The 2009 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 11-15, 2009 St. Louis, USA

Daily Sound Recognition Using Pitch-Cluster-Maps for Mobile Robot Audition

Yoko Sasaki[1], Masahito Kaneyoshi[1], Satoshi Kagami[1], Hiroshi Mizoguchi[2,1] and Tadashi Enomoto[3]

Abstract— **This paper proposes a sound identification method for a mobile robot in home and office environment. We propose a simple sound database called Pitch-Cluster-Maps(PCMs) based on Vector Quantization approach. Binarized frequency spectrum is used for PCMs codebook generation. It can describe a variety of sound sources, not only voice, from short term sound input. The proposed PCMs sound identification requires several tens(msec) of sound input, and is suitable for a mobile robot application which condition is dynamically changing. We implemented the proposed method on our mobile robot audition system equipped with a 32ch microphone array. Robot noise reduction using proposed PCMs recognition is applied to each input signal of a microphone array. The performance of daily sound recognition for separated sound sources from robot in motion is evaluated.**

I. INTRODUCTION

This paper considers sound identification and verification in the home, office room and factory environment. In such situations, there are many sounds/noises as well as the voices of people around a robot. It is necessary for a robot system to respond to a correct sound selectively. We expect for a future robot to have the functions of sound segregation, sound recognition, speaker recognition and speech recognition.

Focusing attention on sound recognition, there are many researches on speaker recognition [1], [2], the Gaussian Mixture Model (GMM) is traditionally used as a speaker model for speaker recognition and the speaker model is trained by the 25 feature vectors:12MFCC, 12 ΔMFCC and 1Δ power. High speaker recognition rate over 95(%) are obtained for the normal speaking rate utterance.

In computational auditory scene analysis (CASA), some progress has been achieved using a multi-pitch tracking algorithm in monaural segregation and sequential organization [3], [4], but the issue of sequential grouping still remains. On the other hand, in the analysis of musical audio signals [5], the singer is identified using Goto's PreFEst method and the harmonic structure of music [6], and the same method is also applied to speaker identification [7].

As for non-voice sound source recognition, Chen et. al. developed a daily sound recognition system using room microphones attached to the environment [8]. Lukowicz et. al. proposed a workshop activity recognition system using wearable computer equipped with microphones and acceleration sensor [9]. For humanoid robot application, daily sound recognition using principal component analysis of cepstrum data is proposed [10]. On the other hand, it is not verified for environmental changes and multiple sound sources condition.

We propose a new simple method using a database (DB), which we call Pitch-Cluster Maps (PCMs). It adopts Vector Quantization (VQ) approach for real time processing, and the method uses only instant pitch data (not using time sequence information). One PCM is constructed from a known sound, which is defined in Section II-A. The processing method of sound recognition as a kind of open-set identification is discussed in Section II-B.

Short term input requirement is the feature of the proposed system. The PCMs sound identification is applicable for several tens(msec) of sound input, and is suitable for a mobile robot application which conditions are dynamically changing. Some applications for mobile robot are performed in Section V.

Experiments of speaker and daily sounds recognition have been done in two kind of real rooms with reverberation. These experiments have been done using a microphone array embedded on a mobile robot. Detailed results are in Sections IV and V. The final section is devoted to some discussion, as well as, some future works.

II. SOUND IDENTIFICATION ALGORITHM

A. Pitch-Cluster-Maps (PCMs)

The voices of known persons and several kinds of known sounds are used to make the pitch-cluster maps database as references. A known sound signal is segmented into N-points short-time Fourier transforms (STFTs).

$S(f, tn)$ are the absolute value of STFTs of the sound where $f \in (1, 2, \cdots, N) \times (f_s/N)$is a discrete frequency. f_s is a sampling frequency and $tn = (1, \cdots, T)$is a time-frame index. The amplitude $S(f, tn)$ is converted to $FS(f, tn)$ where $FS(f, tn) = 1$ if $S(f, tn)$is larger than a threshold P_{thld} and $FS(f, tn) = 0$ in the other case. Binarized value FS $(1 \le i \le N/2, 1 \le j \le T)$ is described in Eq. (1)

$$FS(f_i, j) = \begin{cases} 1 & \text{if } S(f_i, j) \ge Pthld_j \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Binarization threshold $Pthld_j$ for each time slot is defined as Eq. (2):

$$Pthld_j = med[S(f, j)] + std[S(f, j)] \tag{2}$$

where $med[\cdot], std[\cdot]$ denotes median value and standard deviation.

[1] 1: Digital Human Research Center, National Institute of Advanced Industrial Science and Technology, 2-41-6, Aomi, Koto-ku, Tokyo, 135-0064, Japan. {y-sasaki, m-kaneyoshi, s.kagami}@aist.go.jp

[2] 2: Faculty of Mechanical Engineering, Department of Science and Technology, Tokyo University of Science, 2641 Yamazaki, Noda-shi, Chiba, 278-8510, Japan. hm@rs.noda.tus.ac.jp

[3] 3: Kansai Electric Power Co. Inc., 3-11-20, Nakoji, Amagasaki, Hyogo, 661-0974, Japan. enomoto.tadashi@b5.kepco.co.jp

978-1-4244-3804-4/09/$25.00 ©2009 IEEE

These $M \times N$ elements of FS-data are grouped into K-clusters by k-means method as follows :

$$FS(f, tn) \longrightarrow \{CN(tn), CF(f, cn), CD(f, cn)\} \quad (3)$$

where $cn = (1, \cdots, K)$ is the number of cluster and K is a constant number such as 26(a to z) for human voice. $CN(tn)$ is a cluster number$(1, \cdots, K)$ which data of a time-slot tn belongs to. $CN(tn)$ may be utilizable as a time series index of known sound elements. $CF(f, cn)$ is the center-cluster frequency spectrum of each cluster cn. $CD(f, cn)$ is the frequency distance from the cluster center. $CF(f, cn)$ is mainly used as references of grouping of input sound in the following sections.

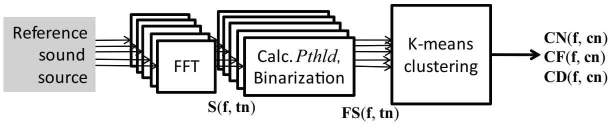
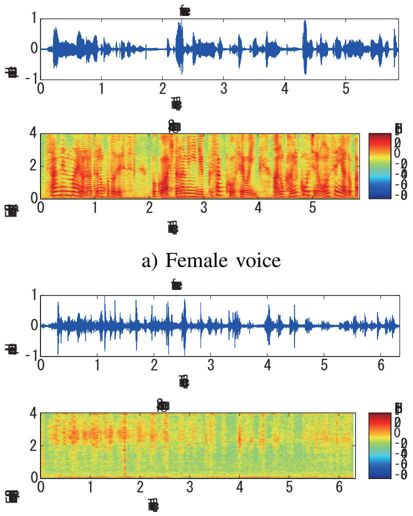


Fig. 1. PCMs generation flow

Examples of PCM models for female voice and jingled keys are shown in Fig. 2, 3 and 4. First, reference sound sources, not including other signals, are collected for PCMs model generation (Fig. 2). From collected sound sources, the pitch spectrum FS is generated. Fig. 3 shows FS-data using a blue dot when $FS(f_i, j) = 1$ after binarization. The centered spectrum CF of grouped FS-data for each cluster is shown in Fig. 4. For these examples, cluster number $K = 10$, data length $N = 1024$, and sampling frequency $fs = 16$(kHz).



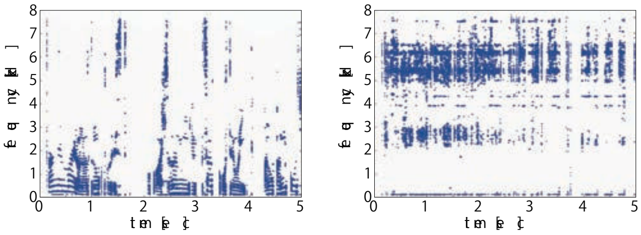
a) Female voice

b) Jingle keys

Fig. 2. Examples of PCMs model generation:Time waves of reference sound sources

B. Sound Identification based on PCMs-DB

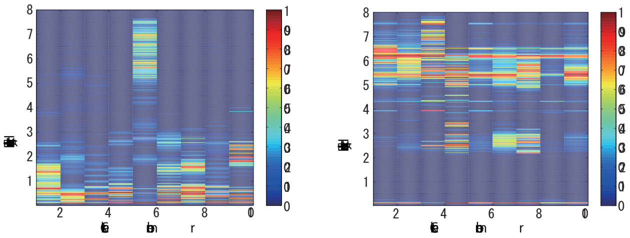
Sound recognition is performed for the input signal after sound localization and separation. $I(f, tn)$ which are the absolute amplitude after STFTs are converted to $IF(f, tn)$ using Eq. (1).



a) Female voice

b) Jingle keys

Fig. 3. Examples of PCMs model generation:Pitch spectrum FS after binarization



a) Female voice

b) Jingle keys

Fig. 4. Examples of PCMs model generation:Center spectrum CF

Similarity-distance estimation: $SD_k(tn, cn)$ is performed as the spectral square-root range calculation between input N-points STFTs $IF(f, tn)$ and the PCMs model $CF_k(f, cn)$.

$$SD_k(Mn) = sqrDistance(IF(f, tn), CF_k(f, cn)) \quad (4)$$

where $sqrDistance(x(f, i), y(f, j))$ stands for matrix calculation of the square-root distance between $x(f, i)$ and $y(f, j)$ at the same-frequency bin f, and $Mn = (1, 2, \cdots)$ is model number. The minimum distance $SD(Mn)$ for each model is defined as Eq. (5).

$$SD(Mn) = min(SD_k(Mn)) \quad (5)$$

Using this similarity-distance, the candidate model number at each time-slot tn is decided as follows:

$$G_{SD} = argmin_{Mn}(SD(Mn)) \quad (6)$$

As a final decision, if the spectral distance $SD(G_{SD})$ is smaller than $Thld_{match}$ which is described in Eq. (7), input signal is recognized to model index G_{SD}, otherwise it is recognized that it is not included in the model codebook. Recognition threshold $Thld_{match}$ is defined in Eq. (7):

$$Thld_{match} = meanDist(G_{SD}) + 3 \times stdDist(G_{SD}) \quad (7)$$

where $meanDist(\cdot), stdDist(\cdot)$ denotes average and standard deviation of square-root distance between corresponding CF and FS including the cluster in PCMs model generation.

C. Time-slot Voting

When the above condition is achieved, the sequential voting-box $V(m)$ is incremented +1. $V(m)$ is a voting box for the m-th sound reference. After a few second processing, $V(m)$ is divided by the number TSN which is the total time-slot number of signal existence state in the processing period. The cumulative voting recognition result VR is calculated from the voting rate of the index-number $m(m = 1, \cdots, RN)$ as follows:

$$VR = argmin_m(V(m)/TSN) \qquad (8)$$

D. PCMs based Robot Noise Reduction

For a mobile robot application, self generated noise reduction is an important function. Robot noise reduction for each microphone input is performed using a PCM model of robot noise. PCM model for robot noise CF_R is generated using recorded sounds from robot embedded microphone array in several motion condition when other sound sources are not existing. Let $X(\omega)$ be a input signal of microphone array in frequency domain, the signal after robot noise reduction is described in Eq. (9).

$$X'(\omega) = X(\omega) - CF_R(\omega, cn_{min}) \times X(\omega) \qquad (9)$$

III. MICROPHONE ARRAY BASED ROBOT AUDITION SYSTEM

This section describes the mobile robot audition system, which used for the proposed PCM recognition. Separated sound sources from a mobile robot are used for input signal of PCM sound recognition. Some experimental results are explained in section V.

A. Sound Localization and Separation

Sound localization and separation is based on Delay and Sum Beam Forming(DSBF) using a microphone array.

Main-Lobe Fitting(MLF) [11] is used for multiple sound source localization. The method detects point sound sources using the main-lobe model of a microphone array. Selecting reliable peaks by MLF rejects deformed peaks caused by reflection or interference, or deformed peaks between two close sound sources.

Frequency Band Selection(FBS) [12] is used for localized sound source separation. The method assumes that the frequency component of each existing sound source are not overlapped, and separate the target sound sources by selecting larger frequency components from DSBF enhanced signals.

Proposed sound recognition is applied to separated sound sources at each time slot. Fig. 5 shows the calculation flow of our robot audition system including PCMs recognition.

B. The Robot Embedded Microphone Array

The proposed sound recognition system is tested using a 32 channel microphone array attached on the mobile robot. Fig. 6 shows the mobile robot embedded microphone array and its microphone arrangement. The microphone array has 32 omni-directional electlet condenser microphones and can sample 32 data channels simultaneously. Sampling frequency is 16(kHz) and resolution is 16(bit).

For calculation, data length is set to 1024 points and shift length is 512 points. Each 64(msec) of sampled data is recognized every 32(msec).

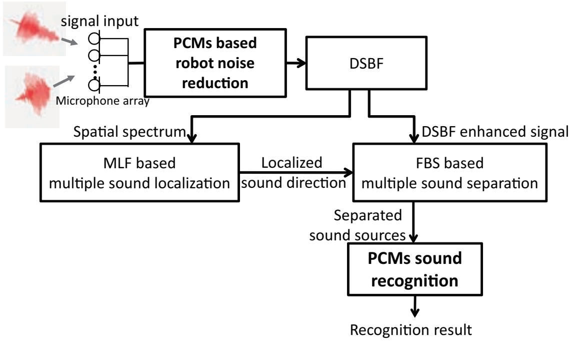
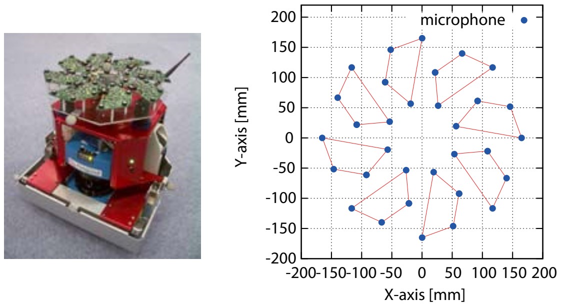


Fig. 5. PCMs recognition flow at a time slot applied for a robot audition system



a) Mobile Robot b) Microphone arrangement

Fig. 6. 32ch microphone array attached on a mobile robot: pen2r

IV. BASIC IDENTIFICATION EXPERIMENT

In this section, the proposed sound recognition is evaluated using clean data (not separated sound sources).

A. Speaker Identification

Evaluation for speaker recognition is performed by using DET-curve plotting software[1]. For PCMs-db generation, A total of 20 people's voice; 10 male and 10 female, are selected from JEITA a monosyllabic speech database. The cluster number K is set to 25, sampling frequency fs = 16(kHz) and data length $N = 1024$ points(64(msec)).

The evaluation is performed for the monosyllabic voice data, which is not used in PCMs-db generation. Fig. 7 shows the speaker recognition result of 40 monosyllabic voice data for each person. It shows adequate performance needed to identify the speaker from several words utterance.

B. Daily Sound Recognition

An experiment on recognition for non-voice sound sources was performed. A total of 232 sound sources existing in living environment are used for evaluation. The reference data is recorded using Roland WAVE/MP3 Recorder R-09. Sounds of possible action sequences are recorded, and each data contains different frequency spectrum. For example, water sound covers different flow rate, and curtain sound covers varied patterns of opening/closing motion. The recording

[1] NIST Speaker Recognition Evaluation
http://www.nist.gov/speech/tests/sre/2008/index.html
Tool:DETware v2.1(DET-Curve Plotting software for use with MATLAB)

TABLE I

DAILY SOUND DATA FOR PCMS CODE BOOK EVALUATION AND RECOGNITION RATE

Data set	Num,	Grp.	Data example	recognition rate(each/grp.)	
Kitchen	27	23	water running, fridge/bin closing, boiling water, dish care	0.894	0.937
Dining	45	20	pour water, clatter dishes, plates, broken glass	0.921	0.936
Bottle	32	19	open a can, (un)cap bottles, crash empty can	0.923	0.959
Cleaning	25	21	clothes washer, cleaner, put water in a bucket,	0.901	0.896
Bag	14	4	plastic garbage bags, snack box open, cardboard box	0.897	0.938
Door	27	8	door open(steel/wood), turn doorknob, mailbox closing	0.822	0.798
Sliding Door	33	13	sliding screen(fusuma/shoji) closing, shutter	0.704	0.794
Window	18	5	window/screen closing, knock glass window	0.712	0.769
Curtain	11	4	curtain closing, raise/lower blinds, accordion curtain open	0.960	0.994
Total	232	117		0.794	0.813

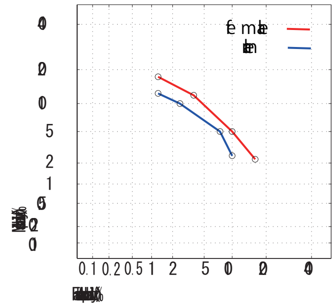


Fig. 7. Evaluation result for speaker recognition

TABLE II

RECOGNITION TIME FOR 64 (MSEC) DATA

Num of model	ave.(msec)	max. (msec)	max.(msec)
2	1.21	1.89	1.17
10	3.81	7.88	3.63
230	20.08	29.9	25.84

environment is the experimental housing space ”Holone”, which has 5 rooms in 14 × 10(m) area. The reverberation time(RT$_{20}$) is 250(msec) and the background noise level (L_A) is 35(dB). The recorded sound length is 4.4(sec) in average (0.2(sec) minimum and 30.8(sec) maximum), and SNR from back ground noise level is more than 25 (dBA). For PCMs model generation, cluster number K is set to 10, and other parameters are same as the above experiment.

Table I shows the data set and its recognition result. its includes various technique From 232 data samples, some sound sources are combined into 1 group when the sound is generated from same situation, such as window opening and closing, clatter sound of different dishes, and different curtain closing sound. In Table I, the second column shows the number of recorded data and the third column shows the number of sound groups. The fifth column shows sound recognition rate for every PCMs codebook including that data set from each recording data, and the sixth column shows that of grouped data. The last line is the result for the PCMs codebook including all data sets. From 117 PCMs models, 80 (%) of input signals are correctly recognized on average.

C. Calculation Time

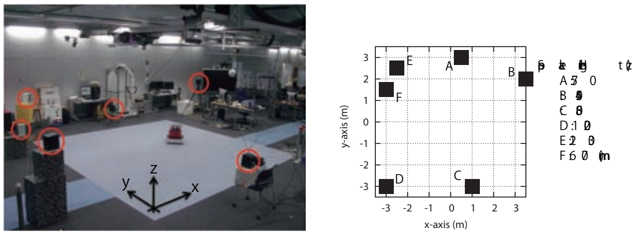
Computational loading for the proposed method is performed. The computer has a 2.1GHz Intel Core2 CPU and 2GB main memory. For evaluation, frequency spectrum is used as input data, because sound localization and separation are usually calculated in frequency domain. The calculation time for 1024 points (64msec) data is shown in Table II. It lists the response time of ”PCMs sound recognition” module in Fig. 5. Cluster number K is set to 15 and number of models are 2, 10 and 230. The calculation time depends on model size and input data size. In this condition, it spend less than half time of input data length for 230 models.

V. EXPERIMENT ON MOBILE ROBOT

This section describes some experiments for separated sound sources from a mobile robot embedded microphone array.

A. Experimental Setup

6 loudspeakers playing sounds included in PCMs codebook are used as the sound sources. Fig. 8 shows the experimental setup a) and loudspeakers arrangement b). The reverberation time (RT_{20}) is 170(msec) and background noise level is about 50(dBA) containing computer fan or exhaust fan in the room.



a) Exp. setup

b) Speakers arrangement

Fig. 8. Experimental setup for multiple sound recognition from robot in motion

The robot is manually controlled using a joystick. The average velocity is 0.4 (m/s) (1.1(m/s) maximum) and maximum angular velocity is 1.3 (rad/s).

Table III shows the list of sound sources included in the PCMs codebook. Total 26 sound sources are used for mobile robot experiment in this section. Daily sounds are recorded in ”Holone” using R-09, and office sounds are recorded in our laboratory which has a reverberation time(RT_{20}) of 150(msec) and background noise level of 35(dBA).

TABLE III

PCMs CODEBOOK FOR MOBILE ROBOT EXPERIMENT

Data set	Num.	Data example
Daily sounds	12	running water, curtain, refrigerator
Office sounds	8	voice, jingle keys, open/close door
Chirp sounds	6	birds chirping, frogs, cicadas
Total	26	

B. PCM Recognition for Separated Sound Sources

First, PCM recognition in static condition is performed. The noise reduction module described in Fig. 5 is not used in this experiment. The robot is set at $(x, y) = (0.0, 0.0)$(m), and separates each sound source using FBS from the MLF localization result.

Fig. 9 shows the average sound recognition rate for separated sound sources with the number of sounds from 1 to 6 (speaker A to F in Fig. 8 b)). The data is recorded for 60(sec) in each experiment. The recognition rate for 1ch input of the microphone array is dynamically decreased with increasing number of sound sources.

For DSBF enhanced signals, recognition rate of less than 2 sound sources are more than 70(%), and is dropped to about 60(%) for more than 3 sound sources. For FBS separated signals, recognition rate is around 80 (%), and is constant with number of sound sources.

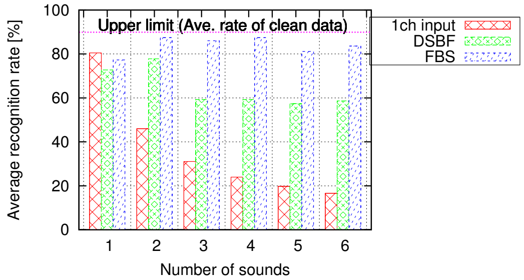


Fig. 9. Sound recognition rate for separated sound sources

C. PCM based Robot Noise Reduction

PCM recognition while robot motion is performed. Fig. 10 shows the recognition result when noise reduction module is not used. By comparison with the static condition in Fig. 9, the recognition rate is decreased overall. The result of 1ch input is around 10(%), and the DSBF result shows about 5(%) improvement from 1ch input. For FBS separated sound, the recognition rate is increased with the number of sound sources. The lower recognition rate for smaller sound sources is caused by robot noise which remains in separated sound, especially in silent intervals.

Fig. 11 shows PCM recognition results for croaking of frogs(model number is 13) at each time-slot in the 6 sound source experiment. The vertical axis shows the model number and model number = 0 means that it is recognized as a sound not included in the model. The recognition result is shown as a blue dot, and the correct model is drawn with a red line. For DSBF enhanced signal, there are error recognition to model number 1 to 3 (which means male voice, bounce ball and fan noise). FBS separated signal performed better than DSBF signal, and robot noise reduction decreases error recognition of model number 6 (which means

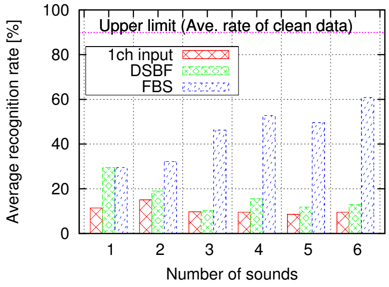
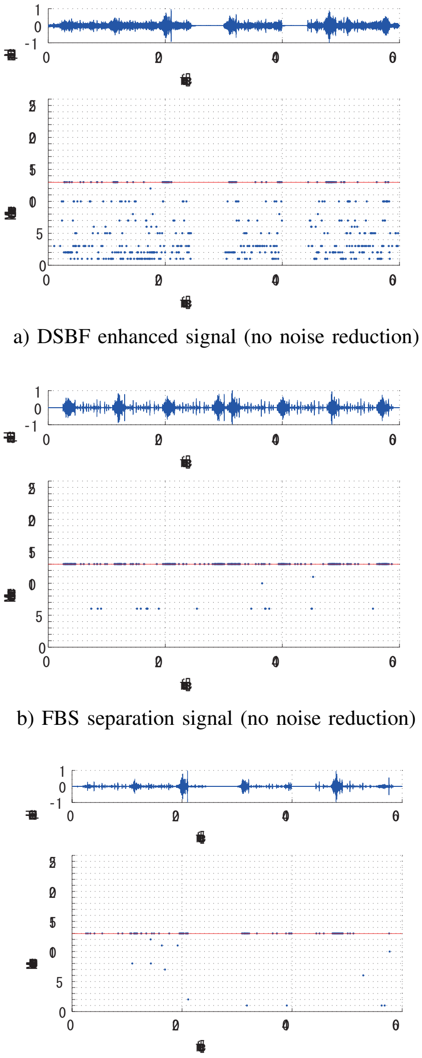


Fig. 10. Sound recognition rate for separated sounds while robot moving



a) DSBF enhanced signal (no noise reduction)

b) FBS separation signal (no noise reduction)

c) FBS separation signal after noise reduction

Fig. 11. Sound recognition result for croaking of frogs at each time slot (separated sound from 6 sound sources)

footsteps).

The average recognition rate using noise reduction is shown in Fig. 12. The PCM model for robot noise is generated from 60(sec) recordings while robot moving. The noise reduction is performed in each microphone's input before sound source localization.

The result shows higher recognition rate in comparison with Fig. 10, especially for separated sound sources of less than 2 sound sources. In this condition, the mobile robot simultaneously recognized 1 to 6 sound sources with 50 to 60(%) recognition rate in average from each 64(msec) input signal.

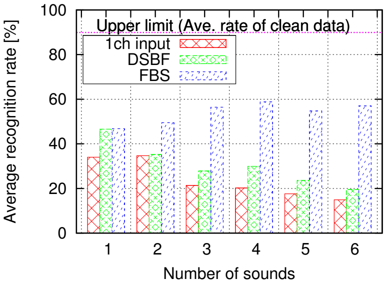


Fig. 12. Sound recognition rate after robot noise reduction

D. Cumulative Voting Recognition

In sections V-B, V-C, PCMs recognition for each time-slot input was evaluated. The instant result is useful for mobile robot application such as tracking moving sound sources. For recognizing sound, accumulation of instant recognition results increases recognition performance.

Fig. 13 shows the relationship between accumulating data time and the recognition rate for each number of sound sources. The instant recognition result in Fig. 12 is used for the evaluation. From cumulo-voting result of 1(sec) data, the average recognition rate is more than 80(%).

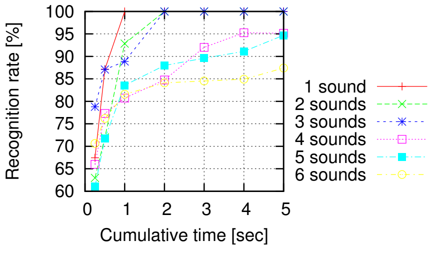


Fig. 13. Cumulative voting result for Fig. 9

VI. CONCLUSIONS AND FUTURE WORKS

This paper proposed a simple sound source identification method using a Pitch-Cluster-Map (PCMs) database. The method provides identification of sound sources and a solution of the issue: ”What is the sound?”, in computational auditory scene analysis. It works with short term signal input, and is suitable for mobile robot application.

The experimental result of simultaneous multiple sound recognition while robot moving shows that the proposed recognition system performed 50 to 60(%) recognition rate for less than 6 sound sources from each time-slot input. These instant recognition results are useful for a mobile robot application such as tracking moving sources. For sound stream recognition, accumulation of each instant recognition increases recognition performance. From 1(sec) accumulation, the system shows more than 80(%) recognition rate for 6 sound sources.

Robot noise reduction using PCMs is applied to input signal of the microphone array. Using preliminary recorded robot noise in motion, the system rejects the noise without robot motion information at each time.

Future work is needed to improve the method. First, some optimization method to decide the cluster number K is needed. Second, the K-means method is used for clustering in this paper, but other clustering algorithm should be considered.

REFERENCES

[1] S. Furui. 50 years of progress in speech and speaker recognition. In *Proceedings of SPECOM2005*, pp. 1–9, Patras, Greece, 2005.

[2] T. Matsui and K. Tanabe. Comparative study of speaker identification methods : dplrm, svm and gmm”. *IEICE Transaction on INFOMATION and SYSTEMS*, Vol. 89-D, No. 3, pp. 1066–1073, March 2006.

[3] N.Roman and D.L.Wang. Pitch-based monaural segregation of reverberant speech. *Journal of Acoustics Sciety of America*, Vol. 120, No. 1, pp. 458–469, July 2006.

[4] Yang Shao and DeLiang Wang. Model-based sequential organization in cochannel speech. *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 14, No. 1, pp. 289–298, January 2006.

[5] Masataka Goto. Analysis of musical audio signals. In D. L. Wang and G. J. Brown, editors, *Computational Auditory Scene Analysis:Principles, Algorithms, and Applications*, pp. 251–295. Wiley-IEEE Press, 2006.

[6] Hiromasa Fujihara, Tetsuro Kitahara, Masataka Goto, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno. Singer identification based on accompaniment sound reduction and reliable frame selection. In *Proceedings of 6th International Conference on Music Information Retreival (ISMIR2005)*, pp. 329–336, London, U.K., September 2005.

[7] Hiromasa Fujihara, Tetsuro Kitahara, Masataka Goto, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno. Speaker identification under noisy environments by using harmonic structure extraction and reliable frame weighting. In *Proceedings of International Conference on Spoken Language Processing (Interspeech2006)*, pp. 1459–1462, Pitsburgh PA, USA, September 2006.

[8] Jianfeng Chen, Alvin Harvey Kam, Jianmin Zhang, Ning Liu, and Louis Shue. Bathroom activity monitoring based on sound. *Pervasive Computing: Lecture notes in Computer Science*, Vol. 3468, pp. 47–61, May 2005.

[9] Paul Lukowicz, Jamie A Ward, Holger Junker, Mathias Stager, Gerhard Troster, Amin Atrash, and Thad Starner. Recognizing workshop activity using body worn microphones and accelerometers. *Pervasive Computing: Lecture notes in Computer Science*, Vol. 3001, pp. 18–32, May 2004.

[10] Satoru Tokutsu, Kei Okada, and Masayuki Inaba. Discrimination of daily sounds for humanoids understanding situations(in jalpanese). In *Proceedings of the 25th annual conference of the Robotics Society of Japan*, p. 1H36, September 2007.

[11] Yoko Sasaki, Satoshi Kagami, and Hiroshi Mizoguchi. Simple sound source detection using main-lobe model of microphone array (in japanese). In *Proceedings of the 25th annual conference of the Robotics Society of Japan*, p. 1N13, Chiba, Japan, September 2007.

[12] Yuki Tamai, Yoko Sasaki, Satoshi Kagami, and Hiroshi Mizoguchi. Three ring microphone array for 3d sound localization and separation for mobile robot audition. In *Proceedings of 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS2005)*, pp. 903–908, Edmonton, Canada, August 2005.