# An Eigenvalue-based Method for Estimating the Number of Simultaneous Speakers

Ehsan Fouladi
Electrical and Computer
Engineering Department
Yazd University
Yazd, Yazd, Iran 89195–741
Email: Ehsan.Fouladi@stu.yazd.ac.ir

Hamid Reza Abutalebi
Electrical and Computer
Engineering Department
Yazd University
Yazd, Yazd, Iran 89195–741
Email: habutalebi@yazd.ac.ir

***Abstract*—We address the problem of speaker number detection from speech signals of simultaneous speakers, collected by uniform linear microphone arrays (ULAs). Using the signal captured by microphone array a covariance matrix can be constructed. To take advantage of the information that covariance matrix provides, we employ eigenvalue decomposition on the covariance matrix. By finding the notable gap between decreasingly sorted eigenvalues of the covariance matrix, the number of the speakers can be determined. While many existent method for speaker number detection are faulty in the presence of the noise, our simulations on different numbers of simultaneous speakers demonstrates the robustness of the proposed method against both noise and reverberation.**

***Keywords-Microphone array speech processing; Speech overlap; Speaker number detection*.**

## I. Introduction

Determining the number of sources and estimating the direction of arrival (DOA) are two major topics in sensor array signal processing. Source number detection is essential for estimating the DOAs in multi-source scenarios [1]. In the case of speech signal, the problem is to detect the number of speakers in multi-speaker signals, where two or more speakers are speaking simultaneously [2].

During the last decades, several methods have been proposed to estimate the number of speakers. In 2003, Arai [3] suggested a procedure to find the so-called "equivalent number of speakers". A typical speech signal has a modulation characteristic with a peak around 4-5 KHz. It is claimed [3] that the peak changes according to the number of speakers. When the number of speakers increases, the value of peak will decrease consequently. In another work, Swamy et al. [2] proposed a method based on excitation source information, that uses cross-correlation of Hilbert envelop to achieve more accurate estimation of time delays. Based on the fact that the location of peaks in the histogram defers from speaker to speaker, counting the number of notable peaks gives an estimation of speaker number.

In 2011, Kumar and Balakrisha [4] proposed a similar method using Bessel coefficients of speech signal. A band-limited signal is demonstrated by Bessel coefficient series and time delays are calculated by applying cross-correlation function. The number of speakers is estimated by using the obtained coefficients of Bessel expansion and cross-correlation function between microphones.

In [5], Sayoud and Ouamour presented another method that determines the speaker number via an experimental investigation using the statistical properties of the $7^{th}$ Mel coefficient of the speech signal.

Firouzabadi and Abutalebi [6], proposed a method that estimates the number of speakers before localizing simultaneous speakers. The method employs the fact that speech signals are W-Disjoint Orthogonality (W-DO), so it is assumed that in any time-frequency bin, there is only one active speaker; the method then estimates the speaker number by applying K-means clustering and silhouette criterion.

In this research, unlike most of previous methods that only use one or two microphone, we propose a method that utilizes uniform linear microphone array to detect the number of speakers using an efficient model order selection method to separate the signal and the noise subspace. This method which is called SORTE (Second ORder sTatistic of Eigenvalues), has been primarily proposed to estimate the number of components in multivariate data analysis. This method was successfully employed to exploit the number of clusters for n-way probabilistic clustering [7]. The main idea of our work is to separate the signal and the noise subspace and then determine the number of speakers. This method has shown several advantages over previous methods. The most important one is its robustness against noise and reverberation. As it is shown in section IV, the algorithm works fine in noisy and reverberant environments.

The rest of this article is structured as follows: First the signal model for near field scenarios will be presented. Then we review the basic concepts of SORTE for counting source number in section II . In section III our work of using SORTE in speech signal for estimating the number of speakers is proposed . In section IV, our proposed method is evaluated not only in presence of noise in different SNRs, but also in reverberant and noisy-reverberant environments. Section V contains conclusions.

## II. Basic Concepts

As it is known, the covariance matrix of a sensor array could be divided into signal subspace and noise subspace. Suppose we have $M$ sensors and $K < M-2$ sources, then the covariance matrix would be $M$by$M$, and $K$ columns of that matrix would construct the signal subspace while the noise subspace would be formed by the rest $M-K$ columns. If the covariance matrix are divided correctly, then by counting the number of vectors of signal subspace sub-matrix, the number of sources could be easily detected. One way to solve this, is to find a gap between decreasingly sorted eigenvalues of covariance matrix. As in it is declared in [7], this could be done by using SORTE. In the following, we firstly present the signal model and then explain the SORTE algorithm.

### A. Signal Model

In some of practical applications of array signal processing, specially in speech signals processing, the criterion of *far-field* assumption is not satisfied. That is:

$$r > \frac{2L^2}{\lambda} \tag{1}$$

where $r$ is the radial distance from the source, the wavelength $\lambda$ is related to $c$, the wave velocity, by the simple relation $\lambda = c/f$ and $L$ is the array length. In the abovementioned situation, the source is said to be located in *near-field* of the array [8]. Suppose that the speaker is in the position $(r_0, \phi)$, as it is shown in Figure-1. Let $y(t)$ be the received signal of the reference microphone, then the received signal of the $m^{th}$ microphone is obtained as:

$$y_m(t) = \frac{r_0}{r_m} y(t - (\frac{r_m - r_0}{c})) \tag{2}$$

where $r_0$ and $r_m$ are the distance between the source and the reference and $m^{th}$microphone respectively.

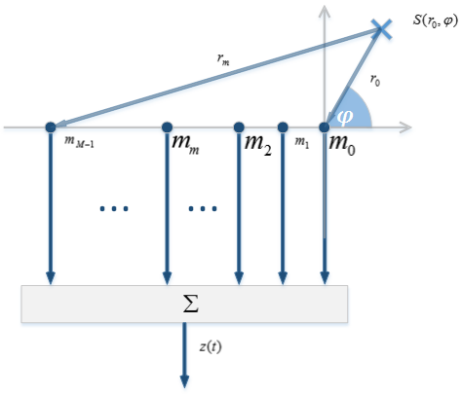


Fig. 1. Array Structure in Near-field Assumption.

$$r_m = \sqrt{(md)^2 + r_0^2 + 2r_0 md\cos(\phi)} \tag{3}$$

The output clean signal is defined as follows:

$$z_c(t) = \sum_{m=o}^{M-1} \frac{r_0}{r_m} y(t - (\frac{r_m - r_0}{c})). \tag{4}$$

In *near-field* situation, the amplitude of the microphone signals differs unlike the case in *far-field* assumption.
Let $n(t)$ be an additive white Gaussian noise that is uncorrelated with the source signal. It is also assumed that the noise of each microphone is uncorrelated with the noise at the other microphones; for the $m^{th}$ microphone signal we have:

$$z_m(t) = y_m(t) + n(t) \tag{5}$$

Suppose $T$ snapshots i.e.
$\mathbf{z}(t) = [z_0(t), z_1(t) \dots z_{M-1}(t)]^T \quad t = 1, 2, \dots, T$
are available. Stacking all measurements together, we have the matrix form of the received signals:

$$\mathbf{Z} = [\mathbf{z}(1), \mathbf{z}(2), \dots, \mathbf{z}(T)]_{M \times T} \tag{6}$$

then a good estimation of covariance matrix could be:

$$\hat{\mathbf{C}}_{zz} = \frac{1}{T}\mathbf{Z}\mathbf{Z}^H \tag{7}$$

### B. SORTE

The goal is to find the number of vectors constructing signal subspace submatrix of received signal covariance matrix. This can be aimed finding a gap between eigenvalues of covariance matrix [7]. Applying eigenvalue decomposition on the covariance matrix, we have:

$$EVD(\hat{\mathbf{C}}_{zz}) = \mathbf{U}\Lambda\mathbf{U}^H$$

where

$$\Lambda = diag(\lambda_1, \lambda_2, \dots, \lambda_M) \tag{8}$$

are the eigenvalues and

$$\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_M] \tag{9}$$

is the corresponding eigenvectors matrix.
Consider that eigenvalues are sorted in descending order as:

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_K > \lambda_{K+1} = \dots = \lambda_M$$

a gap measure is defined:

$$SORTE(K) = \begin{cases} \frac{var\left(\{\bigtriangledown\lambda\}_{i=K+1}^{M-1}\right)}{var\left(\{\bigtriangledown\lambda\}_{i=K}^{M-1}\right)} & var\left(\bigtriangledown\lambda\right)_{i=K}^{M-1} \neq 0 \\ +\infty & var\left(\bigtriangledown\lambda\right)_{i=K}^{M-1} = 0 \end{cases} \tag{10}$$

where $K = 1, \dots, M, \bigtriangledown\lambda_i = \lambda_i - \lambda_{i+1}$ and

$$var\left(\{\bigtriangledown\lambda\}_{i=K}^{M-1}\right) = \frac{1}{M-K}\sum_{i=K}^{M-1}\left(\bigtriangledown\lambda_i - \frac{1}{M-K}\sum_{j=K}^{M-1}\bigtriangledown\lambda_j\right)^2 \tag{11}$$

then the number of sources is the $K$ that minimizes $SORTE$:

$$\hat{K} = \underset{K}{argmin}(SORTE(K)) \tag{12}$$

## III. Speaker Number Detection

Unlike the previous approaches, we utilize microphone array for detecting the number of speakers. Using microphone array, we can take advantage of valuable information of covariance matrix. After eigenvalue decomposing the covariance matrix and sorting eigenvalues decreasingly, we consider that in no noise condition, then the $M-K$ smallest eigenvalues are zero; so we can detect speaker number by counting nonzero values of $\Lambda$; however as it is declared in [7], the true values of $\lambda_1, \ldots, \lambda_K$ are not available since the true $C_{zz}$ is not known in practice. In presence of noise and approximating $C_{zz}$ form (6), the eigenvalues are in the form of

$$\lambda_1 \geq \ldots \geq \lambda_K > \lambda_{K+1} \approx \ldots \approx \lambda_M \approx \sigma_\epsilon^2 \qquad (13)$$

where $\sigma_\epsilon^2$ is white Gaussian noise power.
The expression (13) states that there is a detectable gap, between $\lambda_K$ and $\lambda_{K+1}$, if $\lambda_K$ is notably larger than $\lambda_{K+1}$. Using equations (10) and (12) we can find that gap, this is used in this work, for detecting the number of sound sources. Finding the gap, is equivalent to detection of the speakers number. The process can be described as in Table-I.

TABLE I
Algorithm for Detecting the Number of Speakers.

1: *Construct the covariance matrix* $\hat{\mathbf{C}}_{zz}$ *from* (7).
2: *Apply eigenvalue decomposition on* $\hat{\mathbf{C}}_{zz}$ *to get eigenvalues matrix* $\Lambda$.
3: *Sort the eigenvalues decreasingly.*
4: *Find the* $K$ *by solving* (12).

## IV. Experimental Results

For the evaluation, we simulated a ULA of 8 microphones with inter-microphone distances of 4 cm and overlapping speech signal from 2, 3 and 4 speakers. The distance between the speakers to reference microphone is 1.5 m, and sources angles to the array axis are -10, 35, 85 and 170 degrees respectively. While most of the previous works report the evaluation results for only different noise levels, we consider both noise and reverberation in our evaluations. The results in different SNRs and RT60s have been measured with 10000 trials using $T = 1000$ Monte Carlo simulations. The detection accuracy is calculated as:

$$\%Accuracy = \frac{Number\ of\ Correct\ Detection}{Number\ of\ Total\ Trials} \times \%100.$$

Evaluation is done using completely overlapped multi-speaker signals with the sampling frequency of $16kHz$.

In the first experiment, detection accuracy is calculated for signals with 2, 3 and 4 simultaneous speakers, in the presence of white Gaussian noise. The results have been shown in Figure-2 (a-b-c) for the case of 2, 3 and 4 simultaneous speakers, respectively.

As it is shown in Figure-2, in very low SNRs, the detection accuracy is low. This can be justified by considering that the noise affects the eigenvalues and makes the gap between them undetectable; but in higher SNRs (less noise power), it is observed that the algorithm works precisely. When the speaker number increases the detection accuracy in low SNRs decreases which was expected due to spatial aliasing.

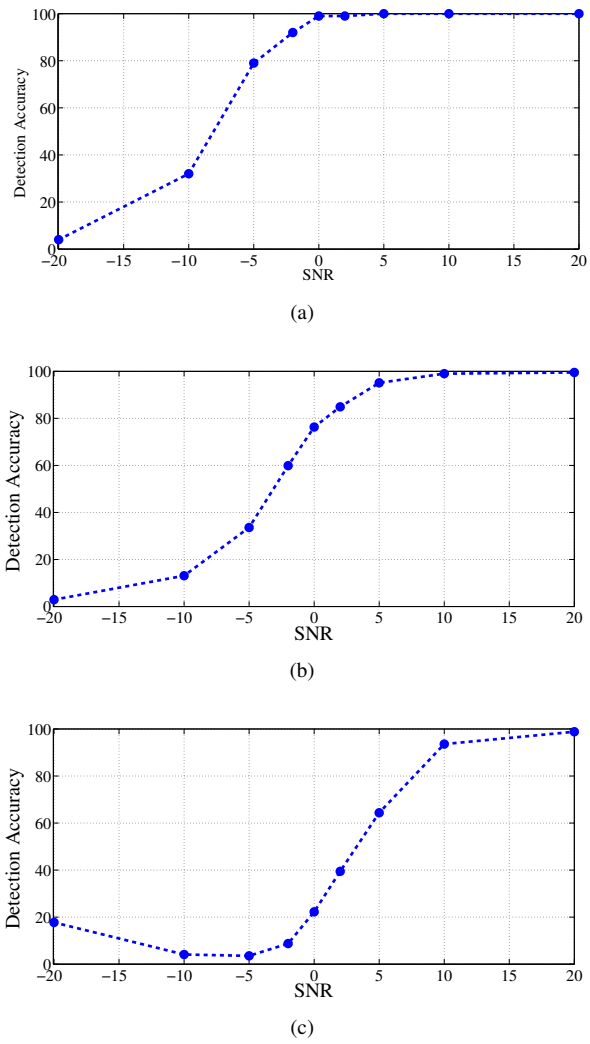


Fig. 2. Speaker number detection in different SNRs for (a) two, (b) three and (c) four speakers, speaking simultaneously.

Table II shows the values of $SORTE$ for 2, 3 and 4 simultaneous speakers in $5dB$ SNR. As we expected the minimum of values takes place in the column number equal to the number of speakers.

TABLE II
The $SORTE$ values for 2,3 and 4 simultaneous speakers in $5dB$ SNR.

| | $K=1$ | $K=2$ | $K=3$ | $K=4$ | $K=5$ | $K=6$ |
|---|---|---|---|---|---|---|
| 2 Speakers | 1.0348 | **0.0263** | 0.4171 | 0.8095 | 0.1063 | 0.6402 |
| 3 Speakers | 1.2041 | 0.4320 | **0.0319** | 0.0921 | 0.5603 | 0.7814 |
| 4 Speakers | 0.1452 | 0.1008 | 0.0810 | **0.0628** | 0.2184 | 0.6834 |

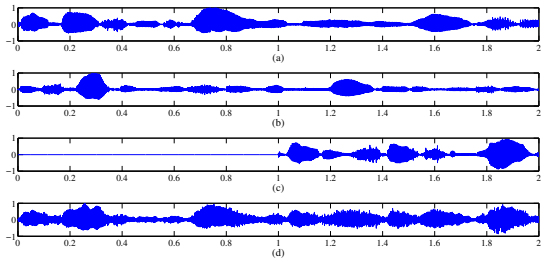

Fig. 3. Time-waveform of speech signal in the $2^{nd}$ experiment. (d) is the summation of (a), (b) and (c).

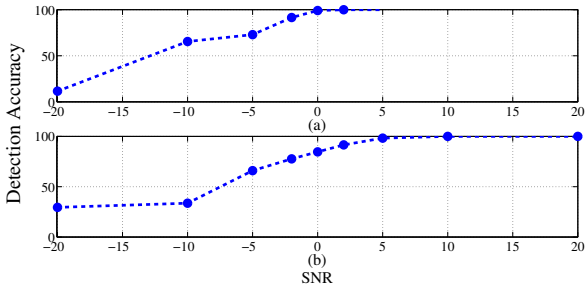


Fig. 4. Detection accuracy in SNRs, (a) first part of signal with 2 speakers, (b) second part containing 3 speakers.

In the second experiment the employed speech signals have two parts: in first part we have two simultaneous speakers, while in the second part there are three speakers. This is to check the algorithm ability to detect changing number of the speakers. Figure-3 demonstrates the time waveform of these speech signals and the mixture of them. The detection accuracy for this experiment is depicted in Figure-4. As seen, in both cases of 2 and 3 simultaneous speakers, the method has been able to detect the true number of speakers in moderate and high SNR values. In lower SNRs (less than 0dB), the method has encountered some errors in detecting the number of speakers.

In the third experiment we used Room Impulse Response Generator (Version 2.0.20100920) [9] to simulate a reverberant environment and evaluate the algorithm in the this situation. As Figure-5 shows, the proposed method is very robust against reverberation. The results are approximately the same in RT60s 200ms, 400ms and 600ms, so it could be judged that the reverberation has minor effect on the performance of the method.

## V. Conclusion

In this paper we employed SORTE method to take the advantage of eigenvalues of the covariance matrix for determining the number of speakers. We evaluated the algorithm in different numbers of simultaneous speakers with ULA microphone array. The results showed that in the open environments (with no reverberation), if there is no noise, the method detects

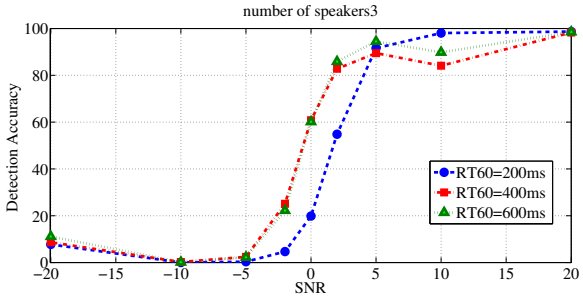


Fig. 5. Detection accuracy in SNRs, (a) first part of signal with 2 speakers, (b) second part containing 3 speakers.

the true number of the speakers; also, in the presence of the noise (signal to noise ratio above 0dB), the detection accuracy is very high. We also demonstrated the robustness of the proposed method in the reverberant environments. As shown, different values of reverberation time had negligible effect on the results.

## References


[1] K. Han and A. Nehorai, "Improved source number detection and direction estimation with nested arrays and ULAs using jackknifing," *IEEE Transactions on Signal Processing*, vol. 61, no. 23, pp. 6118-6128, 2013.

[2] R. Swamy, K. Murty and B. Yegnanarayana, "Determining number of speakers from multispeaker speech signals using excitation source information," *IEEE Signal Processing Letters*, vol. 14, no. 7, pp. 481-484, 2007.

[3] T. Arai, "Estimating number of speakers by the modulation characteristics of speech," *Acoustics, Speech, and Signal Processing, Proceedings. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), , IEEE International Conference on.* Vol. 2. IEEE, 2003.

[4] A. Kumar, P. V. Balakrishna, Ch. Prakesh, and S. V. Gangashetty, "Bessel features for estimating number of speakers from multispeaker speech signals," *Systems, Signals and Image Processing (IWSSIP), IEEE International Conference on.*, 2011.

[5] H. Sayoud, and S. Ouamour, "Proposal of a new confidence parameter estimating the number of speakers-an experimental investigation," *Journal of Information Hiding and Multimedia Signal Processing* vol. 1, no. 2, pp. 101-109, 2010.

[6] A. D. Firoozabadi, and H. R. Abutalebi, "A novel nested circular microphone array and subband processing-based system for counting and DOA estimation of multiple simultaneous speakers," *Circuits, Systems, and Signal Processing* vol. 34, pp. 1-29, 2015.

[7] H. Zhaoshui, C. Andrzej and C. Kyuwan, "Detecting the number of clusters in n-way probabilistic clustering," *Pattern Analysis and Machine Intelligence, IEEE Transactions On* vol. 32, no, 11, pp. 2006-2021, 2010.

[8] I. A. McCowan, "Robust speech recognition using microphone arrays," *PhD Thesis, Queensland University of Technology, Australia*, 2001.

[9] J.B. Allen and D.A. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal Acoustic Society of America*, vol. 65 no.4, pp. 943, 1979.