et al. proposed weighted - Cross-power Spectrum Phase (CSP) coefficients based on an average speech spectrum [5]. Ichikawa et al. proposed a harmonic structure based weighting method for robust speech source localization [6]. Both of these methods have shown success in source localization for speech sources. Since they are designed for speech sources, however, they are not suitable for non-speech acoustic events such as breaking glass or alarm sounds.

Shon proposed Selective Frequency Bin (SFB) method based weighting function [3]. Pre-trained selected frequency bin represents well each abnormal acoustic event properties. However, due to the weights being fixed, there can be situations when high noise levels present in the selected frequency bins.

To address this issue, selecting (or weighting high) desired signal frequency bin in non-stationary noise, we propose a novel weighting function using Non-negative Matrix Factorization (NMF). The first step of the proposed process is to construct each abnormal event bases and their weights of NMF matrix pair by training with labeled set of clean abnormal event data base. Once converged, the basis matrix can then be used on a noisy input spectrum to extract desired abnormal event components. The proposed approach is based on an assumption that noise bases cannot sparse represent desired abnormal event signal spectrum.

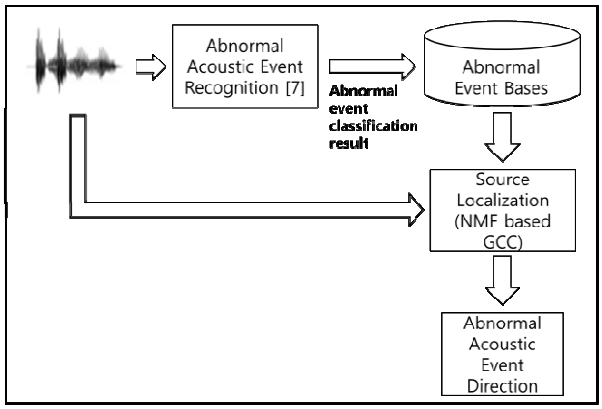

Figure 1: Block diagram of the proposed abnormal acoustic event localization system

2. Abnormal Acoustic Event Localization

An overview of the proposed approach is sketched in Figure 1. First, an abnormal acoustic event recognition step is needed. A hierarchical structure based abnormal acoustic event recognition approach, developed by Choi, is adopted

978-1-4799-4871-0/14/$31.00 ©2014 IEEE

in this study [7]. After recognizing the acoustic event into one of the pre-defined abnormal events, weighting function was determined based on the reconstructed abnormal event spectrogram. Finally, source localization is performed based on the weighting function, which eventually leads to the direction of an abnormal acoustic event.

The GCC of the l-th and q-th microphone signals is

$$R_{lq}(\theta) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \Psi_{lq}(t,\omega) X_l(t,\omega) X_q^*(t,\omega) \mathrm{e}^{j\omega\tau_{lq,\theta}}\, d\omega \qquad (1)$$

where $\tau_{lq,\theta}$ is the Time Delay Of Arrival (TDOA) between the *l*-th and *q*-th microphones when the source is at θ degree. The θ that has maximum amplitude of the GCC determines the direction of the sound source. $X_l(t,\omega)$ is the Short Time Fourier Transform (STFT) of the *l*-th microphone input signal in the *t*-th frame at frequency ω, and Ψ_{lq} denotes a weight function. The PHAT function is defined as $\Psi_{lq}(t,\omega) = 1/|X_l(t,\omega)X_q^*(t,\omega)|$. This approach, however, is insufficient since some of the frequency bins may not contain any desired abnormal event signal. For improving algorithm, equation (2) can be developed from PHAT function by putting a weight $W_{lq}(t,\omega)$.

$$\Psi_{lq}(t,\omega) = W_{lq}(t,\omega) \,/\, |X_l(t,\omega)X_q^*(t,\omega)| \qquad (2)$$

As mentioned above, in previous methods, the weight $W_{lq}(t,\omega)$ was obtained by the frequency bins with power in excess of the average power of corresponding frame in frequency domain. [3] However, due to the weights being fixed, there can be situations when high noise levels present in the selected frequency bins.

To address this issue, we propose an NMF based weighting function. The procedure of proposed system is shown by figure 2.

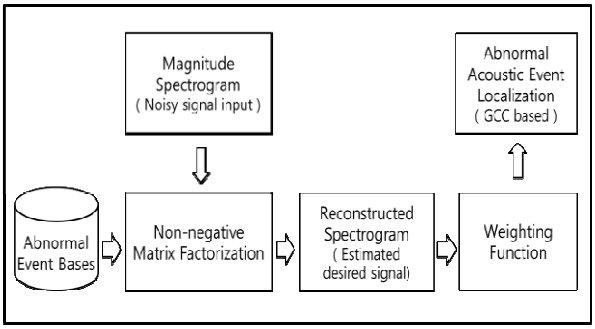

Figure 2: The procedure of proposed system

3. Proposed algorithm

The idea in NMF is to factorize the data matrix (in this work, magnitude spectrogram of microphone input $\mathbf{X}^{mag}$) as

$$\mathbf{X}^{mag} \approx \mathbf{DH}$$
$$X^{mag}(t,\omega) \approx (\mathbf{DH})_{\omega t} = \sum_{k=1}^{r} D_{\omega k} H_{kt}. \qquad (3)$$

A matrix **X** can be approximately factorized into two matrices **D** and **H** with dimensions $w \times r$, and $r \times t$, respectively. Matrices **D** and **H** are referred to as the basis vectors and weight of basis, respectively. r denotes number of basis vectors and $\mathbf{A}_{ij}$ denotes the i-j element of matrix **A**. Each column of **D** constitutes a source specific basis and **H** contains the corresponding weight of the basis determined by the amplitudes of each basis used in each time frame. It is well known that this method is quite useful for sparse representation of a database [8-9]. By assuming that each source is additive, we have equation (4) where subscripts, **S** and **N** indicate desired signal and noise.

$$\mathbf{X}^{mag} = \mathbf{X_S}^{mag} + \mathbf{X_N}^{mag} \approx [\mathbf{D_S}\ \mathbf{D_N}] \begin{bmatrix} \mathbf{H_S} \\ \mathbf{H_N} \end{bmatrix} = \mathbf{DH} \qquad (4)$$

There are various approaches to factorizing equation (3). In this work, we apply Euclidean distance measure which is popularly used. Updating rule of **D**, **H** is as follows.

$$H_{kt} \leftarrow H_{kt} \frac{(\mathbf{D}^{\mathrm{T}}\mathbf{X}^{mag})_{kt}}{(\mathbf{D}^{\mathrm{T}}\mathbf{DH})_{kt}},\ D_{\omega k} \leftarrow D_{\omega k} \frac{(\mathbf{X}^{mag}\mathbf{H}^{\mathrm{T}})_{\omega k}}{(\mathbf{DHH}^{\mathrm{T}})_{\omega k}} \qquad (5)$$

Following equation (5), we pre-trained only desired signal bases $\mathbf{D_S}$ using a clean abnormal event training database. When applying NMF to an input noisy signal after pre-training, we modify parts of equation (5) into (6) for updating noise bases only.

$$H_{kt} \leftarrow H_{kt} \frac{(\mathbf{D}^{\mathrm{T}}\mathbf{X}^{mag})_{kt}}{(\mathbf{D}^{\mathrm{T}}\mathbf{DH})_{kt}},\ (\mathbf{D_N})_{\omega k} \leftarrow (\mathbf{D_N})_{\omega k} \frac{(\mathbf{X}^{mag}\mathbf{H_N}^{\mathrm{T}})_{\omega k}}{(\mathbf{D_N}\mathbf{H_N}\mathbf{H_N}^{\mathrm{T}})_{\omega k}} \qquad (6)$$

Noise bases are initialized randomly (0~1 value) for data independence of the noise source. Through iterating equation (6), randomly initialized noise bases are updated to represent noise components of input noisy spectrogram. We assumed that noise components of input spectrogram cannot be sparse represented by desired signal bases, and confirmed experimentally that this assumption is appropriate. Consequently, we can reconstruct desired signal dominant spectrogram $\hat{\mathbf{X}}_S^{mag}$ as follows: $\hat{\mathbf{X}}_{\mathbf{S}}^{mag} = \mathbf{D_S}\mathbf{H_S}$

(matrix multiplication of desired signal bases $\mathbf{D}_S$ and $\mathbf{H}_S$ denoting amplitude of each desired signal basis used in each time frame).

In previous research [9], they modified equation (6) to updating $\mathbf{H}_S$, $\mathbf{H}_N$ separately, but in this work we observed experimentally that equation (6) performs better. There had been some previous work of factorization methods based on multi-microphone [8]. However in this work, we use average magnitude of spectrograms from microphones for simple NMF, since number of sources and transfer function between each microphone and source have little significance for finding desired signal dominant frequency bins. Finally, we derived GCC weight $W_{lq}(t,\omega)$ by normalizing elements of reconstructed spectrogram $\hat{X}_S^{mag}(t,\omega)$.

$$W_{lq}(t,\omega) = \frac{\hat{X}_S^{mag}(t,\omega)}{\max((\hat{\mathbf{X}}_S^{mag})_{t\omega})} \quad (7)$$

4. Experiments

For the purpose of evaluation, we used street environmental noise for surveillance camera system. We consider the direction of abnormal event as that corresponding to the highest GCC value. The room for the experiment is 10m x 10m x 10m and a Uniform Linear Array (ULA) consisting of 4 microphones with inter-spacing distance of 20cm to each other is located at the center of the room. The database was created with a desired abnormal event source at DOA of 30˚ at 1m from the ULA and stationary street noise sources [10] at 60˚ and non-stationary noise sources (car horn [11]) at 0˚. We set the same two abnormal events and the database from Choi [7] as shown in Table 1. The test database composed of noise level corresponding to SNR 0, 5, 10dB, respectively, in 16kHz sampling rate were used here. We used the abnormal event of the data base for the pre-training and division rate for train/test set was 3:1.

The performance was evaluated with two metrics, namely Root Mean Squared Error (RMSE) and Probability Of Success (POS). Each utterance was considered a success if the estimated DOA is within an angular tolerance of 10˚. Figure 7-10 show the direction estimation performance in terms of RMSE and POS under each experimental environment. It shows that the proposed method attained excellent performances under the non-stationary noise environment. In stationary noise condition, the performance is shown comparable. The conventional method shows good performance in high SNR, but the performance drops when SNR is lowered, especially in non-stationary noise environment. This is because the conventional methods used not only desired signal dominant frequency bands but also inadvertently captured noise concentrated frequency bands.

Method	Total Number	Total Duration [sec]
Scream (Male, Female)	414	888
Breakage of Glass	127	182

Table 1. Abnormal acoustic event DB information.

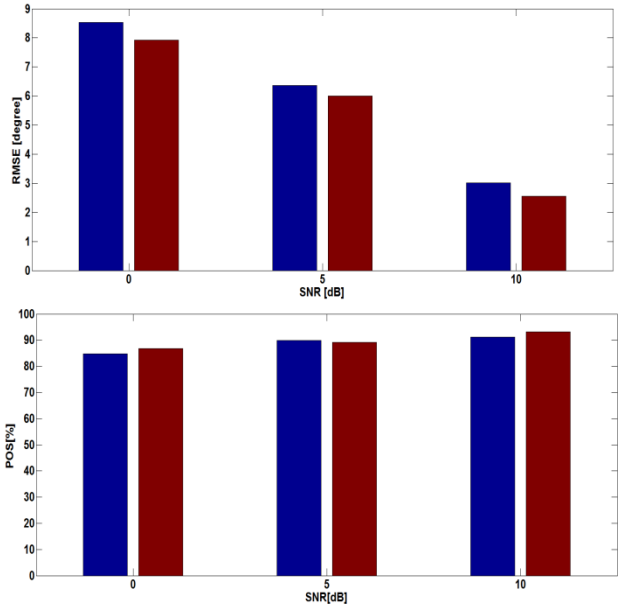

Figure 3: Direction estimation performance in terms of RMSE and POS (Scream in Street [10] environment)

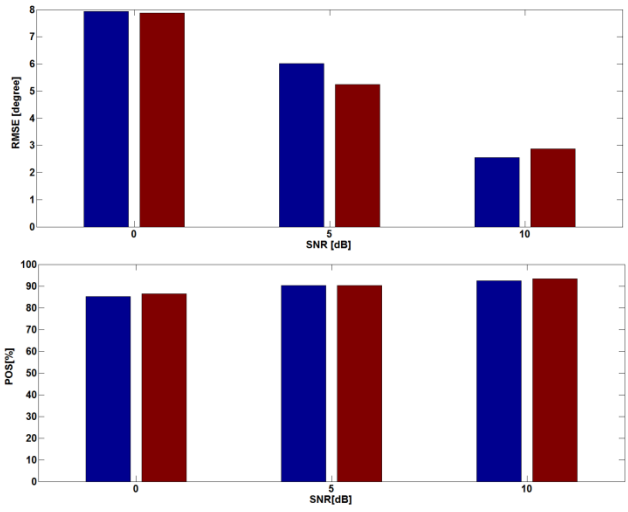

Figure 4: Direction estimation performance in terms of RMSE and POS (Glass breakage in Street [10] environment)

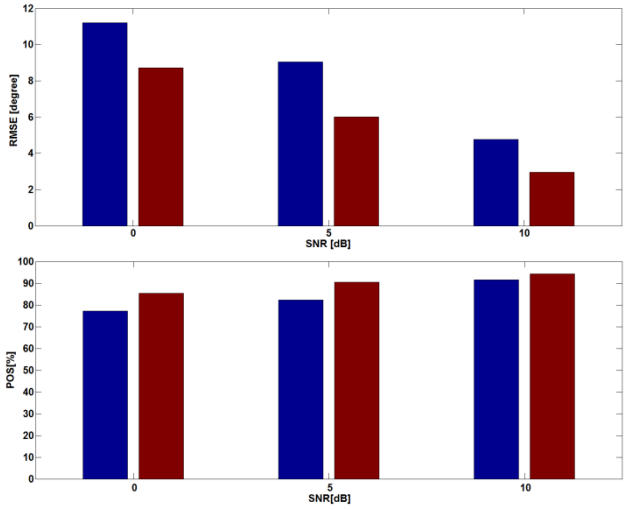

Figure 5: Direction estimation performance in terms of RMSE and POS (Scream in Street [10] + Car horn [11])

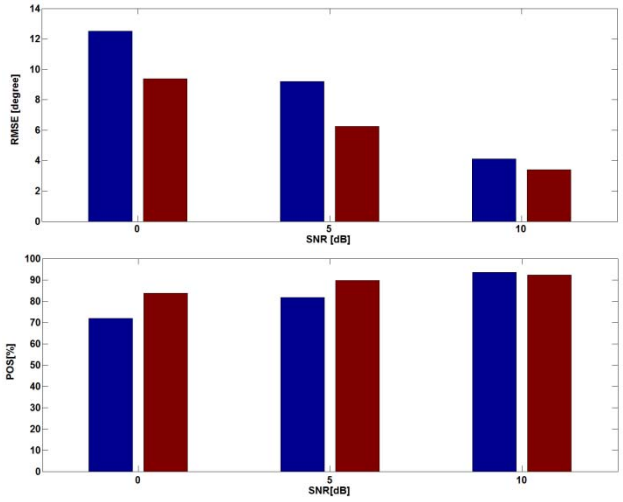

Figure 6: Direction estimation performance in terms of RMSE and POS (Glass breakage in Street [10] + Car horn [11])

5. Conclusions

This paper proposes a noise robust sound source localization using an NMF based weighting function. A high level of DOA accuracy was possible by establishing and incorporating weights of desired abnormal event frequency bins over those bins dominated by noise. The direction estimation experiments confirm that the proposed method is robust under surveillance camera environment, and expected to perform well in non-stationary noise environment.

6. Acknowledgements

This research was supported by Seoul R&BD Program (WR080951).

References

[1] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti, “Scream and gunshot detection and localization for audio-surveillance systems” in Advanced Video and Signal Based Surveillance (AVSS) 2007. IEEE Conference on, pp. 21–26, 2007

[2] S. Shon, E. Kim, J. Yoon, and H. Ko, “Sudden noise source localization system for intelligent automobile application with acoustic sensors” in Consumer Electronics (ICCE), 2012 IEEE International Conference on, pp. 233–234, 2012.

[3] S. Shon, D. K. Han, and H. Ko, “Abnormal acoustic event localization based on selective frequency bin in high noise environment for audio surveillance” in Advanced Video and Signal Based Surveillance (AVSS) 2013. IEEE Conference on, 2013, pp. 87–92.

[4] S. Shon, D. K. Han, J. Beh, and H. Ko, “Full Azimuth Multiple Sound Source Localization wigh 3-Channel Microphone Array,” *IECE Trans. on Fundamentals*, vol. E95-A, no. 4, pp. 745–750, 2012.

[5] Y. Denda, T. Nishiura, and Y. Yamashita, “Robust Talker Direction Estimation Based on Weighted CSP Analysis andMaximum Likelihood Estimation,” *IEICE Trans. on Information and Systems*, vol. E89-D, no. 3, pp. 1050–1057, 2006.

[6] O. Ichikawa, T. Fukuda, and M. Nishimura, “DOA Estimation with Local-Peak-Weighted CSP,” *EURASIP Journal on Advances in Signal Processing*, vol. 2010, pp. 1–10, 2010.

[7] W. Choi, J. Rho, D. K. Han, and H. Ko, “Selective Background Adaptation Based Abnormal Acoustic Event Recognition for Audio Surveillance,” *2012 IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance*, pp. 118–123, Sep. 2012.

[8] A. Ozerov, et al., “Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation”, *IEEE Trans. Acoust. Speech. Signal Process.*, 18(3): pp. 550-563, 2010.

[9] M Schmidt, et al., “Wind noise reduction using non-negative sparse coding”, Proc. Workshop on MLSP, pp.431-346, August 2007.

[10] ETSI : EG 202 396-1 V1.2.2 , September 2008.

[11] Sound Ideas, http://www.sound-ideas.com/