Suresh Gyan Vihar University Journal of Engineering & Technology
(An International Bi-Annual Journal)
Vol . 7, Issue 2, 2021, pp 08-11
ISSN: 2395-0196

Speech Sound Detection In Noisy Environment For Voice Controlled IoT Systems

Rajendra B. Mohite[1]

[1]Research Scholar, Department of Electronics and Communication Engineering, Suresh Gyan Vihar University, Jaipur, Rajasthan, India.

mohiterajendra52@gmail.com

Dr. Onkar S. Lamba[2]

[2]Professor, HOD,Departmrnt of Electronics and Communication Engineering, Suresh Gyan Vihar University Jaipur, Rajasthan, India

Onkar.lamba@mygyanvihar.com

Abstract: The blind source separation of audio signals using features and classifier based technique is the main contribution of this paper. The audio signals that get mixed into one another with respect room impulse response which depends on specific location in the room show specific characteristics. These characteristic features extraction and classification using neural network model is shown in this paper. The direction of arrival estimation and selecting particular signal as for required command in voice controlled IoT application is shown using room impulse response based mixing model and sound decible level estimation for direction of arrival. The features are extracted such as mel frequency cel cepstrum and are used to train the support vector machine (SVM) and decision tree for various commands and various surrounding sound samples from which SVM shows significantly good performance while executing the recognized command.

Keywords: Room Impulse Response (RIR), Direction of Arrival (DOA), Convolutive Blind Source Separation (CBSS), Mel Frequency Cepstral Coefficients (MFCC), Support Vector Machine (SVM) Classifier etc.

I. Introduction

The convolutive blind audio source separation problem arises when an array of sensor microphones is placed in a room, so that as well as recording a mixture of the source signals, multipath copies of the sources are also present. Many methods have been proposed for convolutive source separation, including time-domain deconvolution and frequency-domain ICA [8]. One approach that has been found to be successful in practical blind audio source separation applications is the degenerate unmixing estimation technique (DUET) [5]. DUET is a time-frequency (TF) masking method designed to address the underdetermined blind source separation (BSS) problem, where there are fewermixtures than sources. It separates an arbitrary number of source signals from two mixtures [5], under the assumption that in the time-frequency domain each time-frequency point of a mixture signal is due only to one of the sources, a property denoted as W-disjoint orthogonality [10]. To estimate the dominating source at each time-frequency point, DUET assumes anechoic mixing, i.e. that only delays and attenuations are present in the mixture, with no echoes. The method applies CBSS to the observed time-domain data to find a set of basis functions (dictionary elements), and then assigns eac basis function to one of the sources present in the sound field using a dependency analysis. In [4], this approach is modified, so that clustering is performed based on the estimated direction of arrival (DOA) of the sources.

II. Proposed Work

The proposed work consist of various steps as shown in figure 1. The input signals are first mixed into each other with consideration of room impulse response. Due to effect of accoustic properties of room, room impulse response is most important factor that affects the sound signals while capturing and estimating position of their respective sources. When there are multiple processing nodes involved in the processing, the selective processing is done based on Direction of Arrival (DOA). Room impulse response based mixing is responsible to generate real time accoustic effects when are compared with direct caprturing principle of microphone array. The instantaneous mixing model, noise effects and then separation of sound using CBSS method are detailed further in this section.

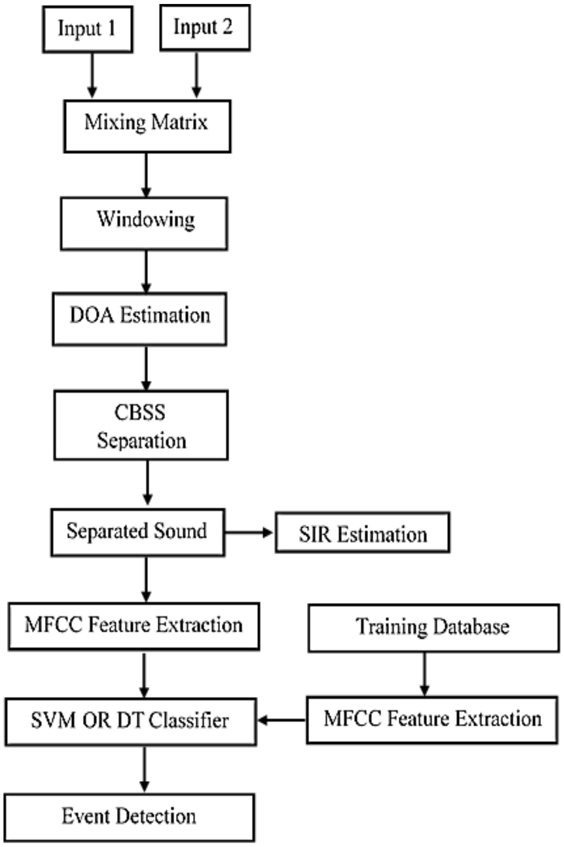

Figure 1: System work flow

The entire processing is based on steps enlisted as,

1. Input sounds
2. Mixing
3. DOA Estimation
4. CBSS based separation
5. SIR estimation of separated sound signals
6. MFCC feature extraction of separated sound signals
7. SVM or DT based classification using train database MFCC features
8. Event detection.

Mixing of sounds:

The mixing of sounds can be modeled by assuming that all the signals arrive at the sensors at the same time without being filtered, the convoluted mixture model (2) simplifies to

$$x(t) = As(t) + v(t). \quad (4)$$

This model is known as the instantaneous or delay less (linear) mixture model. Here, A = A0, is an M × N matrix containing the mixing coefficients. Many algorithms have been developed to solve the instantaneous mixture problem, see e.g. [17, 24]. Delayed Sources: Assuming a reverberation-free environment with propagation delays the mixing model can be simplified to

$$x_m(t) = \sum_{n=1}^{N} a_{mn} s_n(t - k_{mn}) + v_m(t)$$

Direction of arrival for Incoming sound:

To solve the permutation problem, an estimation of the direction of arrival of the source signal is utilized. By this estimation, each mixing vector into the correct column of the mixing matrix can be placed. Based on (4) and [12], we compute the DOA of the ith source signal as,

$$DOA(h_i) = \cos^{-1}\left(\frac{\arg\left[\frac{h_i(r)}{h_i(s)}\right]}{2\pi f c^{-1} d}\right)$$

Where and denote a pair of microphones which are nearest in the sensor array, and denotes the distance.

Convolution Blind Source Separation (CBSS)

Blind source separation algorithms are based on different assumptions on the sources and the mixing system. In general, the sources are assumed to be independent or at least decorrelated. The separation criteria can be divided into methods based on higher order statistics (HOS), and methods based on second order statistics (SOS). In convoluted separation it is also assumed that sensors receive N linearly independent versions of the sources. This means that the sources should originate from different locations in space (or at least emit signals into different orientations) and that there are at least as many sources as sensors for separation, The objective of blind source separation is to find an estimate, y(t), which is a model of the original source signals s(t). For this, it may not be necessary to identify the mixing filters Ak explicitly. Instead, it is often sufficient to estimate separation filters Wl that remove the cross-talk introduced by the mixing process. These separation filters may have a feed-back structure with an infinite impulse response (IIR), or may have a finite impulse response (FIR) expressed as feed-forward structure.

Feature Extraction:

MFCC Features

The first stage of speech recognition or event or command detection is to compress a speech signal into streams of acoustic feature vectors, referred to as speech feature vectors. The extracted vectors are assumed to have sufficient information and to be compact enough for efficient recognition[5].The concept of feature extraction is actually divided into two parts: first is transforming the speech signal into feature vectors; secondly is to choose the useful features which are insensitive to changes of environmental conditions and speech variation[6].However, changes of environmental conditions and speech variations are crucial in speech recognition systems where accuracy has degraded massively in the case of their existence. As examples of changes of environmental condition: changes in the transmission channel, changes in properties of the microphone, cocktail effects, and the background noise, etc. Some examples of speech variations include accent differences, and male-female vocal tract difference. For developing robust speech recognition, speech features are required to be insensitive to those changes and variations. The most commonly used speech feature is definitely the Mel Frequency Cepstral Coefficients (MFCC)features, which is the most popular, and robust due to its accurate estimate of the speech parameters and efficient computational model of speech[7].Moreover, MFCC feature vectors are usually a 39dimensional vector, composing of 13 standard features, and their first and second derivatives.

Performance Evaluation:

The proposed system experimentation is done using MATLAB based implementation. The room impulse response is responsible for actual direction of arrival estimation. The room impulse response is used with respect to two microphones of node 1 and two microphones of node2. These two nodes located randomly in a room will possess differing room impulse responses and hence for experimentation we have used combinations of the required impulse response data. Figure 4 and 5 shows the room impulse response plot.

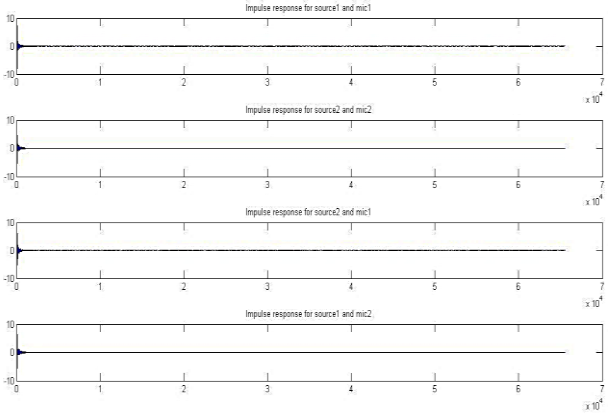

Figure 4: Room impulse response for node 1

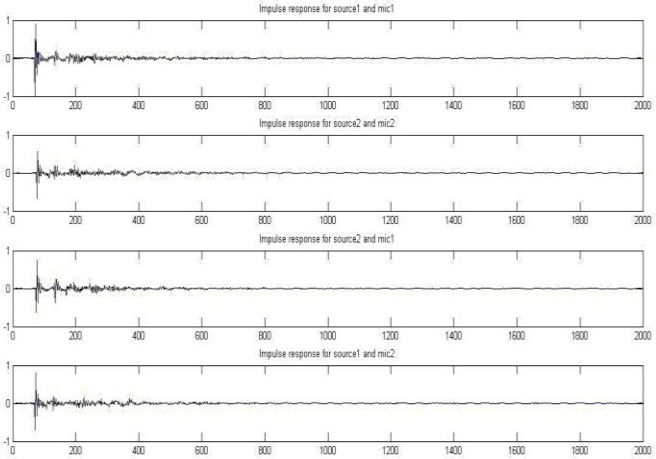

Figure 5: Room impulse response for node 2

Figure 6 shows the input speech signals used for experimentation. The MFCC coefficient are presented in spectrum graph as shown in figure 7.The mixed signal is obtained based on room impulse response as shown in figure 8 and 9.

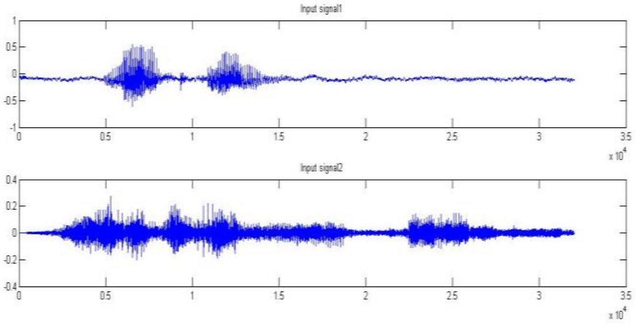

Figure 6: Input speech signals

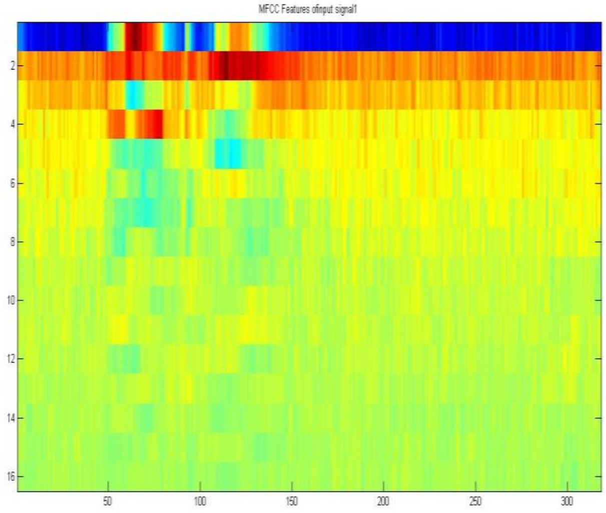

Figure 7: MFCC coefficient

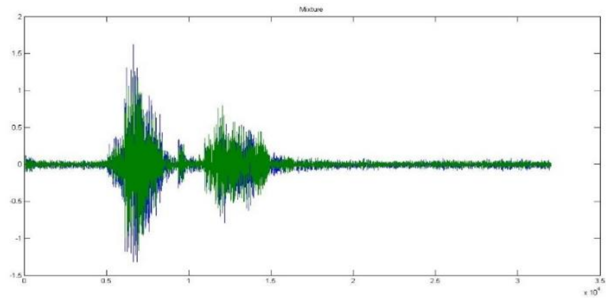

Figure 8: Mixed signal with respect to node 1 room impulse response

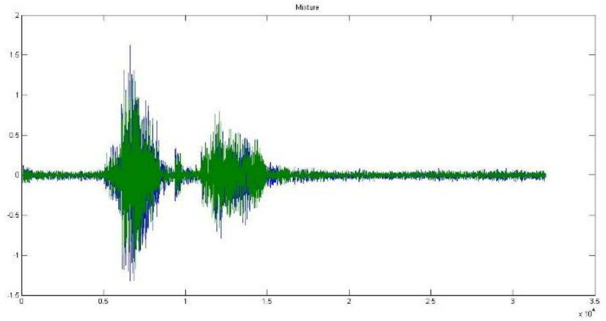

Figure 9: Mixed signal with respect to node 2 room impulse response

Event Detection Accuracy analysis:

The event detection accuracy is analyzed using SVM and Decision tree classifiers.

The comparative analysis is shown in figure 10.

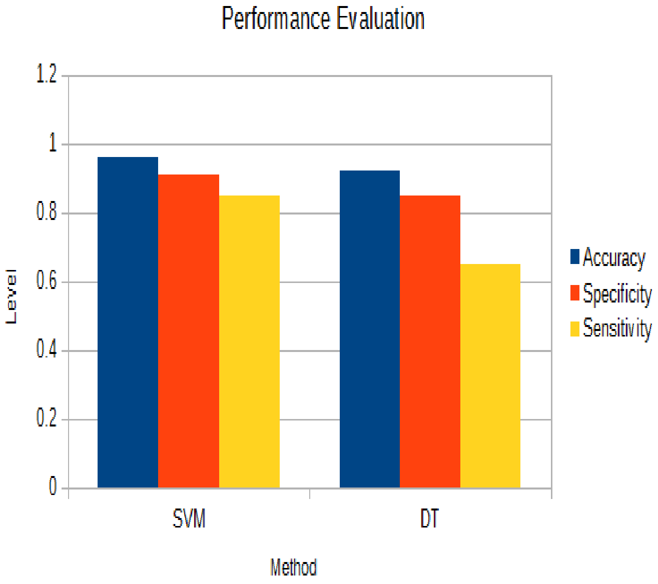

Figure 10: Performance evaluation

III. Conclusion

This paper focuses on sound separation and event detection in voice controlled home automation environment. The results obtained for event detection are satisfactory which ultimately depends on the separated sound signal quality. The result of event detection using MFCC coefficient and SVM classifier show the satisfactory outcome of the proposed system.

References:

[1] A. Mansour, N. Benchekroun, and C. Gervaise, “Blind separation of underwater acoustic signals,” in ICA’06, 2006, pp. 181–188.

[2] S. Cruces-Alvarez, A. Cichocki, and L. Castedo-Ribas, “An iterative inversion approach to blind source separation,” IEEE Trans. Neural Networks, vol. 11, no. 6, pp. 1423–1437, Nov 2000.

[3] K. I. Diamantaras and T. Papadimitriou, “MIMO blind deconvoluition using subspace-based filter deflation,” in ICASSP’04, vol. IV, 2004, pp. 433–436.

[4] D. Nuzillard and A. Bijaoui, “Blind source separation and analysis of multispectral astronomical images,” Astron. Astrophys. Suppl. Ser., vol. 147, pp. 129–138, Nov. 2000.

[5] J. Anem¨uller, T. J. Sejnowski, and S. Makeig, “Complex independent component analysis of frequency-domain electroencephalographic data,” Neural Networks, vol. 16, no. 9, pp. 1311–1323, Nov 2003.

[6] M. Dyrholm, S. Makeig, and L. K. Hansen, “Model structure selection in convolutive mixtures,” in ICA’06, 2006, pp. 74–81.

[7] C. Vay´a, J. J. Rieta, C. S´anchez, and D. Moratal, “Performance study of convolutive BSS algorithms applied to the electrocardiogram of atrial fibrillation,” in ICA’06, 2006, pp. 495–502.

[8] L. K. Hansen, “ICA of fMRI based on a convolutive mixture model,” in Ninth Annual Meeting of the Organization for Human Brain Mapping (HBM 2003), 2003.

[9] E. C. Cherry, “Some experiments on the recognition of speech, with one and two ears,” J. Acoust. Soc. Am., vol. 25, no. 5, pp. 975–979, Sep 1953.

[10] S. Haykin and Z. Chen, “The cocktail party problem,” Neural Computation, vol. 17, pp. 1875–1902, Sep 2005.

[11] M. Miyoshi and Y. Kaneda, “Inverse filtering of room acoustics,” IEEE Trans. Acout. Speech. Sig. Proc., vol. 36, no. 2, pp. 145–152, Feb 1988.

[12] Rajendra B. Mohite and Dr. Onkar S. Lamba, “Blind Source Separation Survey”, in IJSTR vol.8, no.11, pp.340-344, Nov.2019.