DETECTION OF OVERLAPPING SPEECH IN MEETINGS USING SUPPORT VECTOR REGRESSION

[1]*Kiyoshi Yamamoto,* [2]*Futoshi Asano,* [3]*Takeshi Yamada and* [3]*Nobuhiko Kitawaki*

[1]kyama-conf@mmlab.cs.tsukuba.ac.jp
[1,3]University of Tsukuba, Graduate School of Systems and Information Engineering,
1-1-1 Tennoudai, 305-8573, JAPAN
[2]AIST, Information Technology Research Institute, Central 2, 1-1-1 Umezono, 305-8568, JAPAN

ABSTRACT

A method of detecting overlapping speech in meetings is proposed in this paper. The eigenvalue distribution of the spatial correlation matrix reflects information on the relative power of sound sources. By applying Support Vector Regression to a set of input eigenvalues, the relative power of sources is estimated. Based on this, overlapping speech is then detected. The proposed method was evaluated using recorded meeting data, which showed that the detection rate increased by around 9% compared with the conventional Support Vector Machines approach.

1. INTRODUCTION

Conversation in meetings often suffers from interference by overlapping speech (OS). When automatic speech recognition (ASR) is applied to recordings of meetings including OS segments, these segments result in severe deterioration of the transcription. Therefore, the detection of OS segments is an important issue in automatic transcription of meetings. Once OS segments are detected, these segments can be omitted for the transcription, or sound source separation can be applied to reduce overlap.

The eigenvalue distribution of the spatial correlation matrix calculated from a microphone array input reflects the information on the number and power of sound sources [1]. When the difference of the relative power of sources is small, the number of dominant eigenvalues roughly corresponds to the number of active sound sources. Using this property, the authors previously proposed a method of estimating the number of sources by clustering the eigenvalues using Support Vector Machines (SVM) [2]. Using this method, the OS segments can be detected. A problem of this method, however, is that when sources with weak power are included, the difference in relative power becomes large, resulting in failure in detecting the weak sources.

In this paper, a method of detecting OS segments using Support Vector Regression (SVR) (e.g., [3]) is proposed. In this method, the eigenvalue distribution is utilized for estimating the relative power of sound sources. The regression function in SVR is designed based on the training data set of eigenvalue distributions for estimating the relative power of sources. The OS segment is then detected using the estimated relative power. The advantages of the proposed method is that the sensitivity of detecting OS segments can be controlled simply by changing threshold value of the relative power. An experiment was carried out to evaluate the performance of this method using data of an actual meeting.

2. EIGENVALUE DISTRIBUTION AND THE SVM APPROACH

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 an 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.

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

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

where $\cdot^H$ denotes the complex conjugate transpose.

When the noise $\mathbf{n}(\omega, T)$ is uncorrelated from the source $\mathbf{s}(\omega, T)$ and the noise is spatially white,

$$\mathbf{R}(\omega) = \mathbf{A}(\omega)\mathbf{P}(\omega)\mathbf{A}^H(\omega) + \sigma\mathbf{I}, \qquad (3)$$

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

$$\lambda_1, \cdots, \lambda_M = \overbrace{\gamma_1 + \sigma, \cdots, \gamma_N + \sigma}^{N}, \overbrace{\sigma, \cdots, \sigma}^{M-N} \qquad (4)$$

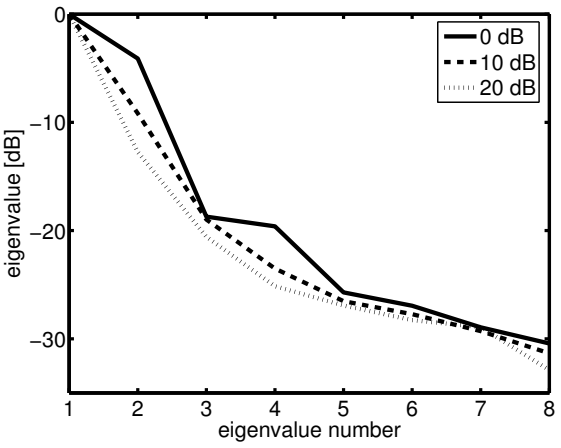

Figure 1: *Example of eigenvalue distribution.*

Assuming that the power of the source $\mathbf{s}(\omega, T)$ is greater than that of the background noise $\mathbf{n}(\omega, T)$, equation (4) shows that the eigenvalue distribution reflects information of the number of sound sources. The simplest way of estimating the number of sound sources is to count the number of eigenvalues whose value is over σ.

In a real acoustic problem in a reverberant sound field, however, the above assumptions, i.e., that $\mathbf{s}(\omega, T)$ and $\mathbf{n}(\omega, T)$ are uncorrelated and that $\mathbf{n}(\omega, T)$ is spatially white, do not hold. Therefore, the approach of counting the eigenvalues over the threshold, or of using AIC/MDL criteria does not work successfully.

We previously attempted to classify the eigenvalue distribution according to the number of sound sources using SVM [2]. The problem with this approach is that the eigenvalue distribution varies according to the relative power of the sound sources. Figure 1 shows the eigenvalue distribution of a two-sound-source case when the relative powers are 0 dB, 10 dB and 20 dB. When the difference in the relative power is large, the eigenvalue distribution becomes close to that for a single sound source, resulting in failure in detecting the second sound source with small power.

3. DETECTION OF OVERLAPPING SPEECH USING SUPPORT VECTOR REGRESSION

In this section, a method of detecting overlapping speech in a meeting situation based on information on the number of *active* sound sources is developed. For detecting overlap in conversation, precise estimation of the number of active sources (denoted as N_a hereafter), such as distinguishing $N_a = 2$ and $N_a = 3$, is not required. On the other hand, even if the relative power of one of the sources in overlapping segments is small, such segments should be correctly detected. Thus, instead of SVM, SVR is introduced to detect the overlapping segments for which the power difference is relatively large.

3.1. Relative Power Estimation by Support Vector Regression

For the sake of simplicity in explanation, it is assumed that the number of sources is two and that the power difference of these sound sources can be estimated by SVR. When the power difference exceeds a certain value, the number of active sound sources N_a is estimated as $N_a = 1$. When the power difference is within this value, $N_a = 2$.

Let us suppose that the training data set, i.e., the set of the eigenvalue distributions $\{\lambda_i\}$ and the corresponding relative power $\{d_i\}$, is available.

The regression function of SVR is written as

$$f(\mathbf{x}) = \sum_{i=1}^{l} (\alpha_i^* - \alpha_i) K(\mathbf{x}_i, \mathbf{x}) + b, \tag{5}$$

where $K(\cdot, \cdot)$ is a Mercer kernel. The symbols, α_i^* and α_i, are the optimal solution of the following optimization problem [4] ($C > 0$ and $\epsilon \geq 0$ are constants),

$$\begin{aligned} \text{maximize } & -\epsilon \sum_{i=1}^{l} (\alpha_i^* - \alpha) + \sum_{i=1}^{l} (\alpha_i^* - \alpha_i) d_i \\ & - \frac{1}{2} \sum_{i,j=1}^{l} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j) K(\lambda_i, \lambda_j) \end{aligned} \tag{6}$$

$$\text{subject to } \sum_{i=1}^{l} (\alpha_i^* - \alpha_i) = 0, \tag{7}$$

$$0 \leq \alpha_i^*, \alpha_i \leq C \ (i = 1, \ldots, l), \tag{8}$$

The symbol b is calculated using the equation

$$b = d_i - \sum_{j=1}^{l} (\alpha_j^* - \alpha_j) K(\mathbf{x}_j, \mathbf{x}_i), \tag{9}$$

in which i satisfies $0 < \alpha_i^*, \alpha_i < C$.

Figure 2 shows an example of estimation of the relative power of the two sound sources using SVR. In this figure, the lateral axis represents the sample number sorted according to the relative power. The dots represents the relative power of the training samples. The dashed line represents the corresponding estimated value. From this figure, the true value and the estimate are in good accordance within the range of $[10, 30]$ dB. On the other hand, in the range under 10 dB and that over 30 dB, the estimates were saturated. However, this saturation has little effect on estimation of the number of active sound sources, since when the power difference is over 30 dB, the second source with small power is practically inactive.

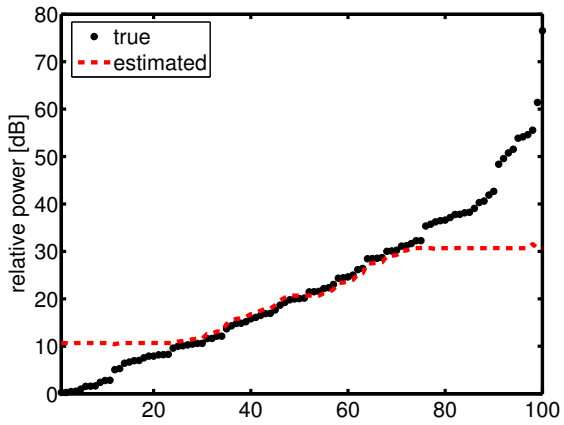

Figure 2: *Example of estimation of the relative power using SVR (training data set).*

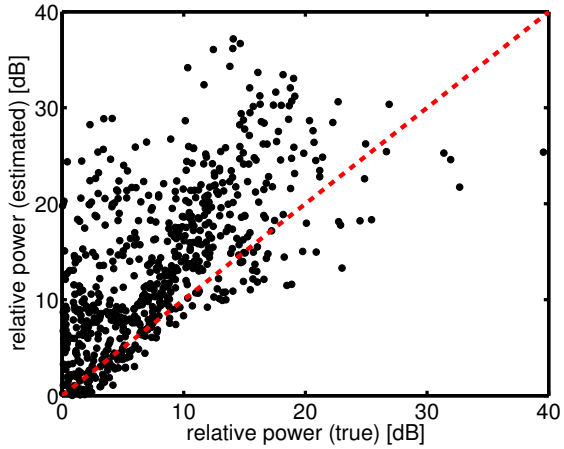

Figure 3: *Scatter plot of the true value and the estimated relative power using SVR (test data set is different from training data set).*

Figure 3 shows the result of the estimation using test data different from the training data. In this figure, the horizontal axis is the true value of the test data and the vertical axis is the estimated relative power of the test data. The variance of the estimation is greater than that with the training data. The regression coefficient was 0.848.

3.2. Detection of Overlapping Speech

Using the function (5) and the eigenvalue distribution $\lambda(\omega)$, the relative power of the two sound sources in each frequency can be estimated. Based on this, the number of active sound sources $N_a(\omega)$ at the frequency ω is determined as follows:

- $N_a(\omega) = 2$ when $\hat{d}(\omega) < P_{th}$.
- $N_a(\omega) = 1$ when $\hat{d}(\omega) \geq P_{th}$.

Here, $\hat{d}(\omega)$ is the estimated relative power of the eigenvalue distribution $\lambda(\omega)$ and P_{th} is the threshold of the power.

Using the estimate of N_a at each frequency, a histogram for $N_a = 1$ and $N_a = 2$ over the frequency range of interest is then obtained. Based on this histogram, it is determined whether the corresponding block is a single speech (SS) section or an overlapping speech (OS) section. When the number of frequency bins classified as $N_a = 2$ is greater than that of $N_a = 1$, this block is judged as an OS segment.

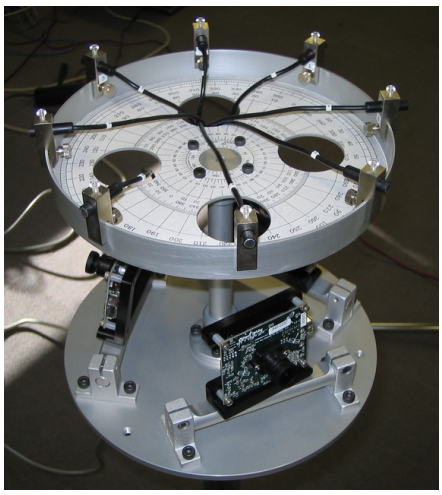

Figure 4: *Microphone array used for recording.*

Table 1: *Parameters of the experiment.*

Sampling frequency	16000 Hz
FFT length	512
FFT shift	128
Frequencies of interest	500 - 4000 Hz
Frame length	0.5 s

4. EXPERIMENT

4.1. Experimental Conditions

A Japanese market research meeting termed “Group Interview” was recorded and used for testing the proposed method. In this meeting, one professional interviewer and five interviewees (university students) participated. The interviewer asked questions such as “What types of cellular phones are you using?,” and the interviewees answered the questions in a discussion manner.

The meeting was conducted in a middle-sized meeting room with a reverberation time of 0.5 s. The six participants sat around a table. The microphone array shown in Figure 4 was located in the middle of the table and consisted of eight microphones in a circular shape with a diameter of 0.2 m. The distance from the center of the array to the participants was approximately 1 - 1.5 m.

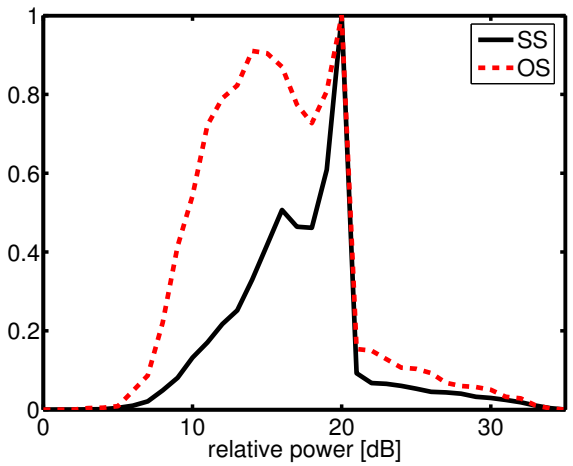

Figure 5: *Distribution of estimated relative power in SS and OS segments.*

Table 2: *Experimental results*

	R_r	R_p	R_f	F	P_{th} [dB]
SVR	0.402	0.424	0.063	0.413	15.2
	0.423	0.423	0.066	0.423	15.3
	0.433	0.404	0.073	0.418	15.4
	0.454	0.400	0.078	0.425	15.5
	0.474	0.393	0.084	0.430	15.6
	0.495	**0.393**	**0.087**	**0.438**	**15.7**
SVM	0.402	0.361	0.081	0.381	-

To obtain the training data set (λ_i) for SVR, two Japanese sentences were convolved with the measured impulse response of the meeting room in which the actual meeting was conducted. The locations of the two sources were 0 degrees and 180 degrees. For the sake of comparison, the method using SVM was also tested. To obtain the training data set for SVM, four Japanese sentences were convolved. The locations of the four sources were 0 degrees, 90 degrees, 180 degrees, and 270 degrees. The parameters of the experiment are shown in Table 1.

4.2. Experimental Results

Figure 5 shows the distribution of the estimated relative power of the SS and OS segments in the test data. From this figure, the estimates for SS segments were concentrated at a higher relative power as expected. Using this difference in distribution, it was possible to determine the P_{th} threshold. In this study, $P_{th} = 15 - 16$ dB is employed.
Table 2 shows recall rate R_r, precision rate R_p, false alarm R_f and F-measure [5] F defined as:

$$R_r = \frac{\text{Number of correctly detected OS segments}}{\text{Total number of OS segments}} \quad (10)$$

$$R_p = \frac{\text{Number of correctly detected OS segments in } N_d}{\text{Number of detected OS segments } (N_d)} \quad (11)$$

$$R_f = \frac{\text{Number of not correctly detected OS segments in } N_d}{\text{Total number of SS segments}} \quad (12)$$

$$F = \frac{2R_rR_p}{R_r + R_p} \quad (13)$$

P_{th} is the threshold used to classify the estimated relative power into SS and OS segments.
When $P_{th} = 15.7$, the F-measure achieved the best score among all of the tested parameters. In this case, the false alarm rate, R_f, was comparable for SVR and SVM, while the recall rate increased by 9% for the SVR case. From this, it can be seen that some of the OS segments which were not detected by SVM were detected by SVR without increasing R_f.

5. CONCLUSION

We have herein presented a method of detecting overlapping-speech segments by estimating the relative power of sound sources using the eigenvalue distribution and SVR. The proposed method was applied to the recorded data of an actual meeting and around 50% of the overlapping segments were detected. One of the advantages of the proposed method is that the sensitivity of detecting OS segments can be controlled simply by changing P_{th}. In the theoretical part of this paper, only the case of two sound sources was treated. In the test data used in this paper, however, a case with more than two sources was included and OS segments were detected to some extent. The theoretical aspects of the case with more than two sources should be addressed in the future.

6. REFERENCES

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

[2] K. Yamamoto et al., in *ICASSP 2003*, April 2003, pp. V–485–V–488.

[3] B. Scholkopf et al., *Advances in Kernel Methods: Support Vector Learning*, MIT Press, 1999.

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

[5] C. J. van Rijsbergen, *Information retrieval (second edition)*, Butterworths, 1979.