https://doi.org/10.20965/jaciii.2007.p1177

Paper:

A Robotic Auditory System that Interacts with Musical Sounds and Human Voices

Hideyuki Sawada and Toshiya Takechi

Department of Intelligent Mechanical Systems Engineering, Faculty of Engineering, Kagawa University
2217-20 Hayashi-cho, Takamatsu-city, Kagawa 761-0396, Japan
E-mail: sawada@eng.kagawa-u.ac.jp

[Received November 13, 2006; accepted August 16, 2007]

Voice and sounds are the primary media employed for human communication. Humans are able to exchange information smoothly using voice under different situations, such as a noisy environment and in the presence of multiple speakers. We are surrounded by various sounds, and yet are able to detect the location of a sound source in 3D space, extract a particular sound from a mixture of sounds, and recognize the source of a specific sound. Also, music is composed of various sounds generated by musical instruments, and directly affects our emotions and feelings. This paper introduces real-time detection and identification of a particular sound among plural sound sources using a microphone array based on the location of a speaker and the tonal characteristics. The technique will also be applied to an adaptive auditory system of a robotic arm, which interacts with humans.

Keywords: musical instruments, microphone array, sound identification, sound localization, mel cepstrum

1. Introduction

Voice and sounds are the primary media employed in human communications. Voice is used not only for simple daily communication, but also for logical discussions. Humans are able to exchange information smoothly using voice under different situations, such as in a noisy environment in a crowd and in the midst of plural speakers. Humans are surrounded by various sounds every day, and yet are able to detect the position of a sound source in 3D space, extract a particular sound from a mixture of sounds, and identify what generates a specific sound. Music is composed of various sounds generated by musical instruments, and directly affects our emotions and feelings. We are able to recognize specific musical instruments used in a musical performance, and are able to imagine a scene or a particular image when listening to a tune.

By realizing a human-like auditory system using a computer, a robot that interacts with humans by employing voice, sound and music will be presented. The techniques developed will be applied for recording a sound with high quality by reducing noise, presenting a clarified and enhanced sound, and realizing microphone-free speech recognition by extracting particular sounds.

Various techniques for detecting and identifying a particular sound from a sound signal have been proposed thus far. These are classified into two approaches; one is sound source separation from a monophonic input [1–3], and the other uses sound source information based on the stereo inputs [4–6]. Since most of the techniques require a sound source model or assume certain special conditions and restrictions, the computational costs become large, which causes difficulties for real-time processing.

This paper presents a robotic auditory system that reacts to and interacts with acoustic sounds and musical sounds. A technique for real-time detection and identification of particular sounds among plural musical sounds using a microphone array, based on the location of a sound source and its tonal characteristics will be described. The algorithm was then installed in a robotic arm, which tracks a particular sound source in the midst of plural sounds.

2. System Configuration

Figure 1 shows the system configuration, which consists of a microphone array, a low-pass filter (LPF), a computer with an A/D card, and a robotic arm. The microphone array is composed of 5 microphones L-Q arranged diagonally, as shown in **Fig. 2**. The microphone array was connected to the computer via the A/D card for the input sound signals. The sampling frequency was set to 16.0 kHz, and an analysis window of 1024 points was chosen in this study. The cut-off frequency of the LPF, which was placed in the inlet of the A/D card, was set to 8.0 kHz.

The system was able to identify the direction of an arbitrary sound source among plural sources by parallel input sound signals from the microphone array, and selectively enhancing a particular sound signal. At the same time, a particular sound was identified based on the direction of the sound source and the tonal characteristics.

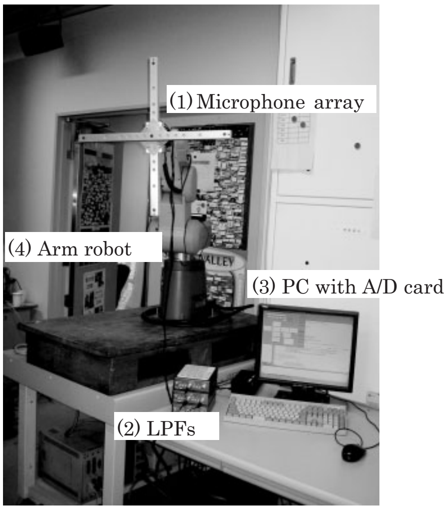


Fig. 1. System configuration.

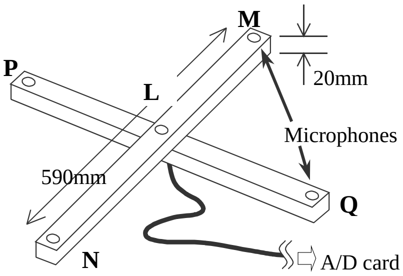


Fig. 2. Microphone array.

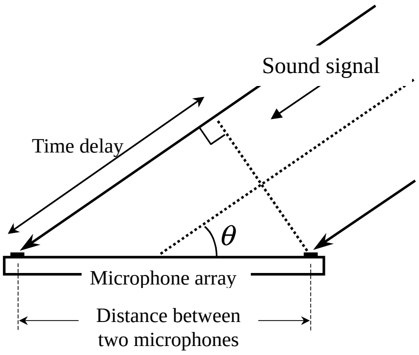


Fig. 3. Time difference reaching two microphones.

3. Estimation of Sound Source Direction

We assume that the source sound is a plane wave, and travels straight. The direction of the sound can be estimated by measuring the time delay between two microphones, as shown in **Fig. 3**. The direction can be estimated by the calculation of CSP (cross-power spectrum) phase analysis coefficients as

$$CSP_{1,2}(k) = DFT^{-1}\left[\frac{DFT\,[x_1(n)]\,DFT\,[x_2(n)]^*}{|DFT\,[x_1(n)]|\,|DFT\,[x_2(n)]|}\right]$$

$$CSP_{1,2,M} = \sum_{n=1}^{M} CSP_{1,2,n}(k) \quad . \; . \; . \; . \; . \; . \; . \; . \quad (1)$$

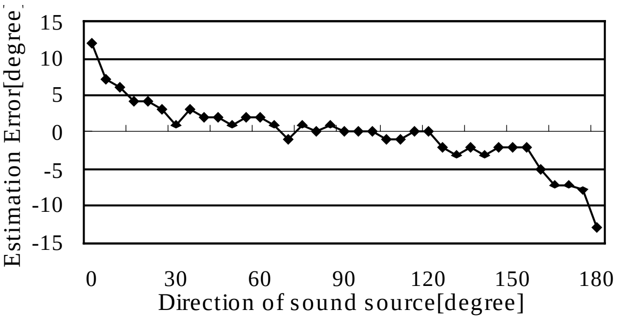


Fig. 4. Experimental result of direction estimation.

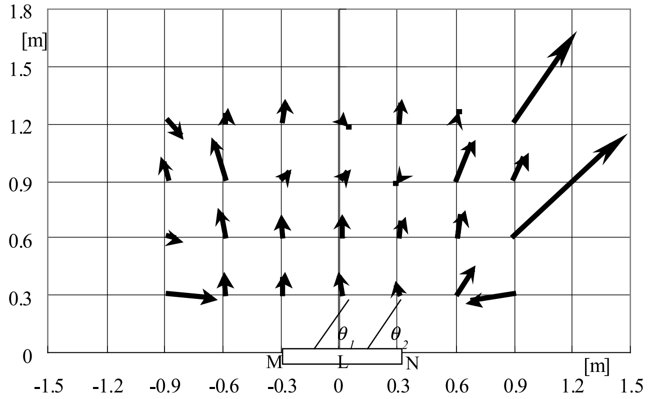


Fig. 5. Experimental results of estimation of location.

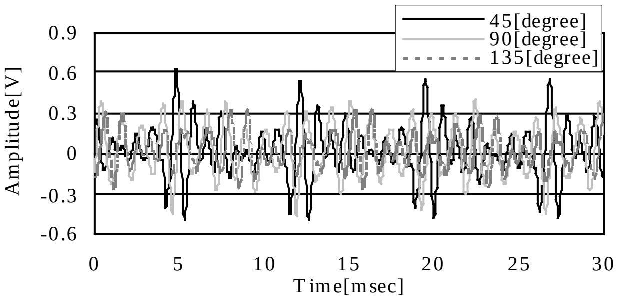
where $x_1(n)$ and $x_2(n)$ are the sampled signals of microphones Nos. 1 and 2, respectively. *DFT* represents the calculation of the discrete Fourier transform, the * represents a complex conjugate, and k corresponds to the time difference between the two signals [7, 8].

An experiment was conducted by placing source sounds in front of the microphone array at intervals of 5° on a circumference with a 150 cm radius, and the direction was estimated in open 3D space. The experimental results are shown in **Fig. 4**. The results presented that the fair estimation was assumed in the direction between 30° and 150° in front of the microphone array.

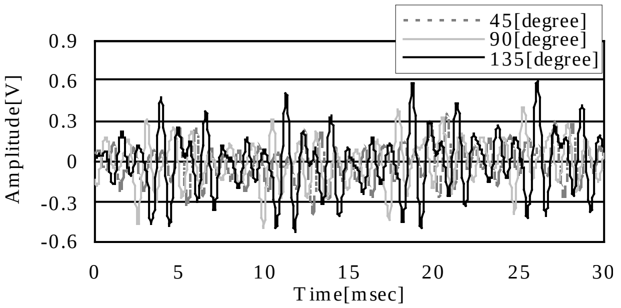
4. Estimation of Sound Location in 3D Space

The direction of a sound source can be estimated by using a pair of microphones placed apart. By using two arbitrary pairs from five microphones, the location of a sound source in 3D space can be calculated based on triangulation measurements. We conducted an experiment to estimate the position of the source. Sound sources were placed in front of the microphone array at intervals of 300 mm. A human voice was output from an acoustic speaker, and its position was estimated using the algorithm.

The experimental results of estimation of the location are shown in **Fig. 5**. Estimation errors are shown as vectors, where the origin of a vector shows the actual position of a source sound, and the end point indicates the estimated source position. The shorter the arrow length, the



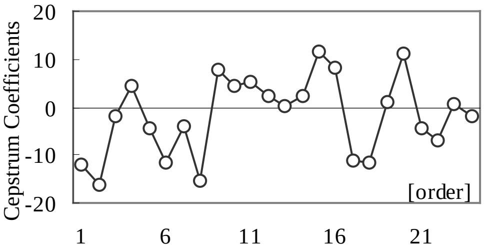
(a) Sound source placed in a 45° direction



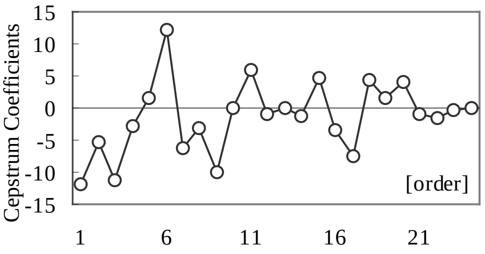
(b) Sound source placed in a 135° direction

Fig. 6. Experimental results of direction estimation.

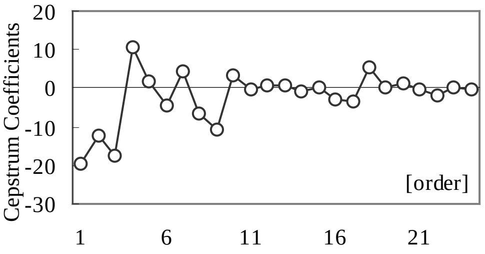
more accurate was the estimation. The results proved satisfactory estimation of the frontal area of the microphone array.



(a) Electric guitar (Clean tone)



(b) Recorder



(c) Human Voice

Fig. 7. Examples of MFCC.

5. Enhancement of a Particular Sound

A sound source located in the direction θ is selectively magnified and enhanced by the calculation of the delay-and-sum beam forming [3] as

$$y(t) = \sum_{i=1}^{M} x_i(t) \exp\left\{ j2\pi f(i-1)\frac{d\cos\vartheta}{c} \right\} \quad . \quad . \quad (2)$$

where M is the total number of microphones, d is the distance between two microphones, c is the speed of sound, and i is the microphone number.

A speech enhancement experiment was carried out by placing a sound source at 45° and 135° directions. The results of the enhancement are shown in **Fig. 6**, wherein sounds traveling from 45°, 90° and 135° direction were enhanced. **Fig. 6(a)** and **(b)** show successful enhancement of sounds from the 45° and 135° directions, respectively.

6. Identification of a Particular Sound

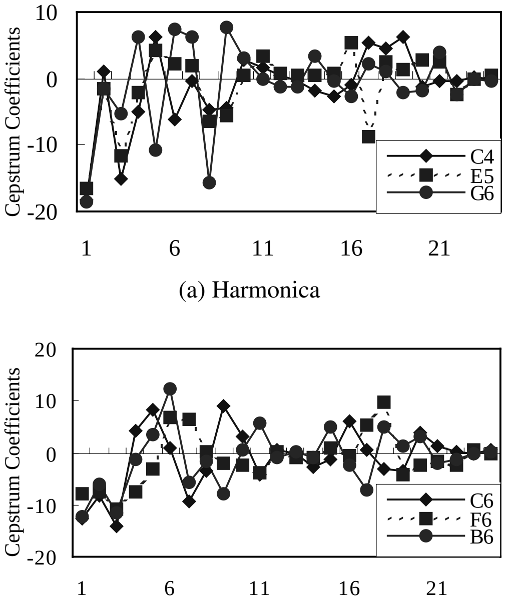
Humans are able to identify and isolate a particular sound from a mixture of sounds, including musical sounds, by using information such as the location of the source and tonal characteristics [9–11]. Computerized sound segregation can be realized by reproducing this procedure. As presented in the section above, the active selection of a particular sound source in 3D space could be achieved using a microphone array.

In this study, we also examined musical instruments and their sounds. Attention was paid to the tonal characteristics and their attenuation in the time-domain for segregation of a sound from multiple sounds, and we examined the template matching method based on templates obtained from the Mel-Frequency cepstrum coefficients (MFCC) with tonal characteristics [11, 12]. We also selectively employed human voices and five musical instruments, including an electric guitar (clean tone sound), an electric guitar (distorted sound), a classical guitar, a harmonica and a recorder.

Figure 7 shows examples of the 24th order MFCC extracted from the electric guitar (clean tone), the recorder and the human voice. The Mel cepstrum coefficients are often used for voice recognition, since different coefficient values are obtained for different speakers and different vowels. We attempted to extract differences from the musical instruments, and found this technique to be effective for feature extraction.

6.1. Characteristic Parameters

For identification of musical instruments, MFCC should be considered with the effects of its change in the



(a) Harmonica

(b) Recorder

Fig. 8. Comparison of MFCC of different pitches.

pitch and sound attenuation in the time-domain. The comparison of MFCC based on the change in pitch is shown in **Fig. 8**, and the difference in MFCC attenuation in the time-domain is presented in **Fig. 9**.

From the figures, we found that the MFCC changes with not only a change in the pitch but also with sound attenuation. In this study, the MFCC from the 1st to the 24th orders were employed to present the tonal characteristics for the identification of musical instruments. Characteristic parameters derived from the MFCC, as listed below, are stored as templates, and are used for feature extraction.

T1) MFCC from 1st to 24th order

T2) Binarized MFCC from 1st to 24th order

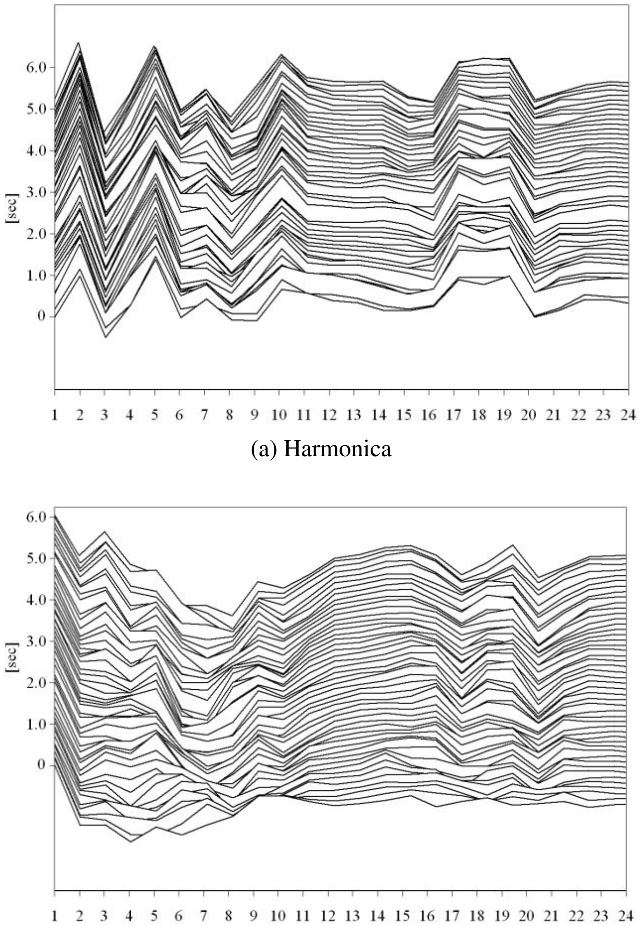
T3) Derivatives of MFCC among adjacent orders from 1st to 23rd order

T4) Binarized derivatives of MFCC among adjacent orders from 1st to 23rd order

Each template was averaged by shifting the window frame 30 times.

6.2. Template Matching Technique

For the template matching method using the characteristic parameters, we first prepared templates of the musical instrument sounds. As described in the section above, the MFCC changes due to a difference in the pitch and attenuation, even if the sounds are obtained from the same musical instrument.



(a) Harmonica

(b) Electric guitar (Distortion)

Fig. 9. MFCC attenuating in the time-domain.

Table 1. Templates of musical instruments and human voice.

Electric guitar (Clean tone)	/1st string/, /2nd string/, /3rd string/, /4th string/, /5th string/, /6th string/
Electric guitar (Distortion)	/1st string/, /2nd string/, /3rd string/, /4th string/, /5th string/, /6th string/
Classic guitar	/1st string/, /2nd string/, /3rd string/, /4th string/, /5th string/, /6th string/
Harmonica	/C4/, /G4/, /E5/, /C6/, /G6/
Recorder	/C6/, /E6/, /F6/, /A6/, /B6/
Human voice	/a/, /i/, /u/, /e/, /o/ (Japanese vowels)

Here, we propose to prepare multiple templates for each instrument with arbitrarily selected pitches. **Table 1** shows the templates selected for each musical instrument. For the string instruments, sound templates from each string were registered, and for the human voice, five Japanese vowels were selected to be stored as templates.

In the template matching, the characteristic parameters are obtained from the MFCC, and the distance to each template e_g is calculated as

$$e_g = \sum_{i=1}^{N} \left(\left| \delta_i - Temp_i^S \right| \right) \quad . . . . \quad (3)$$

Table 2. Experimental results for sound identification (pattern distance values).

		Templates					
		Clean tone	Distortion	Classic	Harmonica	Recorder	Voice
Inputted Sounds	Clean tone	***<u>37.1</u>***	98.8	74.7	133.9	129.1	113.8
	Distortion	96.6	***<u>37.8</u>***	98.6	112.6	118.9	113.5
	Classic	74.2	93.0	***<u>41.3</u>***	128.4	123.3	113.0
	Harmonica	112.8	105.4	122.3	***<u>43.6</u>***	133.5	140.0
	Recorder	126.9	118.5	126.7	140.8	***<u>41.2</u>***	140.2
	Voice	105.2	109.9	104.4	133.5	131.3	***<u>46.4</u>***
Error rate (%)		0.42% (1/240)	0.45% (1/224)	6.2% (16/257)	0.0% (0/164)	0.0% (0/167)	0.0% (0/172)

where δ_i represent the obtained characteristic parameters, Temp_i^s is the s-th template, and i represents the MFCC order.

The minimum e_g then is selected as a candidate. In case, if this is smaller than a predetermined threshold value T_h shown below, then the result of the recognition is confirmed.

$$T_h = \min_{i \neq j} \left\{ \sum_{\alpha} \frac{(V_\alpha^{g_i} - E_\alpha^{g_j})^2}{(\mu_\alpha^{g_j})^2} \right\} \quad . \; . \; . \; . \; . \quad (4)$$

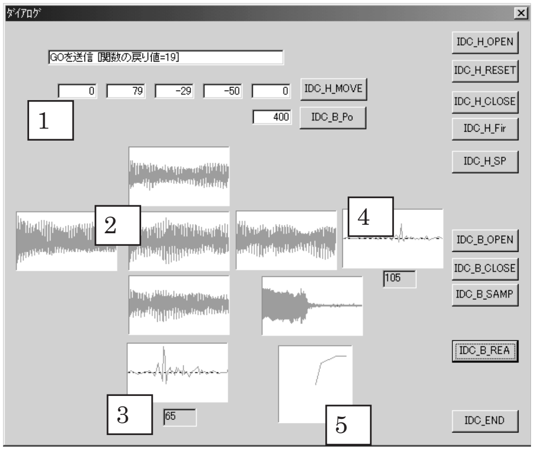
for all i, j
g_i, g_j: Instruments i, j
E: Average of Parameter Value
μ: Standard Deviation of Parameter Value

6.3. Recognition Experiments

An experiment for the identification of a particular musical sound was carried out. For the identification of a sound from an unknown source, the MFCC from the 1st to the 24th orders are extracted and the characteristic parameters are calculated. The distances against each template are then calculated, and the sound with the minimum error is selected as a recognition result.

The identification of sounds from five musical instruments and one human speaker situated in front of the microphone array was examined. All of the sounds were recorded in advance, and the templates were extracted and stored prior to the experiment. The experimental results are shown in **Table 2**.

Correct recognition was achieved for the harmonica, recorder and human voice, and fair results were obtained for the guitars. Appropriate differences were found between the templates of input sound and the other ones of different instruments, resulting in good performance of the identification. In the identification of the three guitar sounds, mis-recognition among the three was observed, likely due to the small pattern distances. Guitar sounds start attenuating just after the string is picked, and the MFCC changes with the attenuation to cause the error.



GOを送信 [関数の戻り値=19]

1. 5 Joint Angles of Robot
2. Input Sound Signals
3. Horizontal Direction Estimation
4. Vertical Direction Estimation
5. Posture of Robot Arm

Fig. 10. Interface panel for sound tracking.

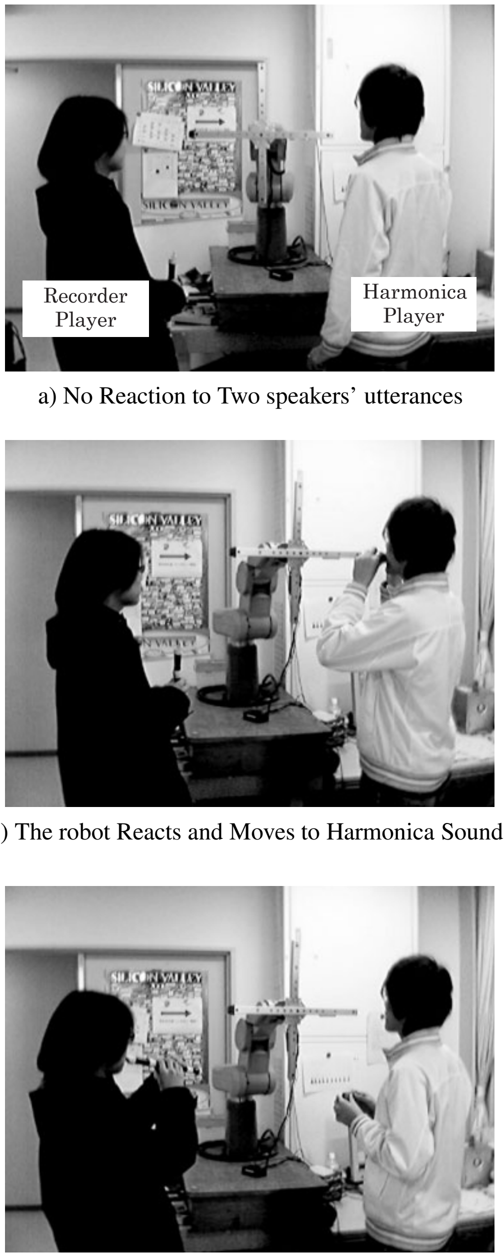
7. Sound Tracking by a Robotic Arm

The sound identification algorithm was installed in a robotic arm to track a particular person or musical instrument among plural sound sources, as already shown in **Fig. 1**. The sound location is estimated based on the two diagonal directions M-L-N and P-L-Q. The microphone array inputs five signals from L to Q simultaneously. An interface panel is shown in **Fig. 10**, in which the sound signals input and the estimated results are displayed in real-time, together with the posture of the robotic arm.

Figure 11 shows a human voice tracking experiment. The robotic arm system identified the two speakers' voices, and tracked only the left speaker, while it did not react to the right speaker. The robot also listens to musical instruments, and selectively tracks a particular musical instrument.

8. Conclusions

Real-time detection and identification of a particular sound among musical sounds generated by musical in-



a) No Reaction to Two speakers' utterances

b) The robot Reacts and Moves to Harmonica Sounds

c) No Reaction to Recorder Sounds

Fig. 11. Tracking of Harmonica Sounds.

struments was presented. Using a microphone array and using location estimation of a sound source as tonal characteristics, the sounds were identified in real-time. First the location of sounds among plural sound sources was estimated, and the sound was enhanced according to the location information. The sound was then identified to track only the targeted musical instrument. Satisfactory results of the identification were thus obtained by the proposed algorithm. The algorithm was installed in a robotic arm together with the microphone array to achieve real-time tracking of a particular sound. The algorithm was installed in a robotic arm together with the microphone array to achieve real-time tracking of a particular sound. Based on the tracking of a particular sound source, the robot is able to listen to the sound and to behave like a human and act as if it is appreciatively listening to music.

Future work is planned for improving the identification algorithm to be applied to realize recognition of unknown sounds which are not included in the templates in advance. We are now working to construct a humanoid robot that interacts with a human using voice and musical sounds which represents new multimodal communication.

References:

[1] M. Unoki and M. Akagi, "A Method of Signal Extraction from Noise-Added Signal," IEICE, Vol.J80-A, No.3, pp. 444-453, 1997.

[2] S. Hayakawa, K. Takeda, and F. Itakura, "Speaker Recognition Using the Harmonic Structure of Linear Prediction Residual Spectrum," IEICE, Vol.J80-A, No.9, pp. 1360-1367, 1997.

[3] A. Nehorai and B. Porat, "Adaptive Comb Filtering for Harmonic Signal Enhancement," IEEE Trans. Acoust., Speech & Signal Processing, Vol.34, No.5, pp. 1124-1138, 1986.

[4] T. Yamada, S. Nakamura, and K. Shikano, "Hands-free Speech Recognition with Talker Localization by a Microphone Array," Information Processing Society of Japan, Vol.39, No.5, pp. 1275-1284, 1998.

[5] F. Asano, S. Hayamizu, and T. Matsui, "A Realtime Noise Reduction System using Delay-and-Sum Beamformer and its Application to Speech Recognition," Electrotechnical Laboratory, 1996.

[6] J. L. Flanagan, A. C. Surendran, and E. E. Jan, "Spatially selective sound capture for speech and audio processing," Speech Communication, Vol.13, pp. 207-222, 1993.

[7] T. Nishiura, T. Yamada, S. Nakamura, and K. Shikano, "Localization of Multiple Sound Sources Based on CSP Analysis with a Microphone Array," IEICE, D-II, Vol.J83-D-II, No.8, 2000.

[8] C. H. Knapp and G. G. Carter, "The generalized correlation method for estimation of time delay," IEEE Trans. Acoust., Speech & Signal Processing, Vol.24, No.4, pp. 320-327, 1976.

[9] T. Funada and T. Tsuzuki, "Feature extraction based on spectral slope for speech recognition," IEICE, D-II, Vol.J82-D-II, No.11, pp. 2184-2187, 1999.

[10] T. Takechi, K. Sugimoto, T. Mandono, and H. Sawada, "Automobile identification based on the measurement of car sounds," Annual Conf. of the IEEE Industrial Electronics Society, TD6-4, 2004.

[11] H. Sawada and M. Ohkado, "Identification and tracking of particular speaker in noisy environment," Int. Conf. on Machine Vision and its Optomechatronic Applications, OpticsEast, SPIE Int. Society for Optical Engineering, pp. 138-145, 2004.

[12] S. Imai, "Cepstral Analysis Synthesis on the Mel Frequency Scale," IEEE Int. Conf. Acoust., Speech & Signal Processing, pp. 93-96, 1983.

Name:
Hideyuki Sawada

Affiliation:
Department of Intelligent Mechanical Systems Engineering, Faculty of Engineering, Kagawa University

Address:
2217-20 Hayashi-cho, Takamatsu-city, Kagawa 761-0396, Japan

Brief Biographical History:
1990 Received B.Eng. degree from Waseda University, Japan
1992 Received M.Eng. degree from Waseda University, Japan
1995-1998 Research Fellow of the Japan Society for the Promotion of Science
1999 Received Ph.D. from Waseda University, Japan
1998-1999 Research Associate of Waseda University
1999- Associate Professor of Kagawa University

Main Works:
- Y. Mizukami and H. Sawada, "Tactile Information Transmission by Apparent Movement Phenomenon Using Shape-memory Alloy Device," Int. Conf. on Disability, Virtual Reality and Associated Technologies, pp. 133-140, 2006.
- S. Hashimoto and H. Sawada, "A Grasping Device to Sense Hand Gesture for Expressive Sound Generation," Journal of New Music Research, Vol.34, No.1, pp. 115-123, 2005.
- D. Coquin, E. Benoit, H. Sawada, and B. Ionescu, "Fusion of Hand and Arm Gestures," SPIE Optomechatronic Machine Vision, Vol.6051, 6051-14, 2005.
- [Hyper Human Tech Award] H. Sawada and M. Nakamura, "Mechanical Voice System and its Singing Performance," IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS2004), pp. 1920-1925, 2004.
- T. Higashimoto and H. Sawada, "Vocalization Control of a Mechanical Vocal System under the Auditory Feedback," Journal of Robotics and Mechatronics, Vol.14, No.5, pp. 453-461, 2002.

Membership in Academic Societies:
- Institute of Electrical and Electronics Engineers (IEEE)
- The Japan Society of Mechanical Engineers (JSME)
- The Robotics Society of Japan (RSJ)
- The Information Processing Society of Japan (IPSJ)
- The Institute of Electronics, Information and Communication Engineers (IEICE)
- The Japanese Society of Instrumentation and Control Engineers (SICE)

Name:
Toshiya Takechi

Affiliation:
Department of Intelligent Mechanical Systems Engineering, Faculty of Engineering, Kagawa University

Address:
2217-20 Hayashi-cho, Takamatsu-city, Kagawa 761-0396, Japan

Brief Biographical History:
2004 Received B.Eng. degree from Kagawa University, Japan
2006 Received M.Eng. degree from Kagawa University, Japan

Membership in Academic Societies:
- Institute of Electrical and Electronics Engineers (IEEE)