SOUND SOURCE CLASSIFICATION USING SUPPORT VECTOR MACHINE

Makoto KUMON * Yoshihiro ITO *
Toru NAKASHIMA * Tomoko SHIMODA *
Mitsuaki ISHITOBI *

* *Department of Intelligent Mechanical Systems,*
Graduate School of Science and Technology,
Kumamoto University,
2-39-1, Kurokami, Kumamoto, 860-8555, JAPAN
E-mail: kumon@gpo.kumamoto-u.ac.jp

Abstract: This paper shows an application of a learning method for acoustic signal classification by an auditory robot. The learning approach provides an unified acoustic signal classification method without considering the characteristics of target signals. Support Vector Machine was adopted to obtain the classifier and the target signal was characterized by Mel-Scale Log Spectrum which was a general form to symbolize acoustic signals. Results of actual experiments to classify 4 class of acoustic signals at single sound source case and to classify 3 class of acoustic signals at plural sound source case showed the validity of the method.

Keywords: Automatic Recognition, Learning Systems, Robot Audition, Sound Source Classification, Support Vector Machine

1. INTRODUCTION

Oral communication is one of the most natural modality for human-machine interface which is getting significantly important for recent robot systems working close to human since we, human, do utilize vocal modality to interact each other in daily life. Softwares to utilize auditory information such as speech recognition, e.g. Julius(Kawahara and Lee, 2005), have been developed. Those developed systems, however, rely on "well-recorded vocal information". This leads only the target voice without noise must be recorded. Hence, users are often required to attach special devices such as head-sets and it is not really "natural" as we are chatting in the real environment where multiple non-target sound sources exist. In order to overcome this difficulty, methods to extract the target sound signal have been proposed, for example, utilizing microphone arrays, beam forming, independent component analysis and so on. Adding to this, it is needed to recognize the sound source of interest among multiple sound sources and the method to achieve this efficiently is also required. Human voice is one of the most important signals to recognize, and Lu(Lu *et al.*, 2001) proposed a method to detect human voice with High Zero-Crossing Rate Ratio, Low Short-Time Energy Ratio, Spectrum Flux, and frequencies of poles and zeros of Linear Predictive Coding model. Kim(Kim *et al.*, 2006) utilized sound location, acoustic cues like Lu's approach, face recognition and optical flow for detecting a moving person. The last two cues were realized by using a CCD camera and those multi-modal cues were integrated by a particle filter with Gaussian Mixture Model. The method succeeded to follow the target person under existence of another person and music. Both of those methods depend on the models which are given by the designer *a priori*. For example, Kim(Kim *et al.*, 2006) derived a model by assuming that all sound signals could be classified to one of human voice, music or noise. On the other hand, there are a lot of acoustic symbols to recognize such as the sound of phones' ringing bell, the acoustic tone by traffic signals for blind people, dogs' barking and so on, in the real world. As the number of sound signals to recognize increases, it becomes difficult for designers to derive appropriate models to classify all sound classes precisely and systematic approaches to obtain sound source classifier are required.

The classification problem which gives a method to design a criteria from a given data set has been studied for a long time. For example, self organizing map is able to recognize the class of the given data after the learning process. Perceptron is another example which derives a classification function. Support Vector Machine(SVM)(Vapnik, 1998; Cristianini and Shawe-Taylor, 2000) is one of the derivatives of perceptron which has a basis on optimization approach. SVM is said to be effective(Cristianini and Shawe-Taylor, 2000) in solving the classification problem since it has an ability to avoid local solutions through the learning process and the obtained information can be structured in a simpler model. This method may have a potential to simplify the procedure in obtaining the classifier for sound sources. The main objective of this paper is to reveal the applicability of the learning approach for sound source classification with real experiments.

This paper is organized as follows. In the next section (Sec. 2), the acoustic cue utilized in this paper is introduced. In Sec.3, a brief summary of SVM is given following the reference (Cristianini and Shawe-Taylor, 2000). By applying SVM for sound source classification, results of experiments is shown in the section 4 which is the main contribution of this paper. Then conclusion follows in the section 5.

2. ACOUSTIC CUE FOR SOUND SOURCE CLASSIFICATION

In this paper, sound signals utilized in humans daily life are considered. This implies that it can be assumed that sound signals of interest have the following properties:

(1) signals sustain for a certain period.
(2) frequency response is able to be assumed constant for about 50msec.
(3) power of the signal is limited in a certain frequency range such as our range of hearing.

Usually it is much efficient to extract acoustic cues to process instead of processing the sound signal itself when those above assumption can hold. Actually, cues based on short time frequency analysis such as linear predictive coding(LPC) or Mel-Frequency Cepstrum Coefficient(MFCC) are utilized for speech recognition. Zero-Crossing Rate Ratio, Low Short-Time Energy Ratio and Spectrum Flux are also possible candidates for sound source classification as mentioned in the introduction. Although those special cues are powerful for sound source classification, the measured signal is not only used for sound source classification but also for other objectives such as the speech recognition when we consider the application to auditory robots. Therefore, sound cues which are specific only for sound classification are not suitable because of the computational efficiency. In this paper, Mel-Scale Log Spectrum(Takeda *et al.*, 2006), which will be denoted as MSLS, is utilized as the acoustic cue for sound source classification since MSLS is commonly used for speech recognition.

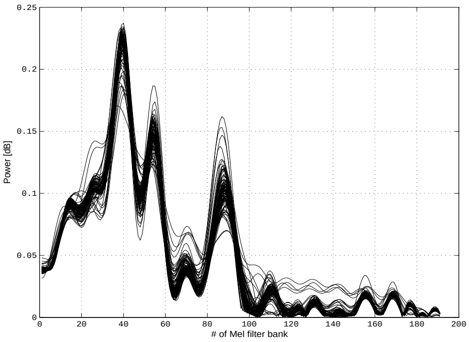
MSLS is defined by the log power spectrum of a short time sound signal which is processed by some appropriate window (hamming window is adopted in this paper) and it is defined on the Mel-frequency domain(Imai, 1996). Mel frequency is defined by the following nonlinear frequency transformation:

$$Mel(f) = 2595 \log \left(1 + \frac{f}{700} \right).$$

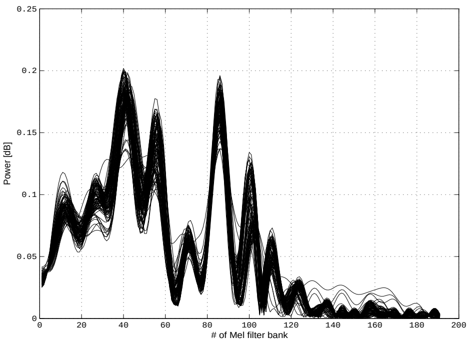
For speech recognition, the sampling frequency is commonly selected to about 16kHz, and the Mel-filter bank with about 25 filters is utilized. In this paper, the sampling frequency for signal processing is broadened to 40kHz. 1024 points are processed as the frame of the frequency analysis. 191 filters are adopted for Mel-filter bank and the moving average on the Mel-frequency domain is adopted in order to reduce the effect of the noise.

Examples of measured MSLS by two male subjects uttering "a" in Japanese are shown in Fig.1. The frame which has the power more than the given threshold is considered as the active frame which contains the vocal information. For those frames, the mean of the frame is subtracted in order to remove the offset, and processed to derive the MSLS of the frame. In Fig.1, all active frames are shown. In both cases, stable MSLS are shown and it can be found that some peaks or notches characterize the subject although the rough shape of curves resemble each other.

In the following, a method to classify the sound source by the measured MSLS is introduced. Basically, the method matches the pattern of the measured MSLS and the MSLS of the memo-



Subject A



Subject B

Fig. 1. MSLS of subjects (male) : /a/

rized sound source. SVM is utilized to obtain the classification function in the matching process by considering the MSLS of the frame as one data point.

3. SOUND CLASSIFICATION BY SUPPORT VECTOR MACHINE(SVM)

SVM is a method of machine learning that extends linear classification approach with implicit non-linear transformation, so-called "*Kernel trick*", for nonlinearly separated data sets. Classification model is given as a discriminant function, e.g. $\mathrm{sgn}(\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}) - \rho)$, where $\boldsymbol{x}$ shows the feature to classify and $\boldsymbol{w}$, $\boldsymbol{\phi}$ and ρ are parameters learned by SVM. The function $\boldsymbol{\phi}(\boldsymbol{x})$ gives the nonlinear transformation to make the set of data points becomes linearly separated. SVM allows this procedure without explicitly finding the nonlinear transfer function of $\boldsymbol{\phi}$. Parameters $\boldsymbol{w}$ and ρ are obtained through optimization of dual problem. In most cases, not the whole training data points but only a part of them are needed to derive the discriminant function. This implies that the solution of SVM learning process may provides a simpler model than other methods(Cristianini and Shawe-Taylor, 2000).

The proposed framework is sketched in the figure 2. As mentioned in the previous section, the sound signal measured by the robotic head is sampled at 40kHz rate and digitized with 12bit

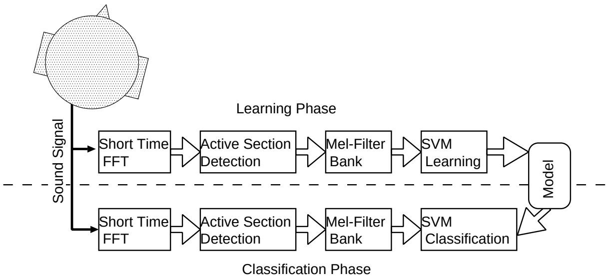


Fig. 2. Proposed SVM based Sound Source Classification Scheme

resolution. 1024 points of the signal is treated as one frame and the signal is transformed to the frequency domain by Fast Fourier Transformation after filtered by a hamming window. Then, the 1024 element is processed by a filter bank of 191 filters to obtain MSLS. Hence, the feature vector of one frame contains 191 elements. The feature vector is brought to SVM as is described below and SVM memorizes and classifies the signal at the learning phase and at the classification phase respectively.

MSLS of one active frame contains 191 data points and it is treated as a feature vector for SVM. Let T_S represent the training data set with respect to the subject S. i-th element of S is defined as $(\boldsymbol{x}_i, y_i)$, where $\boldsymbol{x}_i$ represent the MSLS vector and where y_i is 1 if the data belongs to the subject S and -1 if not. Denote the number of the training data as N. Consider the following minimization problem:

$$\begin{aligned}&\min_{\boldsymbol{\xi}, \boldsymbol{w}_S, \boldsymbol{\rho}_S} < \boldsymbol{w}_S, \boldsymbol{w}_S > + C\sum_{i=1}^{N} \xi_i^2 \\ &\text{subject to} \\ &\quad y_i\,(< \boldsymbol{w}_S, \boldsymbol{x}_i > + \rho_S) \geq 1 - \xi_i \quad \text{for} \ \ \forall i,\end{aligned}$$

where $< \boldsymbol{x}, \boldsymbol{y} >$ represents an inner product of vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ and ξ_i is a slack variable. C represents the positive constant which controls the sensitivity to noise over the training set. The constraint assures that the function $< \boldsymbol{w}_S, \boldsymbol{x}_i > + \rho_S$ gives the same sign of y_i if ξ_i is close to zero, which implies that the function with parameters $\boldsymbol{w}_S$ and ρ_S is a classifier of the subject S. The minimization aims to find the discriminant hyperplane far from all training data which implies that the obtained discriminant function is expected to be robust.

From the theory of SVM(Cristianini and Shawe-Taylor, 2000, Proposition 6.11), the above optimization problem can be transformed into a dual problem of the parameter $\boldsymbol{\alpha}$ instead of $\boldsymbol{w}_S$ as follows:

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} y_i y_j \alpha_i \alpha_j < \boldsymbol{x}_i, \boldsymbol{x}_j > + \frac{1}{C} \delta_{ij}$$

subject to

$$\sum_{i=1}^{N} y_i \alpha_i = 0, \quad \alpha_i \geq 0 \text{ for } \forall i,$$

and the discriminant function, denoted as $f(\cdot)$, is given as

$$f(\boldsymbol{x}) = \operatorname{sgn} \left(\sum_{i=1}^{N} y_i \alpha_i^* < \boldsymbol{x}, \boldsymbol{y} > + b^* \right), \qquad (1)$$

where α_i^* is the i-th element of the optimal solution of α and b^* is the value such that $y_i f(\boldsymbol{x}_i) = 1 - \frac{\alpha_i^*}{C}$ $(\alpha_i^* \neq 0)$ for any i. Furthermore, the inner product $< \boldsymbol{x}, \boldsymbol{y} >$ can be generalized as "kernel" which is possible to be utilized as generalized distance. Gaussian radial basis function is adopted as the kernel in this paper:

$$< \boldsymbol{x}, \boldsymbol{y} >= \exp\left(-\gamma \|\boldsymbol{x} - \boldsymbol{y}\|^2\right),$$

where γ is a positive parameter which represents the radius of the basis.

Once the above optimization problem is solved, it is possible to compute Eq.(1) with the obtained parameters $\boldsymbol{\alpha}^*$ and b^* for the measured MSLS data $\boldsymbol{x}$. If Eq.(1) is 1, the measured data $\boldsymbol{x}$ is classified as the sound signal of the subject S and if Eq.(1) is -1, it is not.

From the above, only the training data of $\alpha_i^* \neq 0$ can effect to the discriminant function and those vectors are named as support vectors. SVM parameters γ and C are needed to be tuned. For the simplicity C is kept 1 in this study. In order to obtain a simple model for realtime classification, in the following experiment, γ was tuned as the number of support vector becomes small through the learning process.

4. EXPERIMENTS

The above algorithm was evaluated by an actual auditory robot shown in the figure 3. Two microphones were installed inside the head of the robot. Only the right microphone was used in the single sound source case and both were utilized in the plural sound source case.

Two male subjects, one female subject and pseudo white signal which was generated by a loud speaker were considered as sound sources to classify. Japanese vowels "a", "i", "u", "e" and "o" were uttered for a several 10 seconds by subjects. The sound signals were recorded by a computer and generated by loud speakers as sound sources. Two data sets of each subject were recorded. The one was used for the learning and the other was used for the test. Although the test data only contains "a" vowel, the system was not informed which vowel would be uttered. The active frames were scanned and MSLS for those frames were computed. Then SVM learned MSLSs for each speaker as it did not distinguish the difference of vowels but the difference of subjects. The total active section of Subject A, B, C and White for learning were 542, 340, 400 and 492 respectively. Following the criteria described above, the parameter γ was defined as 10 after several trials so as to make the number of support vectors small. The number of obtained support vectors was 326 which was only 18.4% of the whole training data.

Fig. 3. Auditory robot(left: front view, right: side view)

The classifier based on the model obtained with the training set was tested in 1) single sound source case and 2) multiple speaker case as shown below. The total active section of Subject A, B, C of test data sets were 468, 168 and 165 respectively.

4.1 Classification of Single Sound Source Case

The result of the experiment is shown in Table1. Subject A and B represent two males and C represents the female subject. White shows the case of the pseudo white signal by the loud speaker. Each row shows the ratio that the classified result by the algorithm for a given sound source. Human voice and pseudo white signal were distinguished perfectly. Signals of subjects were also well clas-

Table 1. Result of sound source classification experiment

		Classified Class			
		Subject A	Subject B	Subject C	White
Sound Source	Subject A	99.57	0.43	0.0	0.0
	Subject B	0.0	100.0	0.0	0.0
	Subject C	0.0	0.0	100.0	0.0

sified and the algorithm succeeded to find the correct answer almost perfectly. Only the case when the Subject A voice was tested, 0.43% of the signal was classified as Subject B's voice. This could happen because both Subject A and B were male and had close vocal patterns. The result showed that the method had an ability of sound source classification.

The algorithm was also implemented for real time classification by using libSVM API. Results of experiments also showed the effectiveness of the approach.

4.2 Classification of Plural Sound Source Case

The proposed algorithm was also tested for plural sound source cases. Sound signals were same as at the single case. Vowel 'a' of two subjects and a pseudo white signal were used. Although it rarely happens that two persons utter the same vowel for a while in natural conversation, it is a good challenge for the system to distinguish plural vocal signals generated simultaneously because the performance of sound classification can be measured without the help of other cues such as dynamic sequence of sound transition. Different to the above single sound case, both of left and right microphones were utilized in both learning phase and classification case. Assuming that the system had information about the location of the sound source, the sound signal from the nearer microphone was used to classify the sound. Since the objective of these experiments was to reveal the classification performance under the contamination of two signals, the learned signals themselves instead of the test signals were utilized as the test patterns. Models were computed for each microphones and utilized for the classification separately. The parameters γ were also tuned. As it was shown in the single case, the difficult combination of subjects, i.e. A and B case, was tested with the model learned subject A, B and the pseudo white signal. All signals were generated one by one from 30 degree right and left to the robot head.

The robot head has pinnae, or outer ears, and they perform as filters to effect the frequency charac-

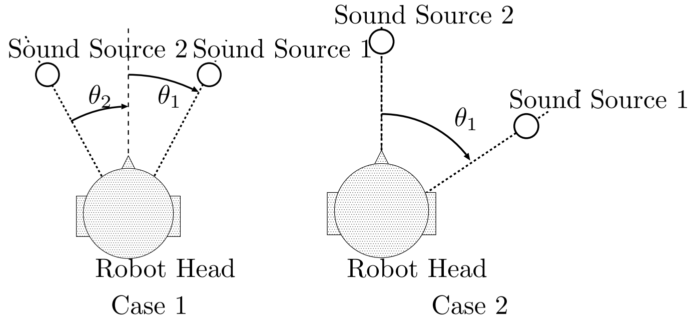


Fig. 4. Plural Sound Source Classification Tests

Table 2. Result of sound source classification experiment

		Classified Class	
		Subject A	Subject B
Sound Source	Subject A as Source 1	94.2	5.8
	Subject B as Source 2	0.9	99.1
Sound Source	Subject A as Source 2	38.8	61.2
	Subject B as Source 1	5.6	94.4

teristics of signals depending on the location of the sound source(Shimoda *et al.*, 2006). The head itself also has same kind of acoustic effect and the location of the sound source may effect the performance. Taking the fact into account, the performance tests were carried out in several cases with different location of the sound sources.

Figure 4 shows two cases of the experiment. Sound sources were located in front of the robot with azimuthal angle of θ_1 and θ_2 from the median plane. The distance from the robot to the speakers was almost 0.5m. Case 1 shows the situation that two sound sources located left and right to the robot and Case 2 shows more natural situation that the robot faced to the one speaker and the other speaker stayed at the right to the robot.

4.2.1. Case 1 The experiment result of the case when two sound sources, Subject A and B were located at $\theta_1 = 30$[deg] and $\theta_2 = -30$[deg] respectively, is shown. The experiment with swapping sound source location were also conducted.

The difference between the training set and the test pattern in this situation was that sound signals were generated simultaneously at the test experiments in order to clearify the effect of the crosstalk. The result is summarized in Table 2.

The method succeeded to classify signals propely except when the sound of the subject A coming from left(source 2). The right microphone, which would be utilized for main purposes on other tasks, showed acceptable performance. Since the robot's left microphone was less sensitive to the right because of its structure, it could happen that

Table 3. Result of sound source classification experiment

		Classified Class	
		Subject A	Subject B
Sound Source	Subject A as Source 1	88.0	12.0
	Subject B as Source 2	0.3	99.7
Sound Source	Subject A as Source 2	24.5	75.4
	Subject B as Source 1	47.8	52.2

the model for the left microphone was not well tuned through the learning process.

4.2.2. Case 2 The second case, when the robot was pointed to the one sound source and the other was located right to the robot ($\theta_1 = 60$[deg]), was more difficult than the first case because two source were closer and because the location of one sound signal was differed from the training set. As same as the case 1, the experiment with swapping the location of sound sources was also conducted.

The result is summarized in Table 3 as in Table 2. The performance was worse in most cases.

In both two cases, the method succeeded to distinguish human voice from the white signal. The preformance could be improved by tuning learning parameters γ and C if the test signal could be utilized. When $\gamma = 0.0035$ and $C = 1$, the worst performace of the above case was 77.1%, which is rather acceptable. Alghouth this approach was not practical since it required the test signals, the possibility to use SVM based discriminant function could be seen.

5. CONCLUSION

This paper shows the method to classify the sound signal into one of the known sound pattern by utilizing SVM. Features to learn was selected as Mel-Scale Log Spectrum of a short time sound signal because it was able to be re-used for speech recognition which is one of the most important application of the auditory information for robots. Actual experiment showed that the algorithm succeeded to classify four sound sources appropriately for single sound case and that it still had the ability to classify plural sound signals to some extent.

For further study, it has to be considered that the limitation of the method as the number of target sound signals increases. The learned model including support vectors is also needed to be studied to find the theoretical background of the sound source classification. In this paper, the algorithm was only tested in a silent room where the test signal was dominant compared to the background noise. The algorithm is also needed to be tested even in the more natural, noisy environment.

REFERENCES

Cristianini, N. and J. Shawe-Taylor (2000). *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge University Press.

Imai, S. (1996). *Audio Sound Signal Processing(in Japanese)*. MORIKITA SHUPPAN.

Kawahara, T. and A. Lee (2005). Open-source speech recognition software julius. *Journal of the Japanese Society for Artificial Intelligence* **20**(1), 44–49.

Kim, H-D., K. Komatani, T. Ogata and H.G. Okuno (2006). Real-time auditory and visual talker tracking through an EM algorithm. In: *Proc. of the 24th Meeting of Special Interest Group on AI Challenges*. Kyoto, Japan. pp. 51–58.

Lu, L., H. Jiang and H-J. Zhang (2001). A robust audio classification and segmentation method. In: *Proc. of the ninth ACM International Conference on Multimedia*. Ottawa, Canada. pp. 203–211.

Shimoda, T., T. Nakashima, M. Kumon, R. Kohzawa, I. Mizumoto and Z. Iwai (2006). Spectral cues for robust sound localization with pinnae. In: *Proc. of the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*. pp. 386–391.

Takeda, R., S. Yamamoto, K. Komatani, T. Ogata and H.G. Okuno (2006). Evaluation of the ICA BSS and missing feature theory based-asr with simultaneous speech recognition (in japanese). In: *Proc. of the 24th Meeting of Special Interest Group on AI Challenges*. Kyoto, Japan. pp. 9–16.

Vapnik, V.N. (1998). *Statistical Learning Theory*. John Wiley & Sons.