2012 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 7-12, 2012. Vilamoura, Algarve, Portugal

Sound Sources Selection System by Using Onomatopoeic Querries from Multiple Sound Sources

Yusuke Yamamura, Toru Takahashi, Tetsuya Ogata and Hiroshi G. Okuno

***Abstract*— Our motivation is to develop a robot that treats auditory information in real environment because auditory information is useful for animated communications or understanding our surroundings. Interactions by using sound information need an aquisition of it and a proper sound source reference between a user and a robot leads to it. Such sound source reference is difficult due to multiple sound sources generating in real environemnt, and we use onomatopoeic representations as a representation for the reference. This paper shows a system that selects a sound source specified by a user from multiple sound sources. Users use onomatopoeias in the specification, and our system separates a mixed sound and converts separated sounds into onomatopoeias for the selection. Onomatopoeais have the ambiguity that each user gives each expression to a certain sound and we create an original similarity based on Minimum Edit Distance and acoustic features for solving its problem. In experiments, our system receives a mixed sound consisting of three sounds and a user's query as inputs, and checks a count of a consistency of a sound source selected by a system and a sound source specified by a user in 100 tests. The result shows our system selects user's required sound source at 49.2%.**

I. INTRODUCTION

Robots need to treat not only human voices and musical sounds but also environmental sounds with respect to understanding our surroundings. Such information helps us determine our next actions to do. For example, a loud sound of an engine tells us the coming of a car and we can listen and pay attention to it. Opening a door is natural if hearing a sound of knocking on the door. In particular, in terms of a detection of an extraordinary sound or understanding our surrounding in a bad view, the importance of auditory information is superior to the one of visual information. In recent years research and development of a robot (system) is advancing, which is manipulated from a remote location by a human (user) such as a telepresence robot. It is helpful that a system provide the required information to a user from actual environment when a user requires information of a specific sound.

A system's provision of the auditory information required by a user is based on the consistency between a sound source specified by a user and a sound source guessed by a system. In this paper, we realize that a user and a system do a unique reference to a sound source in auditory scene. When a user asks a system " Where is this sound comming? ", a user can understand his circumstances if a system tells him a dirrection of " this sound ". This behavior goes well thanks to a proper reference of a sound source.

Y. Yusuke, T. Takahashi, T. Ogata, H. G. Okuno are with Graduate School of Informatics, Kyoto University, Kyoto, 606-8501, Japan. {yamamura, tall, ogata, okuno} @kuis.kyoto-u.ac.jp

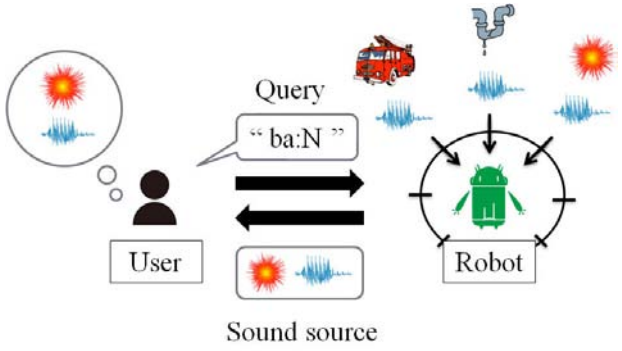

Fig. 1. Reference to a sound by using onomatopoeias

For unique sound source reference, onomatopoeic representation is suitable. Onomatopoeia imitates a sound with a phonemic system of a mother tongue, and enables us to express our impression for a sound. In real environment there are some sound sources timely and spatialy. This representation can refer to individual sounds respectively. Even if a sound is unknown (you don't know what a sound comes from), we can label it with onomatopoeias.

Our system must take the ambiguity of onomatopoeias into account. An onomatopoeic representation depends on users and can be dfifferent from each user. When one user gives an onomatopoeia to certain environmental sound, another user dosen't necessarily give a same representation. This shows that the consistency of two onomatopoeias between a user and a system is difficult.

We deal with this ambiguity problem by making the similarity between two onomatopoeias. The similarity of two onomatopoeias means that a user and a system refer to a same sound. For example, we suppose the situation that a user asks a system " Where is this sound "ba:N" comming? "(N means a syllabic nasal). In this situation, even if the system gives "da:N" to the sound, the similarity of two onomatopoeias "ba:N" and "da:N" leads to a conclusion that the system can refer to a sound specified by the user as shown in the Fig.1.

In this paper, we develop the system that selects one sound source specified by a user's onomatopoeic query from multiple sound sources. A user selects a sound source from multiple sound sources and gives it an onomatopoeia. The system firstly separates multiple sounds into single sounds, and converts each sound to an onomatopoeia. Selecting the most similar onomatopoeia to the user's onomatopoeia, the system judges that the sound having the onomatopoeia is a

user's one.

II. A SELECTION OF A SOUND SOURCE

A. A reference for the selection

Sound source selection problem is the problem that a system selects one sound source refered to by a user from multiple sound sources. Sounds disappear as time passes and we can't refer to their substances. A reference method of sound sources is needed to solve its problem. For example, imagine you select one sound source from three ones. It is difficult for a reference if sound sources don't have any symbols. It is not until we give three sound source symbols of A, B and C respectively that a user and a system can refer to each sound source.

Actual environment has many sound sources at a same time and a proper representation for a sound source reference is required in such case. There are representations such as generating timings[1], [2], arrival directions[3] and class names[4], [5], [6], [7], [8]. The representation by using generating timings enables us to refer to unknown sounds but can't treat the environments where multiple sounds exist at a same time. Arrival directions of sound sources are able to deal with the above environment and on the other hand they are short of the accuracy if a user is at a remote location. We have an advantage of intelligible references and a disadvantage of a limitation of the number of class names in using class names of sound sources. That's why it is hard to say that these representations are suitable for a sound source reference in real environment.

B. A reference with onomatopoeic representations

An onomatopoeia is a representation when listeners imitate environmental sounds with the phoneme system of their mother tongue. This representation makes more delicate expressions than other representations, and helps us refer to sound sources. We can express unknown sounds by using onomatopoeias even if we don't know when, where and from what they come, and give known sounds the more concreteness. The reference of onomatopoeias aren't affected by spatial and temporal ristrictions.

This paper shows the development of a reference method using onomatopoeias. The reference with the use of onomatopoeias means that so many onomatopoeic representations, so many classes of sound sources and we are able to refer to every sound source substantially.

A problem onomatopoeic representations have is an ambiguity that onomatopoeias users give from sound sources depend on the individuals. Even if a system generating onomatopoeias automatically is developed on the basis of a particular generation rule, its system isn't for the public. Mapping onomatopoeias of users and our system is needed and we creat a similarity of them to deal with this problem.

Using onomatopoeias is equal to a visualization of environmental sounds and it leads to a visual understanding of auditory scenes. Shneiderman adovocates that a visualization of information is made of three steps, "Overview first, Zoom and Filter, the Details on demand"[9]. The selection of a sound source by using onomatopoeias is regarded as a new interface in the manipulation "Zoom and Filter". This interface can handle environmental sound sources including unknown sounds in addition to the existing system that visualizes voices[10].

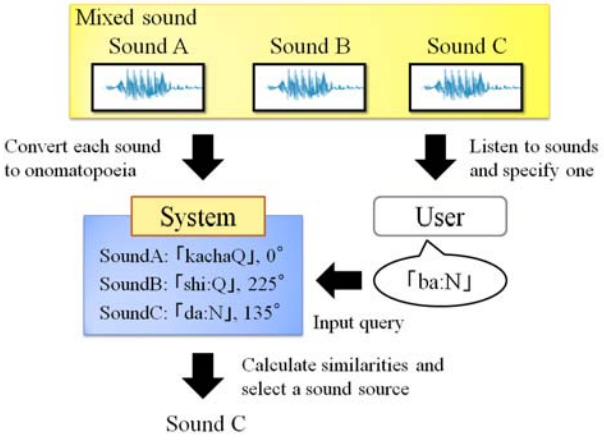

Fig. 2. Image of our sound selection system with an onomatopoeic query

III. SOUND SOURCE SELECTING SYSTEM WITH AN ONOMATOPOEIC QUERY

A. Problem Statement

Our system receives a mixed sound and a user's onomatopoeic query as inputs and shows the user the sound source specified by him. The problem is described as follows:

inputs: A mixed sound
An onomatopoeic query (phoneme sequence)
output: A required sound source

and a movement of our system is shown in Fig. 2. We assume that a user and this system listen to same sounds and the user's onomatopoeia is not his voice but Japanese phoneme sequence. Our system can get some sound sources' information, such as each sound itself, arrival directions and onomatopoeic representations of sound sources, and provide us with auditory information a required sound source has.

B. Assumption

We have the following three assumptions.

1) Each sound is a single directional environmental sound
2) Our system's onomatopoeias have the next grammatical contsruction, /C/+/V/+/QN/
3) User and system have the same phoneme set

Firstly we treat single directional environmental sounds in this paper. Directional environmental sounds frequently act as signals transmitting our environmental changes and give us beneficent information. In fact, we also listen non-directional sounds but we don't treat them because of the difficult separation of them.

Secondly onomatopoeias our system outputs are ristricted by a next form: C (consonants) + V (vowels / long vowels) + QN (a choked sound / a syllabic nasal). This limitation

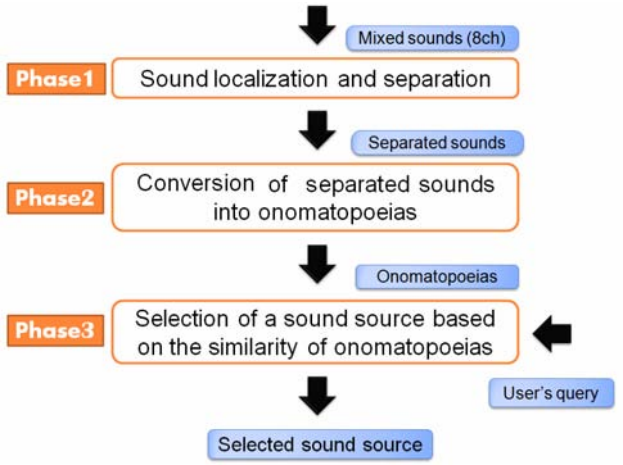

Fig. 3. Processings in our system

of the grammar means that an automatic generalization of general onomatopoeias we often pronounce is difficult. Users' queries aren't affected by this restriction since our system is required to accept various onomatopoeias.

Thirdly the phoneme set of the user's query has to coincide with the system's one. If they have different phoneme sets, our system can't calucualte the similarity between phonemes that are unknown to each other. In this paper, users use the same models that is used the generation of the transcriptions from sound sources.

C. Construction

Figure 3. shows three processes in our system.

phase1sound localization and separation
phase2conversion of separated sounds to onomatopoeias
phase3selection of a required sound source

In the first phase our system separates a mixed sound into some sounds with the use of a sound location and separation. The next phase means that each separated sound is converted into each onomatopoeic phoneme sequence. Our system selects the most similar sound source in the last phase after calculating the similarities between user's query and onomatopoeias and comparing them.

In this paper we focus on the phase3 because of the solution of phase1 and phase2 with already developed methods[11], [12]. Two processes up to the phase3 are detailed below.

D. Sound localization and separation

Our system separates a mixed sound into separated sounds in the first phase. A sound location and separation of a mixed sound is done by using an open source software HARK (HRI-JP Audition for Robots with Kyoto University)[11]. HARK provides us with fair modules for making a robot audition system and assures the highly efficient sound localization and separation, the real time processing, the robustness to noises. A sound localization is performed with MUSIC (Multiple Signal Classification) and a sound separation is done with GHDSS (Geometric High-order Discorrelation-based Source Separation). These methods are packaged as HARK's modules.

TABLE I
DESCRIBED PHONEME SET

/t/, /k-t/, /b/, /p/, /t-ch/, /sh/, /k/, /f-p/, /t-p/, /z-j/, /g/, /r/, /k-p/, /ch/, /k-t-ch/, /b-d/, /j/, /t-ts/, /w/, /ts-ch/, /s-sh/, /k-t-r/, /d-g/, /b-d-g/, /sh-j/, /k-g/, /t-d/, /a-o/, /a/, /i/, /u/, /e/, /o/, /a:-o:/, /a:/, /i:/, /u:/, /e:/, /o:/, /N/, /Q/, /Q-N/

E. Convertion of separated sounds to onomatopoeias

Each separated sound is converted into an onomatopoeia in this phase. The convertion is performed by using Ishihara's method[12]. Ishihara developed a system that automatically converted an environmental sound into an onomatopoeia, which is based on the analysis of the relevance between single environmental sounds and onomatopoeias[13]. Allocattions of phonemes to acoustic signals use the frame of the speech recognition. Using the frame is no problem for environmental sounds according to Cowling's report[14]. Thus Mel-Frequency Cepstral Coefficients (MFCC) and Hidden Markov Model (HMM) are used as acoustic features and recognizers respectively in Ishihara's system.

Considering the different onomatopoeias listeners give, Ishihara made a set including phonemes for onomatopoeias. This set consists of phonemes frequently appearing in descriptions of listeners. For example, you may express "koN" or "toN" when listening a sound of tapping on a table, and a new phoneme "/k-t/" is created in this case. New phonemes for environmental sounds you can express with various onomatopoeias and japanese basic phonemes are all elements of this set. This unique set is shown in TABLE I.

IV. SELECTION OF A SOUND SOURCE WITH THE SIMILARITY OF ONOMATOPOEIAS

A. Problem Statement

In phase3 our sytem selects the most similar sound source in multiple sound sources on the basis of our similarity of onomatopoeias. This is because onomatopoeias are different from individuals and our system. Fig. 4 shows that an onomatopoeia, a systems' output in phase2, is hard to completely corresponds with users' one. The problem statement is as follows:

inputs: some onomatopoeias
an onomatopoeic query (phoneme sequence)
output:a required sound source

The number of onomatopoeias of separated sounds is the number of single sound sources. Our system measure the similarity between the user's query and each onomatopoeias and judges that the most similar onomatopoeia is the user's onomatopoeia.

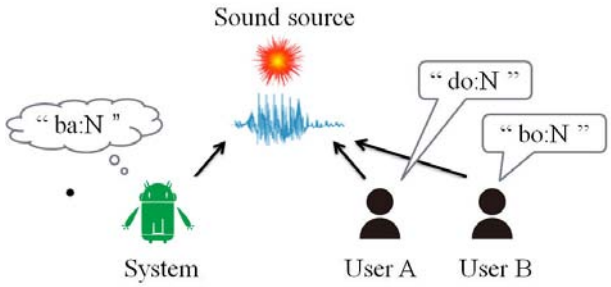

Fig. 4. difficulty of the complete consistency between onomatopoeias

B. Minimum Edit Distance

The similarity of onomatopoeias is created by using the Minimum Edit Distance (MED). The MED is a numerical value indicating what degree of difference two character strings have and a generalized Hamming distance. The MED has three operations to a string; insertion, deletion, and substitution of a character, and the value is defined as a minimum cost of operations in converting one string to the other.

A calculation of the MED of two string is based on dynamic programming. There are two strings A and B whose length are a and b. The MED between A and B is computed as follows:

$$\begin{aligned} M(0,\ 0) &= 0 \\ M(i,\ 0) &= i * I(1 \leq i \leq a) \\ M(0,\ j) &= j * D(1 \leq j \leq b) \\ M(i,\ j) &= min\{M(i-1,\ j-1) + S(A(i),\ B(j)), \\ &\quad M(i-1,\ j) + D,\ M(i,\ j-1) + I\} \end{aligned}$$

Where $A(i)$ is the ith index of the string A and $B(j)$ is the same. $I,\ D$ and $S(A(i),\ B(j))$ are costs of insertion, deletion and substitution respectively, and $S(A(i), B(j))$ is 0 if $A(i)$ is equal to $B(j)$.

In this paper, we define a basic MED's cost as $I = D = 1, S = 2$. A distance of two strings “kaQ” and “kotoQ”, for example, is 4 in this case because of 2 insertion and 1 substitutions.

C. Creation Of Substitution Costs

A basic MED should be extended for the concereteness of the similarity. Considering the next situation, the necessity is clear. There are two sound sources, which are expressed ”pa:N” and ”ja:N” respectively by our system and user's query ”ba:N”. Selecting the former is proper in this case unless a user has a special sensitivity. This is because consonants /b/ and /p/ belong to plosive sounds and /j/ belongs to fricative sounds. A basic MED calculates same values to two similarities, and this result is improper. The reflection of the difference sounds have is required.

In our method, we give different costs to each S of combination of phonemes, given $I = D = 1$. S depends on Kullback-Leibler Divergence (KLD) of two phonemes' probability distributions, which are in the acoustic model used at the phase 2. The reason of using KLD is for its generality and easiness to calculate. KLD has the concrete similarity of probability distributions and a capability of judging what concrete degree of similarity they have. This concreteness enables our system to select the required sound source from multiple separated sound sources.

We calculate the KLD of each phoneme. Phoneme p has its probability distribution $p(x)$, whish is defined as a 16 gaussian mixture model and represented as follows:

$$p(x) = \sum_{k=1}^{16} \pi_k N(x|\mu_k, \Sigma_k) \tag{1}$$

where π_k, μ_k and Σ_k is the kth weight, mean and variance-covariance respectively. μ_k is a 34 dimensional vector and Σ_k is a 34*34 dimensional matrix. The KLD between phoneme p and q is defined as below:

$$KL(P||Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} d\mathbf{x} \tag{2}$$

The analytical solution of this equation is difficult due to the multi-dimensional gaussian distribution. In this paper, the calculation of the equation (2) is regarded as carrying out an expected value of $\log \frac{p(x)}{q(x)}$ by the use of Monte Carlo calculations. Given the probability density function $p(x)$ of a random variable x, an expected value of any function f is

$$E_p[f(x)] = \int f(x)p(x)dx \tag{3}$$

Assigning $\log \frac{p(x)}{q(x)}$ to the function f of this equation shows other expression,

$$KL(P||Q) = E_p[\log \frac{p(x)}{q(x)}] \tag{4}$$

Then we take N samples of $p(x)$ and name them $x_1, x_2, \cdots, x_N$. The equation (2) is approximated with an expected value of discrete random variable;

$$KL(P||Q) \approx \frac{1}{N} \sum_{i=1}^{N} \log \frac{p(x_i)}{q(x_i)} \tag{5}$$

The more the number of the sampling is, the closer the value of the equation (5) is to truth value.

The sampling of the equation (1) has two steps.

1) selecting a gaussian distribution for a sampling with weights π
2) random sampling from the choiced gaussian distribution

In the first step, the weights are probabilities of the selection. $\sum_{k=1}^{16} \pi_k = 1$ holds and we select a gaussian distribution at a rate of a value of its weight. In the second step, we get a sample by using a random sampling. A sample is a random number generating from the multi-dimensional gaussian distribution $N(x|\mu,\ \Sigma)$ choiced in the preliminary stage. We transform uniform random numbers into gaussian random numbers with Box-Muller method and get multivariate gaussian random numbers generated by the use of Cholesky

decomposition. Firstly we get N independent gaussian random numbers $\mathrm{z} = [z_1, \cdots, z_N]^{\mathrm{T}}$, $z_1, \cdots, z_N \sim N(0, 1)$. Cholesky decomposition decomposes A into a product of a lower triangular matix L and its transverse matrix L^{T}, where non-negative Hermitian matrix A is a real symmetric matrix. The decomposition of the variance-covariance Σ gives us to a lower triangular matrix L.

$$\Sigma = \mathrm{L}\mathrm{L}^{\mathrm{T}} \tag{6}$$

A sample of $N(x|\mu, \Sigma)$ is expressed as follows:

$$\mathrm{x} = \mu + \mathrm{L}\mathrm{z} \tag{7}$$

We take 1,000 samples in the above way and caluculate KLDs.

We determine the substitution cost $S(p, q)$ of phoneme p and q. All phonemes are classified as C (consonants), V (vowels / long vowels) and QN (a choked sound / a syllabic nasal). If two phonemes p, q are belonging to different classes, the substitution cost is $I + D$. Otherwise, probability distributions of two phonemes are P, Q respectively and its substitution cost is defined as follows:

$$S(p, \ \ q) = KL_{SYM}(P, Q)\frac{I + D}{KL_{MAX}} \tag{8}$$

where $KL_{SYM}(P, Q)$ is an average of KLDs of P and Q;

$$KL_{SYM}(P, Q) = \frac{KL(P||Q) + KL(Q||P)}{2} \tag{9}$$

The mutual average fills the symmetry of distance. KL_{MAX} is the maximum value of KL_{SYM}s in the same class, and it normalizes $S(p, q)$ such that $0 \leq S(p, \ q) \leq I + D$.

V. EXPERIMENTS

A. Settings

In this experiment, we look up the selection performance of our system. Supposing a real environment, we asked 5 male university students to take part in the experiment. They selected one of three sound sources as users and their onomatopoeic representations were queries to our system.

We use 4,287 non-voice and non-continuous sound files from RWCP Sound Scene Database[15] and all files are 16bits/16kHz samplings. These files include tappings of a cap, sounds of a whistle, sounds of shutting a book, sounds of digging in the sands, and so on. These sounds can be categorized into about 40 classes and each class has similar 100 sound files which are different from each other in terms of its duration, its pitch, and its timbre. All files are labeled onomatopoeic labels by one person in advance. They are splitted into traing data and test data at the ratio of 9:1 randomly. An acoustic model learns by training data and test data are used for the recognition.

B. Procedure

This experiment has the following two step;

1) specify a sound source from three sound sources with an onomatopoeia by a user
2) select a required sound source by our system

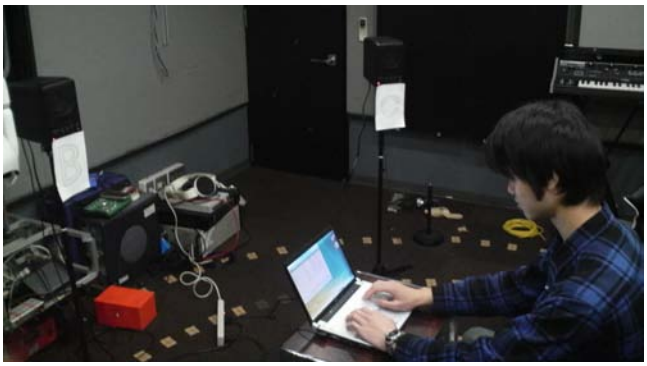

Fig. 5. Experiments by subjects

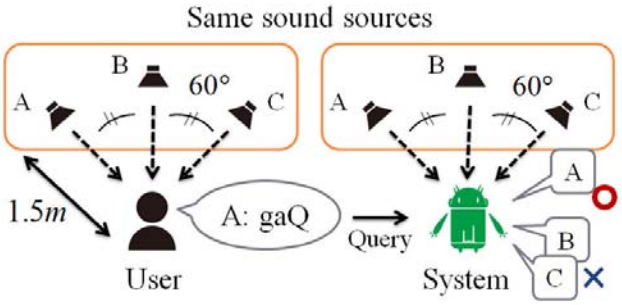

Fig. 6. Illustration of experiments

Firstly 5 subjects gives sound sources onomatopoeias to specify. There are three speakers at 60 degrees and a subject is remote from them. Three speakers generate sound sources at a same time that are chosen randomly from the test data. A subject listens to them and specify one sound source with its onomatopoeia. This image is shown the left side of Fig. 6. This specification is one trial and 100 trial is executed in this experiment.

Secondly our system recieves a mixed sound, which consists of three sound sources in the same condition of the following step, and user's query to one of them as shown the right side of Fig. 6 Three sound sources are convolved with $(0°, \ 60°, \ 300°)$ impulse responses respectively and added in making a mixed sound. From user's query our system selects one sound source.

A valuation basis is the ratio of the consistency of two sound sources refered to by a user and our system. In Fig. 6, for example, if a subject specifies a sound source A, selecting A contributes to the high accuracy rate.

C. Result

Table II shows each ratio of the precision. The concrete creation of substitution costs results in the correct selection of sound source at 49.2%, but this value isn't sufficient as long as you consider the chance rate 33.3% There are the folowing reasons:

- the dependency of the similarity on the train data
- the error of convertion due to the distortion of the sound separation

TABLE II

ACCURACY RATE IN EACH CONDITIONS

subjects	user1	user2	user3	user4	user5	Ave
Accuracy Rate	45%	45%	48%	54 %	54%	49.2%

- the ratio of the weights of the transforming costs

The biggest factor is the dependency of our similarity on the train data. KLDs are easy to reflect the tendency of user's onomatopoeic labeling to the train data because of their derivation from the probability distributions of phonemes. Our system can't select a required sound source if other users, which are different from the user that gives onomatopoeic labels, input their queries. The train data's labels are given by only one user, and in contrast the sounds' labels of speakers are given by other 5 users. The labels they attached are various and our method couldn't deal with its variety well.

Other factor is having a lot of errors of convertion due to a contamination of leaking noise and should be solved. In fact, we saw many examples that a sound of whistle "pi:Q" invades other separated sounds and phonemes "p" appear in onomatopoeias of these sounds.

We also have an alternative approach for an improvement of the accuracy rate and it is a creation of insertion, deletion costs of each phonemes. This is because in listenig to an environmental sound and giving an onomatopoeia to it, we don't insert or delete each phoneme uniformly. For example, A sound "shaQ", a sound of cutting a paper by a scissor, may be expressed "kashaQ" by some people but is rarely expressed "pashaQ". A phoneme "k" is easier to insert than a phoneme "p" and the insertion cost of "k" is set low in this case.

D. Future Works

At first, we have to collect many sound data many users give onomatopoeic labels to. As discribed above, a general and precise system needs an acoustic model, which is made of these train data.

In real environment there are not only multiple sounds at same but also bachground noises that persistant at a long time. A mixed sound of a single sound and a background noise is difficult to be separated successfully. Yamakawa[16] proposes the extraction method of acoustic features that has robustness to noises by using Matching pursuit (MP) and Formant-wave function (FoF). Considering the good movement of our system in the environment having multi noises, we need to think about this method.

We suppose the situation in this paper that there are multiple environmental sounds at a same time. The real environment has single sounds and multiple sounds generated in a certain time frame, and a user usually has the chance of selecting one sound source. If single sounds occur multiple times, it becomes a sound selection problem. For a real time processing, our system needs to deal with a continuity of time.

VI. CONCLUSION

This paper presents a sound source selection system from multiple sound sources with users' onomatopoeic query. For a behavior in real environment there are three processings in our system; a sound separation of a mixed sound, a convertion to onomatopoeias and selection of a required sound source. Considering the ambiguity of users' onomatopoeic representations, our system creates the original MED based on acoustic features of environmental sounds. Our system selects user's required sound source at the rate of 49.2% in the experiment. The improvement of the performance will be reported in the near future.

REFERENCES

[1] Komatani, K. and et al: Spoken Dialogue System that Users Information on Locutionary Acts to Interpret User Utterances(In Japanese), *IPSJ Journal*, Vol. 52, pp. 3374–3385 (2011).

[2] Matsuyama, K. and et al: Analyzing User Utterances in Barge-inable Spoken Dialogue System for Improving Identification Accuracy, *Eleventh Annual Conference of the International Speech Communication Association* (2010).

[3] Okuno, H. and Nakadai, K.: Machine Audition Technology that Listens to Multiple Voiced Speech at Once, *Journal of The Institute of Electrical Engineers of Japan*, Vol. 131, No. 3, pp. 159–163 (2011).

[4] Jahns, G., Kowalczyk, W. and Walter, K.: Sound analysis to recognize individuals and animal conditions, pp. 1–8 (1998).

[5] Zhang, T. and Kuo, C.: Audio-guided audiovisual data segmentation, indexing, and retrieval, *Proceedings of SPIE*, 3656, p. 316 (1998).

[6] Ntalampiras, S., Potamitis, I. and Fakotakis, N.: A Multidomain Approach for Automatic Home Environmental Sound Classification, *Eleventh Annual Conference of the International Speech Communication Association*, 10, pp. 2210–2213 (2010).

[7] Arona, R. and Lutfi, R.: An Efficient Code for Environmental Recognition, *The Journal of the Acoustic Society of America*, Vol. 126, No. 1, pp. 7–10 (2009).

[8] Chu, S., Narayanan, S. and Kuo, C.: Environmental sound recognition with time–frequency audio features, *Audio, Speech, and Language Processing, IEEE Transactions on*, Vol. 17, No. 6, pp. 1142–1158 (2009).

[9] B.Shneiderman: *Designing the User Interface (3rd Ed)*, Addison-Wesley (1998).

[10] Kubota, Y. and et al: 3d auditory scene visualizer with face tracking: Design and implementation for auditory awareness compensation, *Universal Communication, 2008. ISUC'08. Second International Symposium on*, IEEE, pp. 42–49 (2008).

[11] Nakadai, K. and et al: Design and Implementation of Robot Audition System'HARK'Open Source Software for Listening to Three Simultaneous Speakers, *Advanced Robotics, 24*, Vol. 5, No. 6, pp. 739–761 (2010).

[12] Ishihara, K. and et al: Disambiguation in determining phonemes of sound-imitation words for environmental sound recognition, *Eighth International Conference on Spoken Language Processing* (2004).

[13] Hiyane, K., Sawabe, N. and Iio, J.: Study of Spectrum Structure of Short-time Sounds and its Onomatopoeia Expression(In Japanese), *IECEJ thecnical report, Electroacoustics*, Vol. 97, No. 586, pp. 65–72 (1998).

[14] Cowling, M. and Sitte, R.: Comparison of techniques for environmental sound recognition, *Pattern Recognition Letters*, Vol. 24, No. 15, pp. 2895–2907 (2003).

[15] Nakamura, S. and at el: Sound scene data collection in real acoustical environments, *JOURNAL-ACOUSTICAL SOCIETY OF JAPAN-ENGLISH EDITION-*, Vol. 20, pp. 225–232 (1999).

[16] Yamakawa, N. and et al: Environmental sound recognition for robot audition using matching-pursuit, *Modern Approaches in Applied Intelligence*, pp. 1–10 (2011).

[17] Mizumoto, T. and et al: Design and implementation of selectable sound separation on the Texai telepresence system using HARK, *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, IEEE, pp. 2130–2137 (2011).