ESTIMATION OF THE NUMBER OF SOUND SOURCES USING SUPPORT VECTOR MACHINES AND ITS APPLICATION TO SOUND SOURCE SEPARATION

K. YAMAMOTO[†], *F. ASANO*[††], *W.F.G. van ROOIJEN*[†††], *E.Y.L. LING*[††], *T. YAMADA*[†], *and N. KITAWAKI*[†]

† University of Tsukuba, 1-1-1, Ten-noudai, Tsukuba, 305-8573, JAPAN
†† AIST, Central 2, 1-1-1, Umezono, Tsukuba, 305-8568, JAPAN
† † † Universiteit Leiden, WSD-bldg. 1173, Postbus 9515, 2300 RA Leiden, THE NETHERLANDS

ABSTRACT

A method of estimating the number of sound sources in a reverberant sound field is proposed in this paper. It is known that the eigenvalue distribution of the spatial correlation matrix calculated from a multiple microphone input reflects information on the number of sources. However, in a reverberant sound field, the feature of the number of sources in the eigenvalue distribution is degraded by the room reverberation. In this paper, Support Vector Machines is applied to classify the eigenvalue distributions which are not clearly separable. The proposed method is then applied to the source separation system and is evaluated via automatic speech recognition.

1. INTRODUCTION

Estimation of the number of sound sources is an important issue in sound localization and separation. In sound localization based on the subspace approach [1], the number of sources is required for determining the signal/noise subspace. In the sound separation, the number of sources is required for the determination of the dimension of the inverse filter (the number of output channels). Moreover, as employed in this paper, when separating an intermittent signal such as speech from continuous noise such as music, estimation of the number of sources can be used to detect the speech segment.

When input signals from multiple microphones (microphone array) are available, it is known that the eigenvalue distribution of the spatial correlation matrix calculated from the multi-channel input reflects information on the number of sources [2]. The number of *dominant* eigenvalues is equal to the number of (dominant) sound sources (1-rank model). Several methods for estimating the number of sources have been proposed based on this knowledge. However, it is sometimes difficult to employ these methods in reverberant sound fields such as those in ordinary rooms or offices as dealt with in this paper. In methods using information criterion such as AIC or MDL [3], the background noise is assumed to be spatially white. For applying this method to spatially-colored background noise, pre-whitening is required. For pre-whitening, the background noise must be independently observed. However, this is not possible in a reverberant sound field since such a field itself is a part of the background noise and cannot be separated from the direct sound. Another way is to estimate the eigenvalue corresponding to the background noise and to count the eigenvalues above the noise eigenvalues. This approach is also difficult since the noise level and, hence, the noise eigenvalues are time-variant.

In this paper, a method of estimating the number of sound sources in a reverberant acoustic field using Support Vector Machines (SVM) (e.g., [4]) is proposed. In this method, a rough shape of the eigenvalue distribution is utilized for estimating the number of sources. The advantage of this method over the conventional method is that it does not require pre-whitening or precise estimation of the background noise level, which may not be available in an actual reverberant sound field. To evaluate the performance of this method, it is combined with use of a maximum likelihood adaptive beamformer (e.g., [5]) in which the detection of speech/non-speech segment is required.

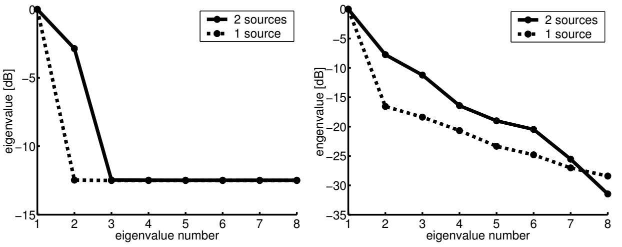
2. ESTIMATING THE NUMBER OF SOUND SOURCES

2.1. Eigenvalue Distribution

Let us consider the short-time Fourier transform of microphone array input $\mathbf{x}(\omega, T) = [x_1(\omega, T) \ \ldots \ x_M(\omega, T)]^T$, where ω is a frequency, T is a frame index and M is the number of microphones. This input signal is modeled as

$$\mathbf{x}(\omega, T) = \mathbf{A}(\omega, T)\mathbf{s}(\omega, T) + \mathbf{n}(\omega, T), \qquad (1)$$

where $\mathbf{A}(\omega, T)$ is a transfer function matrix, the (m, n)th element of which is a transfer function of the *direct* path from a nth source to the mth microphone. The symbol $\mathbf{s}(\omega, T)$ is a source spectrum, and $\mathbf{n}(\omega, T)$ is the background noise spectrum observed at the microphones.



(a) Spatially white. (b) Spatially colored.

Fig. 1. Eigenvalue distributions.

The spatial correlation matrix $\mathbf{R}(\omega)$ is defined as

$$\mathbf{R}(\omega) = E[\mathbf{x}(\omega, T)\mathbf{x}^H(\omega, T)], \quad (2)$$

where $\cdot^H$ denotes the complex conjugate transpose. When the noise $\mathbf{n}(\omega, T)$ is uncorrelated from the source $\mathbf{s}(\omega, T)$, the spatial correlation can be written as

$$\mathbf{R}(\omega) = \mathbf{A}(\omega)\mathbf{P}(\omega)\mathbf{A}^H(\omega) + \mathbf{K}(\omega), \quad (3)$$

where $\mathbf{K}(\omega) = E[\mathbf{n}(\omega, T)\mathbf{n}^H(\omega, T)]$ is the spatial correlation of the noise. The matrix $\mathbf{P}(\omega) = E[\mathbf{s}(\omega, T)\mathbf{s}^H(\omega, T)]$ is the cross-spectrum of the sources. When the noise is spatially white, (3) can be simplified as

$$\mathbf{R}(\omega) = \mathbf{A}(\omega)\mathbf{P}(\omega)\mathbf{A}^H(\omega) + \sigma\mathbf{I}, \quad (4)$$

where $\mathbf{I}$ is an identity matrix. The symbol σ is the variance (power) of the noise. In this case, the eigenvalues of $\mathbf{R}(\omega)$, $\lambda_1, \cdots, \lambda_M$ become

$$\lambda_1, \cdots, \lambda_M = \overbrace{\gamma_1 + \sigma, \cdots, \gamma_N + \sigma}^{N}, \overbrace{\sigma, \cdots, \sigma}^{M-N} \quad (5)$$

Assuming that the power of the source $\mathbf{s}(\omega, T)$ is greater than that of the background noise $\mathbf{n}(\omega, T)$, the typical eigenvalue distribution becomes that depicted in Fig. 1 (a). In this figure which reflects the characteristics shown in (5), N dominant eigenvalues corresponding the number of sound sources is observed. In a real acoustic problem in the reverberant sound field, the above assumptions, i.e., that $\mathbf{s}(\omega, T)$ and $\mathbf{n}(\omega, T)$ are uncorrelated in (3) and that $\mathbf{n}(\omega, T)$ is spatially white in (4), do not hold. However, the above characteristics shown in Fig. 1 (a) can be seen to some extent in the real eigenvalue distribution pattern depicted in Fig. 1 (b). In this paper, this difference in eigenvalue distribution pattern is utilized to estimate the number of sources.

2.2. Support Vector Machines

For classifying the eigenvalue distribution corresponding to the number of sound sources, SVM is introduced. Since SVM is basically a binary classifier, classification of 1-source and 2-source cases is considered in this paper for the sake of simplicity. Theoretically, this binary classification can be easily extended to a case with more sources.

Let us suppose that the training data set, i.e., the set of the eigenvalue distributions $\{\lambda_i\}$ and the corresponding class label $\{d_i\}$ (1-source event or 2-source event), is available. The class label can be either $d_i = +1$ (1-source event) or $d_i = -1$ (2-source event).

The decision function of SVM is written as

$$f(\mathbf{x}) = \sum_{i=1}^{l} \alpha_i^* d_i K(\mathbf{x}, \mathbf{x}_i) + b^*, \quad (6)$$

where $K(\cdot, \cdot)$ is a Mercer kernel, α_i^* is the optimal solution of the following quadratic problem [4] ($C > 0$ is a constant),

$$\text{maximize } -\frac{1}{2}\sum_{i,j=1}^{l} \alpha_i \alpha_j d_i d_j K(\lambda_i, \lambda_j) + \sum_{i=1}^{l} \alpha_i \quad (7)$$

$$\text{subject to } \sum_{i=1}^{l} \alpha_i d_i = 0,\ 0 \le \alpha_i \le C\ (i = 1, \ldots, l), \quad (8)$$

and b^* is calculated using the equation
$b^* = d_i - \sum_{j=1}^{l} \alpha_j^* d_j K(\lambda_i, \lambda_j)$ in which i is $0 < \alpha_i < C$.

Using function (6), an arbitrary eigenvalue distribution λ can be categorized into the following 4 categories. Category 1: $f(\lambda) \in (-\infty, -1]$, category 2: $f(\lambda) \in (-1, 0]$, category 3: $f(\lambda) \in (0, 1)$ and category 4: $f(\lambda) \in [1, +\infty)$. The number of sources is estimated based on the value of this decision function. When the value of the decision function falls into the category 1 or 2, the number of sources is estimated as 1, while if the value fall into category 3 or 4, the number of sources is estimated as 2. However, it should be noted that the region $(-1, 1)$ is termed the *margin of separation* and that when the value of the decision function falls into this region, i.e., category 2 or 3, the eigenvalue distribution is not clearly separable and the decision is somewhat ambiguous. These cases are further considered in the experiment presented in Section 4. Since the eigenvalue distribution is obtained in each frequency bin, the final decision as to whether the corresponding frame is a 1-source event or a 2-source event is made by taking the histogram of the above decision over the frequencies of interest.

3. SOUND SOURCE SEPARATION SYSTEM

In this section, assuming that the one sound source (target) is intermittent while the other (jammer) is continuous, the proposed method of estimating the number of sound sources is applied to the sound source separation system.

3.1. Separating Matrix

The maximum likelihood adaptive beamformer is given by the following equation:

$$y(\omega, T) = \mathbf{W}^H(\omega, T)\mathbf{x}(\omega, T), \quad (9)$$

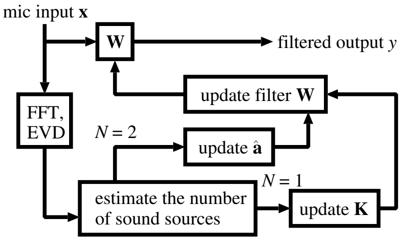


Fig. 2. Block diagram of the source separation system.

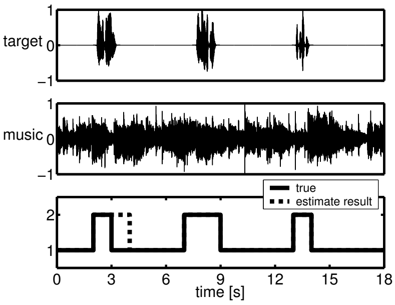


Fig. 3. Source signal and result of estimation.

where beamformer coefficient vector $\mathbf{W}(\omega, T)$ is given by

$$\mathbf{W}(\omega, T) = \frac{\mathbf{R}_J^{-1}(\omega)\hat{\mathbf{a}}(\omega)}{\hat{\mathbf{a}}^H(\omega)\mathbf{R}_J^{-1}(\omega)\hat{\mathbf{a}}(\omega)}. \quad (10)$$

Here, $\mathbf{R}_J(\omega)$ is a spatial correlation matrix when the target source is absent and is estimated in the 1-source event segment. On the other hand, $\hat{\mathbf{a}}(\omega)$ is a transfer function vector for the target source. This vector is estimated in the 2-source event segment using sound localization such as the MUSIC method [6].

Figure 2 shows a block diagram of the sound source separation system with the proposed number-of-sources estimation. The input time-domain signal $\mathbf{x}(t)$ is transformed to the frequency domain by the short-time Fourier transform (FFT), and the spatial correlation $\mathbf{R}(\omega)$ and the eigenvalue distribution λ is calculated (EVD). Then, the number of sources is estimated from the eigenvalue distribution by using the proposed method. Based on this estimation,

- If $N = 2$, $\hat{\mathbf{a}}(\omega)$ is updated.
- If $N = 1$, $\mathbf{R}_J(\omega)$ is updated.

Using the updated $\hat{\mathbf{a}}(\omega)$ and $\mathbf{R}_J(\omega)$, the coefficient vector $\mathbf{W}$ is calculated. This frequency-domain coefficient vector $\mathbf{W}$ is then transformed into the time-domain, and the time-domain input $\mathbf{x}(t)$ is filtered.

4. EXPERIMENT

4.1. Experimental Conditions

As input signals, two sound sources, Japanese words and a music signal were convolved with the measured impulse response of a meeting room with a reverberation time of 0.5 s. The microphone array is a circular array with 8 microphones. The diameter of the array is 0.5 m and the microphone is equally spaced. The position of the speech and the music source is selected from L1, L2 and L3 shown in Table 1. The music source is continuous while the speech source is intermittent as depicted in Fig. 3. The spatial correlation is calculated from the microphone input with a duration of 1 s every 1 s.

Table 1. Location of two sources.

	target source	music
L1	0 degree	280 degree
L2	100 degree	160 degree
L3	300 degree	120 degree

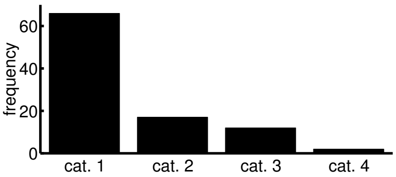


Fig. 4. Example of estimating result in one time frame.

To obtain the training data set, two segments of the input signal, one for the 1-source event and the other for the 2-source event are given. The segments each have a duration of 1 s. This training set is termed a rough training set hereafter. In a practical situation, this data set can be obtained by the human operator (user) by pressing the button once in the 1-source event and once in the 2-source event.

Even in the 2-source event, however, the number of sources may be *effectively* one due to the power difference of the spectrum of the two sources. Therefore, a more precise training set was also prepared in the following manner:

- $\lambda(\omega)$ is classified as a 1-source event when $|P_{s_1}(\omega) - P_{s_2}(\omega)| > P_{threshold}$.
- $\lambda(\omega)$ is classified as a 2-source event when $|P_{s_1}(\omega) - P_{s_2}(\omega)| \leq P_{threshold}$.

Here, $P_{s_1}(\omega)$ is the power of sound source 1, $P_{s_2}(\omega)$ is the power of sound source 2 at the frequency ω, and $P_{threshold}$ is the threshold of the power. This data set is termed a precise training set hereafter. The precise training set is not available in practical situations since the power difference cannot be known from the mixed input signal. This data set was used to investigate the best achievable performance of the SVM classifier.

In the estimation of the number of sources, the conventional threshold method was also employed for the sake of comparison. In this method, the number of sources is estimated by counting the number of eigenvalues above the arbitrary threshold.

Table 2. Rate of discrimination when the data in categories 2 and 3 are included/excluded.

	included	excluded
rough	79%	86%
precise	70%	77%

Table 3. Percentage of 4 categories

	cat. 1	cat. 2	cat. 3	cat. 4
rough	53%	21%	11%	15%
precise	40%	21%	16%	23%

4.2. Experimental Results

As described in Section 2.2, the eigenvalue distributions are classified into 4 categories. Figure 4 shows a representative histogram of these 4 categories over frequencies of interest. Based on this histogram, the final number of sound sources for a certain frame is decided. In this example, the number of sources is determined as one since the number of frequency bins classified as categories 1 and 2 is greater than that of categories 3 and 4. This estimation is conducted in each time frame (the frame length is 1 s). Figure 3 also shows results of estimation and the true classification for the input depicted in Fig. 3.

As described in Section 2.2, when the eigenvalue distributions are classified into categories 2 and 3 (margin of separation), these distributions are not clearly separable. Table 3 shows frequencies of the distribution categorized into the 4 categories for all examined data. Table 2 shows the discrimination rate when the distributions classified into categories 2 and 3 are included or excluded. From this table, it was shown that when these ambiguous data were excluded in the final decision, the discrimination rate was improved by 7%. Based on this result, the data classified into only categories 1 and 4 were utilized in the final decision in the experiment described below.

Figure 5 shows the rate of correct estimation. For the sake of comparison, the results for the threshold method are also shown. For the threshold method, the highest rate was achieved when the threshold was around -13 dB. For the proposed SVM method, performance similar to the best performance achieved by the threshold method was obtained for both rough and precise training.

Furthermore, for the word recognition rate, the SVM method achieved a rate comparable to the highest rate achieved by the threshold method in Fig. 6. The rate for the rough training, however, is lower than that for the precise training by 10%.

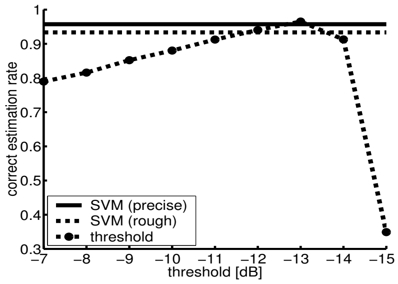


Fig. 5. Correct estimation rate.

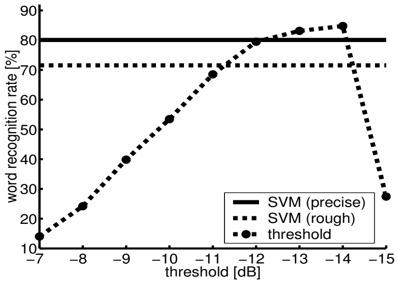


Fig. 6. Word recognition rate.

5. CONCLUSION

We have herein presented a method of estimating the number of sound sources by classifying the eigenvalue distribution using SVM. One of the advantages of SVM is that it can achieve a better performance by excluding ambiguous eigenvalue distributions for the classification, that was verified by the experiment. The proposed method was then applied to the source separation system and was evaluated via ASR. The performance of the system using the proposed method is equivalent to the best performance using the threshold method. Under practical conditions, the threshold method much trial and error may be required to obtain its best performance. Thus, the advantage of the proposed method is that it requires only a short period (2 s) of observed input for the training.

6. REFERENCES

[1] R. O. Schmidt, in *Proc. RADC Spectral Estimation Workshop*, pp. 243-258, 1979.

[2] A. Cantoni and P. Butler, *IEEE Trans. Comm*, vol. COMM-24, pp. 804-809, 1974.

[3] M. Wax and T. Kailath, *ASSP*, vol. 33, pp. 387-392, 1985.

[4] N. Christianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and other kernel-based methods*, Cambridge University Press, 2000.

[5] D. H. Johnson and D. E. Dudgeon, *Array Signal Processing*, Prentice Hall, Englewood Cliffs, NJ, 1993.

[6] F. Asano, H. Asoh, and T. Matsui, *IEICE Trans. Fundamentals*, vol. E83-A, no. 11, pp. 2286-2294, 2000.