Applied Bionics and Biomechanics
Vol. 6, No. 3–4, September–December 2009, 355–367

Taylor & Francis
Taylor & Francis Group

Binaural active audition for humanoid robots to localise speech over entire azimuth range

Hyun-Don Kim*, Kazunori Komatani, Tetsuya Ogata and Hiroshi G. Okuno

Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University, Yoshida-honmachi, Kyoto, Japan

(*Received 29 September 2008; final version received 10 May 2009*)

We applied motion theory to robot audition to improve the inadequate performance. Motions are critical for overcoming the ambiguity and sparseness of information obtained by two microphones. To realise this, we first designed a sound source localisation system integrated with cross-power spectrum phase (CSP) analysis and an EM algorithm. The CSP of sound signals obtained with only two microphones was used to localise the sound source without having to measure impulse response data. The expectation-maximisation (EM) algorithm helped the system to cope with several moving sound sources and reduce localisation errors. We then proposed a way of constructing a database for moving sounds to evaluate binaural sound source localisation. We evaluated our sound localisation method using artificial moving sounds and confirmed that it could effectively localise moving sounds slower than 1.125 rad/s. Consequently, we solved the problem of distinguishing whether sounds were coming from the front or rear by rotating and/or tipping the robot's head that was equipped with only two microphones. Our system was applied to a humanoid robot called SIG2, and we confirmed its ability to localise sounds over the entire azimuth range as the success rates for sound localisation in the front and rear areas were 97.6% and 75.6% respectively.

Keywords: active audition; sound source localisation; humanoid robots; human-robot interaction

1. Introduction

Auditory functions for intelligent humanoid robots have conventionally used a lot of microphones and generally required prior information and learning parameters (Hara et al. 2004; Valin et al. 2007). These technical restrictions have caused problems with hardware capacity and calculation time, and designing auditory systems that were expensive was unavoidable. For this reason, we tried to use only two microphones like human ears and design a robot auditory system without prior information such as impulse response measurement or learning parameters as much as possible. These conditions should enable the robot auditory system to be simply and inexpensively implemented. Moreover, since a pair of microphones can work well on sound devices inside general personal computers, it makes the best use of ubiquitous stereo input. However, two microphones do not have enough performance to cope with noisy environments or sound processing over the entire azimuth range. For example, it is hard for the robot to distinguish whether sound sources are being generated from the front or rear due to its symmetrical head with a pair of microphones installed in its left and right ear positions.

Many robot engineers and researchers have investigated solutions to achieve superior auditory performance with minimum requirements. Although two microphones are currently insufficient for the sound processing of robot audition, we expect that applying the capabilities of hearing in binaural animals and humans to robots will make it possible to improve the performance of the robot auditory system by using two microphones. Thus, in one engineering approach, a pair of microphones has been proposed with motion modelled on the human perception of sound circumstances and it is used in auditory systems designed by using the concept of "active audition" (Nakadai et al. 2002; Berglund 2005; Kim 2008).

Motional theories (Blauert 1996) are of importance to spatial hearing only when the effects of the head and the external ears have already been taken into consideration. The most thorough investigation (Blauert 1996) was undertaken by Thurlow et al. (1967), who observed more than 50 subjects with normal hearing attempting to determine the position of a sound source in an anechoic chamber. The subjects were blindfolded. The sound source radiated narrow-band noise (500–700 Hz or 7.5–8 kHz). Ten sound source positions, well distributed around the chamber, were used. The subjects were permitted to move their heads freely while addressing the task but were instructed to keep their torsos motionless. Their head movements were recorded by a motion picture camera and were subsequently interpreted. A classification in terms of rotating, tipping and pivoting movements was used in the interpretation (Figure 1). The most important results of the interpretation are shown in

*Corresponding author. Email: hyundon@kuis.kyoto-u.ac.jp

ISSN: 1176-2322 print / 1754-2103 online

DOI: 10.1080/11762320903007430
http://www.informaworld.com

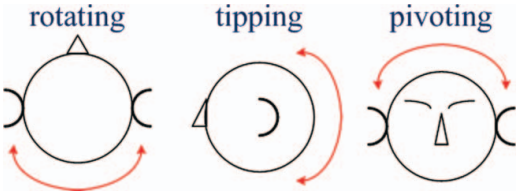

Figure 1. Classes of head movements.

Table 1. As can be seen, the largest average amplitude is that of rotating movements. The most frequent combination of classes of movements involves rotating and tipping movements.

According to these results, we should confirm the validity of experiments on sound localisation over the whole azimuth by movements of microphones. Accordingly, we need to consider how specific attributes of input signals observed from a pair of microphones can be made available to the auditory system for interpretation of localising sounds. We focused on designing such applications to robots in this paper.

1. We applied expectation-maximisation (EM) algorithm to the sound localisation system using two microphones which enables it to simultaneously estimate several sounds and reduce the sound localisation.
2. To evaluate the binaural sound localisation system with motions, we proposed the way to construct the database of moving sounds.
3. We solved the front–rear problem of the binaural sound localisation for humanoid robots by motioning two microphones.

First, to localise sound sources, conventional systems have used many microphones, i.e. an array of microphones, and some have needed to measure the impulse response data beforehand. Thus, these specifications require a relatively complex algorithm and a great deal of calculation time. Also, configuring a lot of microphones in the robot would be problematic and measuring impulse response data in an anechoic chamber is expensive. Therefore, we used cross-power spectrum phase (CSP) analysis (Nishiura et al. 2000; Kim et al. 2007b) to localise the sound source because it can calculate the Time Delay of Arrival (TDOA) between two microphones without impulse response data. In addition, to cope with the ambiguity and sparseness of acquired information picked up by two microphones, we applied EM algorithm (Moon 1996) to localise several sound sources and reduce localisation errors.

Second, since robots move and rotate their bodies and heads in order to track someone and/or perform tasks, the sound localisation method should be able to localise moving sounds while coping with the effects created by moving microphones. Therefore, to design the sound localisation system to cope robustly with moving sounds, we first need database for various moving sounds to evaluate it. As a conventional way to create database for moving sounds, we have recorded sound signals and their positions while moving the speaker manually. Thus it is difficult to create moving sounds which have accurate track information and to repeatedly create the same database with the same condition in order to compare a developed system with other ones for sound localisation regardless of methods types. To solve these problems, we developed the moving sound creation tool by using the API library called SoundLocus of Arinis Sound Technologies Co., Ltd. (http://www.arns.com/english/index.html).

Finally, robots need to improve their cognition abilities (active perception) concerning changing location of sounds while in motion. For example, robots should be able to distinguish whether sound signals are coming from the front or rear if they rotate or move only two microphones placed in the robot's head or body. To solve front–rear problem in binaural sound source localisation, we detect the change in sound localisation by slightly rotating and/or tipping the two microphones. Since such motions have the effect of increasing the number of microphones virtually, a binaural audition system can estimate sound localisation over the entire azimuth range.

This paper is organised as follows: Section 2 explains binaural sound source localisation combining CSP analysis and the EM algorithm. Section 3 describes a tool for freely creating moving sounds to enable sound localisation with motion. In Section 4, we designed active audition to

Table 1. Compilation of the results on the head movements of 23 subjects asked to determine the direction of sound sources (Thurlow et al. 1967).

	Rotating	Tipping	Pivoting	Rotating, tipping, pivoting	Rotating, tipping	Rotating, pivoting	Tipping, pivoting
Average amplitudes and standard deviations							
500–1000 Hz signals	42°± 20.4°	13.1°± 13.5°	10.2°± 9.6°				
7500–8000 Hz signals	29.2°± 18.6°	15.2°± 12.9°	11.6°± 8.3°				
Relative statistical frequencies with which subjects would exhibit specific combinations of movements							
500–1000 Hz signals (in %)	48	13	3	39	70	22	4
7500–8000 Hz signals (in %)	41	15	5	36	62	19	6

distinguish whether sounds were coming from the front or rear by rotating the robot's head. In Section 5, we analysed the effect of sound localisation with tipping androtating and improved the success rate of localising speech over the entire azimuth range. Finally, section 6 concludes the paper with a brief summary and a look at future work.

2. Design of binaural sound source localisation

Although sound source localisation for a humanoid robot has been developed, and its performance has been generally improved, conventional methods still have drawbacks.

1. Localising several sound sources needs many microphones and/or impulse response data.
2. Positioning microphones in the robot is restricted by the shape of its body and/or head.
3. Sound source localisation is usually unreliable in noisy environments.

The latest sound source localisation systems for robots mostly use one of the three methods: head-related transfer function (HRTF) (Cheng and Wakefield 2001; Hwang et al. 2005), multiple signal classification (MUSIC) (Schmidt 1986; Hara et al. 2004) or CSP analysis (Nishiura et al. 2000; Kim et al. 2007b). HRTF and MUSIC typically need impulse response data and an array of microphones in order to localise several sound sources. Impulse response data must be measured for every discrete azimuth and/or elevation before these methods can be applied to robots. Even though many microphones and much impulse response data would improve localisation performance, they would also increase the calculation time. Furthermore, configuring the microphones in the robot would be problematic.

In contrast, CSP analysis does not need impulse response data and can accurately determine the direction of a sound using only two microphones. A system using CSP with two microphones can locate only one sound source of each frame even if several sound sources are present. This is because CSP analysis obtains the sound localisation information from the spatial correlation between two signals. Moreover, CSP analysis is usually unreliable in noisy environments. To overcome these shortcomings, we developed a method based on probability for estimating the number and location of sound sources. We applied an EM algorithm (Moon 1996) to our method in order to estimate the distribution of the data and reduce the error in sound source localisation.

2.1. Cross-power spectrum phase (CSP) analysis

The direction of a sound source can be determined by estimating TDOA between two microphones (Huang et al. 1998; Kim et al. 2007a). When there is a single sound source, it can be estimated by finding the maximum value

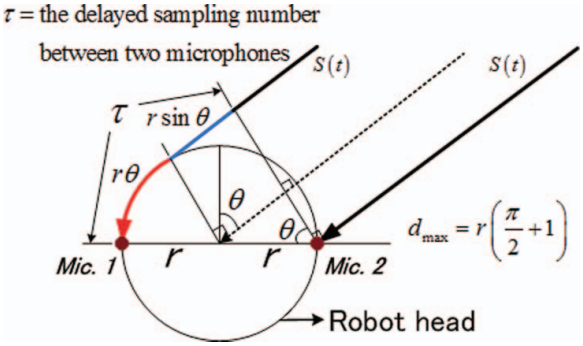

Figure 2. Time delay of arrival (TDOA) for CSP analysis.

of CSP coefficients as derived from the following equations.

$$csp_{ij}(k) = IFFT\left[\frac{FFT[s_i(n)]\,FFT[s_j(n)]^*}{|FFT[s_i(n)]||FFT[s_j(n)]|}\right], \quad (1)$$

$$\tau = \arg\ \max(csp_{ij}(k)), \quad (2)$$

where k and n are the sampling numbers of the delay of arrival between two microphones, $s_i(n)$ and $s_j(n)$ are the signals entering into the microphones i and j, respectively, *FFT* (or *IFFT*) is the fast fourier transform (or inverse *FFT*), * is the complex conjugate operator and τ is the estimated TDOA. The sound source direction is derived from

$$\theta = \cos^{-1}\left(\frac{v \cdot \tau}{d_{\max} \cdot F_s}\right), \quad (3)$$

where θ is the sound direction, v is the sound propagation speed, F_s is the sampling frequency and d_{max} is the distance between the two microphones at which the time delay is maximum. The sampling frequency of our system is 16 kHz. Figure 2 shows the parameters used in Equation (3) for a sound source arriving directly. We assume that the sound waves received at a pair of microphones become plane waves because talkers usually speak at 1 to 3 m from a robot. According to Blauert (1996), as the distance increases beyond 1 m, the spherical waves become more and more plane-like.

2.2. Applying EM algorithm to localising sounds

Figure 3A shows sound source localisation events extracted by CSP according to time or frame lapses. Events that lasted 192 ms are used to train the EM algorithm to estimate the number and localisation of sound sources. The interval for the EM algorithm was experimentally determined as shown in the left part of Figure 7. Figure 3B shows the training process for the EM algorithm to estimate the distribution of sound source localisation events. The EM training results in Figure 3C indicate refined localisations by iterating processes (A) and (B). The interval for EM training is shifted every 32 ms.

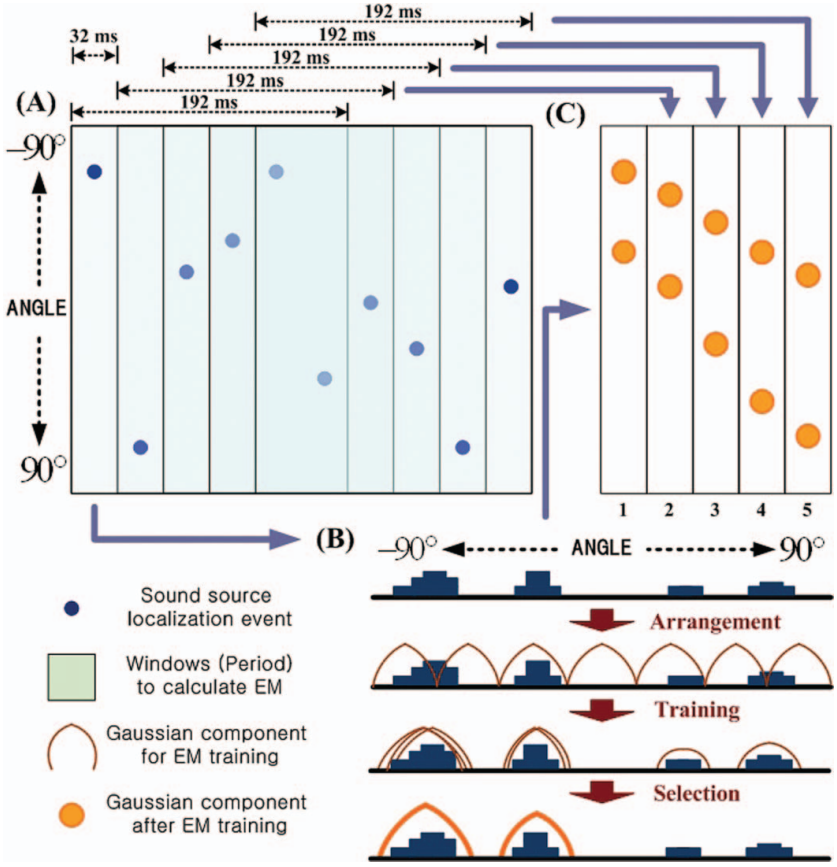

Figure 3. Process to localise sound sources by EM algorithm.

Figure 4 illustrates the process of applying the EM algorithm in more detail. In Figure 4A, the first step is to gather sound localisation events for 192 ms (six frames) for use in EM training. To void errors in the EM training due to lack of events, we executed this step only when each frame had an event, i.e. there were six events for 192 ms. Next, the Gaussian components, defined using Equation (4), for training the EM algorithm, are uniformly arranged on whole angles.

$$P\left(X_m\,|\theta_k\right) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(X_m-\mu_k)^2}{2\sigma_k^2}}, \tag{4}$$

where μ_k is the mean, σ_k^2 is the variance, θ_k is a parameter vector, m is the number of data and k is the number of mixture components. We used nine components, $k = 9$, with a radius, $\sigma_k = 10$. The distance between each μ_k was 20 ($-90 \leq \mu_k \leq 90$). In this step, the μ and σ parameters in the Gaussian components are the respective centre and radius values of each component. The sound localisation events are applied to the arranged Gaussian components to find the parameter vector, θ_k, describing each component density, $P(X_m|\theta_k)$, through iterations of E-step and M-step.

7. E-step, the expectation step, essentially computes the expected values of the indicators, $P(\theta_k|X_m)$, using Bayes' rule, derived as

$$P\left(\theta_k\,|X_m\right) = \frac{P\left(X_m\,|\theta_k\right)\cdot w_k}{\sum_{k=1}^{N} P\left(X_m\,|\theta_k\right)\cdot w_k}, \tag{5}$$

where X_m, θ_k and w_k are the azimuth values of the sound localisation event, the parameter set and the weight value for each component k, respectively, and N is the total number of mixture components.

8. M-step, the maximisation step, computes the cluster parameters that maximise the likelihood of the data, assuming that the current data distribution is correct. As a result, we can obtain the recomputed mean using Equation (6), the recomputed variance using Equation (7)

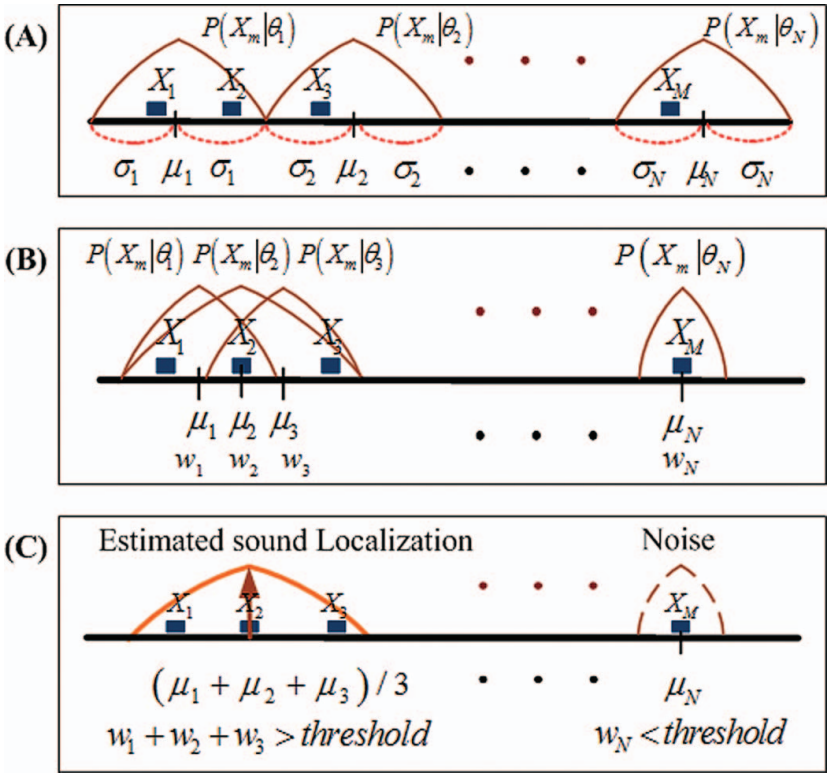

Figure 4. Process of EM algorithm to estimate number and location of sound sources.

and the recomputed mixture proportions (weights) using Equation (8). The total number of data is M.

$$\mu_k = \frac{\sum_{m=1}^{M} P(\theta_k|X_m)\, X_m}{\sum_{m=1}^{M} P(\theta_k|X_m)}, \tag{6}$$

$$\sigma_k^2 = \frac{\sum_{m=1}^{M} P(\theta_k|X_m)\cdot (X_m-\mu_k)^2}{\sum_{m=1}^{M} P(\theta_k|X_m)}, \tag{7}$$

$$w_k = \frac{1}{N}\sum_{m=1}^{M} P(\theta_k|X_m). \tag{8}$$

After the E and M steps are iterated for an adequate number of times (we iterated them 10 times), the estimated mean, variance and weight based on the current data distribution can be calculated. Then, as illustrated in Figure 4B, the weights and means of the Gaussian components are reallocated on the basis of the density and distribution of the sound localisation events. Finally, as illustrated in Figure 4C, if the Gaussian components overlap, the weights of the overlapping components are added. If the weight value is higher than a threshold value, the system can localise the sound source by computing the average mean of the overlapping Gaussian components. Components with small weights, regarded as noise, are moved.

2.3. *Evaluation of our sound localisation with EM algorithm*

To evaluate the performance of the EM algorithm applied to our sound localisation system, we experimentally compared our method, i.e. the CSP method with the EM algorithm, with the CSP only method. We recorded five commands, 'sig', 'ohayogozaimasu', 'konnichiwa', 'konbanwa' and 'oyasuminasai', which mean 'the name of our robot', 'good morning', 'good afternoon', 'good evening' and 'good night', respectively. They were produced at every 10° from −90° to 90°, at a distance of 1.5 m from the head of the robot and at a magnitude of 85 dB (A). Since the robot was at the centre of a square room with sides of about 5 m and background noise of about 55 dB (A), the reverberation effect was neglected. We calculated the average results for each measurement point. As shown in Figure 5, the average errors with the CSP method and the EM algorithm were less than those with the CSP only method for the entire azimuth. Since our method used six frames for EM training, we also evaluated each CSP method using six frames due to the fair comparison of those results. The 'average error' indicates the average of the differences

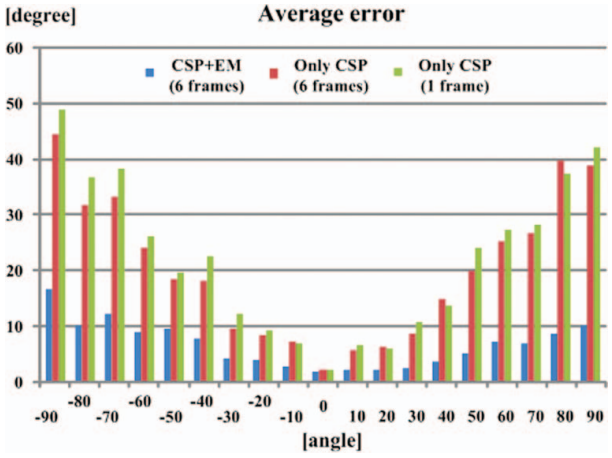

Figure 5. Front localisation errors for CSP only method and CSP + EM method.

between the original point and the observed localisation angles.

3. Evaluation using moving sound creation tool

Robots should localise moving sound sources and fixed sound sources because they should be able to move and rotate their bodies and heads to track people or perform their missions. Therefore, sound source localisation systems for robots should be able to localise moving sounds created by moving microphones as well as moving talkers. Thus, a database for various moving sounds is necessary to design and evaluate the sound localisation systems to robustly cope with moving sounds. After we evaluated our sound localisation system by using created moving sounds, we could design a robust sound localisation system for moving sounds.

3.1. Construction of database for moving sounds

We developed the moving sound creation tool by using the API library from Arnis's technology. We assumed that the validity of this tool was confirmed because theses and patents of Arnis's technology were already presented in its website. This tool can convert audio data of a wave file form into a stereo wav file according to the desired track as shown in Figure 6. Therefore, by designating the velocity and track of moving sounds in advance, we could freely make moving sounds of stereo wave file forms. Since this tool based on HRTF is to create moving sounds for a headphone set, this one does not consider reverberation and ambient noise. Nevertheless, this is effective to evaluate a proposed sound localisation method and compare it with other methods under the same condition, i.e. it is unnecessary to consider the error of a track for moving sounds and to reflect dynamically changed resonance and background noses in real environments whenever doing experiments.

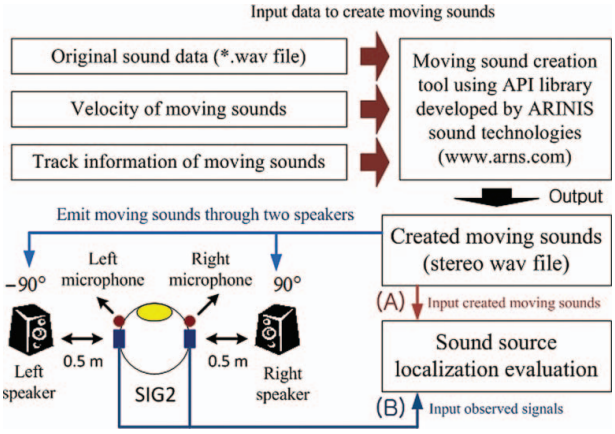

Figure 6. Creating moving sound sources and experimental conditions.

3.2. Evaluation of our method for moving sounds

To evaluate our method for single moving sounds, we created eight moving speech signals, which were rotated from 0° to 359° at 0.25, 0.375, 0.5, 0.625, 0.75, 0.875, 1, 1.25 rad/s at about 2.0 m from the centre position with the humanoid robot SIG2 (refer to Figure 9). The length of each created moving sound was 30 s. We performed sound localisation using these sounds, as shown in Figure 6A. We also localised sounds emitted by stereo speakers, as shown in Figure 6B. We used two omni-directional microphones installed at the left and right ear positions of the humanoid robot SIG2 and used two fixed speakers at 0.5 m from the left and right sides of the microphones. To evaluate our method for two moving sounds, we mixed two moving sounds. One rotated at 2 m from the centre at 0.25, 0.375, 0.5, 0.625, 0.75, 0.875, 1, 1.125 rad/s and the other rotated at 1 m at the half the angular velocity lagging 90° behind the other one. The middle part of Figure 7 shows the track of moving sounds and the results of localising the two moving sounds.

The left part of Figure 7 shows the average error and success rate of localising single moving sound according to the number of frames for training the EM algorithm. The success rate is the total percentage when the difference between the original location of the created moving sound source and the estimated sound localisation was within 30°. All dotted lines in Figure 7 indicate the results of localising sounds as observed from speakers shown in Figure 6B. Here, in six frames for EM, the average error was the least and the success rate was the best. Therefore, we could determine experimentally that the appropriate interval for our system was 192 ms (six frames) as shown in Figure 3. Moreover, we learned that our system can cope with moving sounds slower than 1.125 rad/s. Since one of the purposes of this study was to help robots to localise the voices of walking people, we confirmed that our system can cope with moving speech at the average walking speed, 1.0 m/s

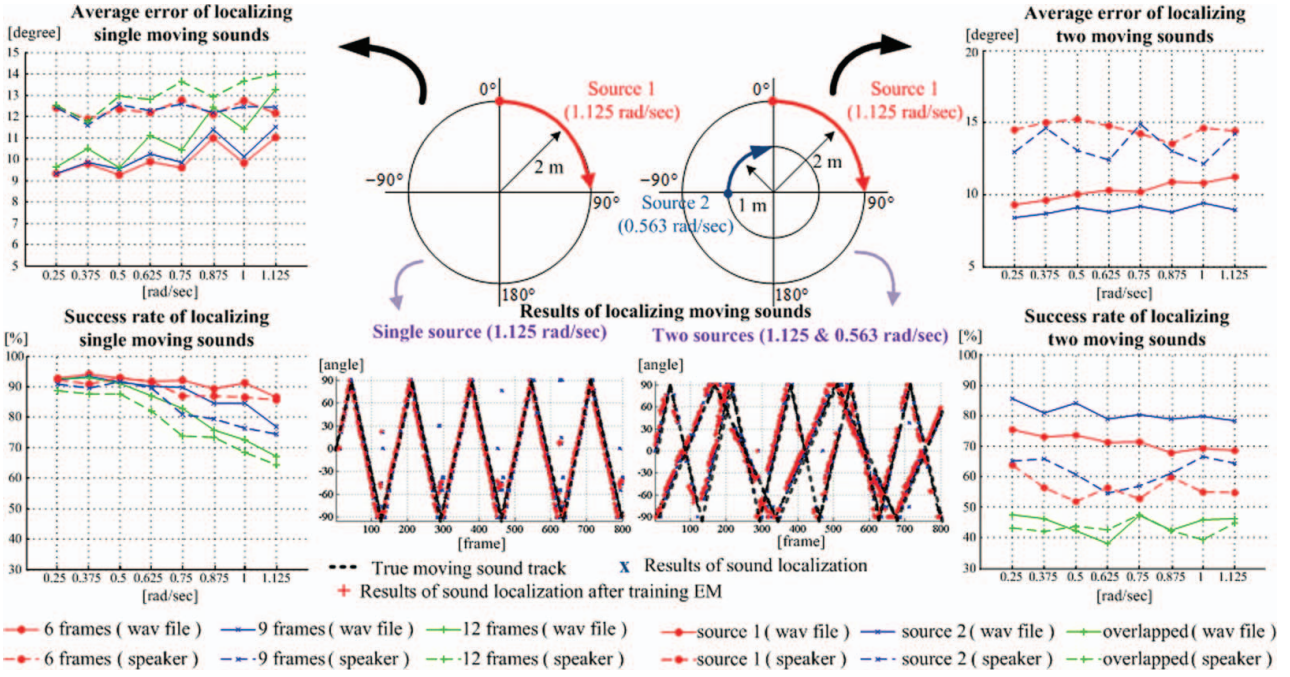

Figure 7. Results of evaluating our system for moving sounds.

(1.0 rad/s at 1 m), of healthy adults. The middle left part of Figure 7 shows that our system localised sounds moving at 1.125 rad/s for 30 s at 2 m. The middle right part of Figure 7 shows that our system localised two sounds moving at 1.125 rad/s and at 0.563 rad/s for 30 s. The right part of Figure 7 shows the average error and success rate of localising two moving sounds when the number of frames for training the EM algorithm was 6 (192 ms). The two sound sources were rotated at different angular velocities. One (source 1) rotated twice as fast as the other (source 2). The average error and success rate was better for the slower source than for the faster one. The overlapped line, in the graph of success rate of localising two moving sounds, indicated the percentage of accurate sound localisation where two sound sources occurred at the same time. Here, two sound sources have some silent intervals severally because we used the sources recorded from common dialogues. In case when two moving sounds were emitted from two speakers, as shown in the right part of Figure 7, the performances were not good because two sounds interfered with each other in the air space.

4. Active audition over entire azimuth range

The target application of our sound source localisation method is robots, and it is natural that robots move and rotate their bodies and heads in order to track someone. Therefore, even though the orientation of the microphones in the robot's head or body will constantly change, the sound source localisation method must be able to cope with the effects created by the moving microphones. Moreover, if moving robots can track sound sources, they may be able to distinguish whether sound signals are coming from their front or rear with only two microphones. This is because the TDOAs and powers obtained for equivalent sound signals coming from the front and back are the same, as shown in Figure 8A. We can overcome this problem by rotating the robot's head while the sound signals are being generated. For example, as shown in Figure 8B, if sound signals are coming from the front, the robot can determine their direction by reducing sound localization angle while turning its head. As shown in Figure 8C, if sound signals are coming from the back, the angle of sound localisation will be increased by turning the robot's head. Given this difference, our method can localise the actual source after the robot's head has turned more than 10°.

4.1. Speech classification by Gaussian mixture model (GMM)

To localise sounds over the entire azimuth range with two microphones, after the robot first classified speech signals, it has to rotate two microphones during the periods of speech signals. Therefore, we developed a voice activity detection (VAD) (Lu et al. 2002) based on Gaussian mixture model (GMM). GMM is a powerful statistical method widely used for speech classification (Bahoura and Pelletier 2004; Shah et al. 2004). We applied 0 to 12th coefficients (total 13 values) and $\Delta 1$ to $\Delta 12$th coefficients (total 12 values) of Mel frequency cepstral coefficients (MFCCs) (Shah et al.

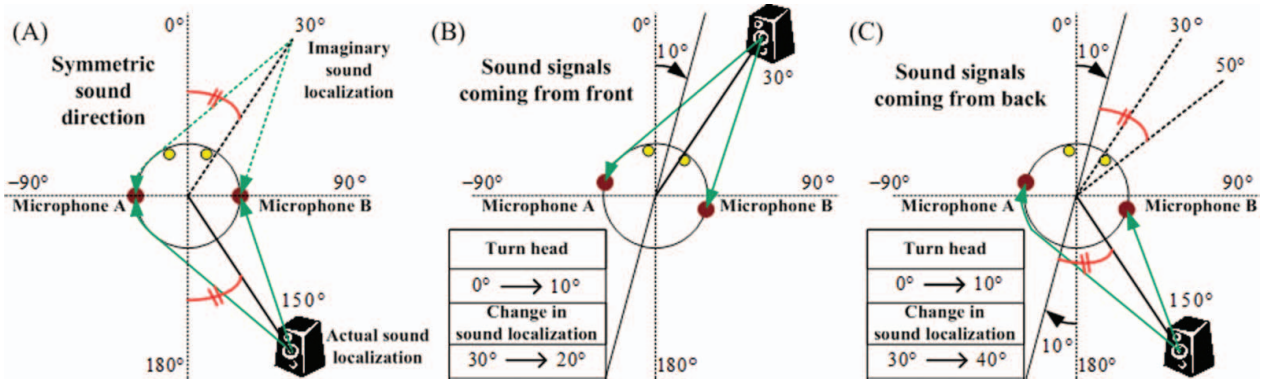

Figure 8. Sound source localisation by rotating robot's head.

2004) to a univariate Gaussian model defined as

$$P(X|\mu,\sigma) = P(X|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X-\mu)^2}{2\sigma^2}}, \quad (9)$$

where P is the component density function, X is data, μ is mean, σ is covariance and θ is the parameter vector. Then, to combine the probability values calculated by Equation (9), we used GMM defined as Equation (10), and the weight is denoted as Equation (11).

$$P_{mixture}(X_{1\sim25}|\theta_{1\sim25}) = \sum_{L=1}^{25} P_L(X_L|\theta_L)\, w(L), \quad (10)$$

$$\sum_{L=1}^{25} w(L) = 1, \quad 0 \le w(L) \le 1, \quad (11)$$

where $P_{mixture}$ is the mixture model of component density functions, $\boldsymbol{L}$ is the number of MFCC parameters, $\boldsymbol{X}$ is the value of the MFCC data of 0 to 12th and $\Delta1$ to $\Delta12$th coefficients and θ is the parameter vector concerning each MFCC value. Moreover, to classify speech signals robustly, we designed two GMM models for speech and noise derived as

$$f = \log(P_s(X_s|\theta_s)) - \log(P_n(X_n|\theta_n)), \quad (12)$$

where P_s is the GMM related to speech and X_s is the MFCC data set at the $t-$th frame belonging to the speech parameter, θ_s. On the other hand, P_n is the GMM related to noise and X_n is the MFCC data set at the $t-$th frame belonging to the noise parameter, θ_n. Finally, if the final value, f, denoted as Equation (12), is higher than the value of the threshold to discriminate the speech signal from GMM, signals at the t-th frame will be regarded as speech signals.

$$IF\ f(t) > thresold\ \text{THEN}\ f(t) = 1\ (speech)$$
$$\text{ELSE}\ f(t) = 0\ (noise). \quad (13)$$

We used 30 speech data (15 males and 15 females) for the speech parameters to train GMM parameters, and 77 noise data generated in home environments, such as the sounds of a door opening or shutting and those of electrical home appliances (e.g. a vacuum cleaner, a hair drier and a washing machine) for the noise parameters. To verify the performance of GMM parameter training, we classified the sound sources using speech and noise data for training. As a result, we obtained a success rate for speech classification of 95.5% and a success rate for noise classification of 72.8%.

4.2. *Design of sound source localisation over entire azimuth range*

According to this result as shown in Table 1, people usually rotate their heads at $42^\circ \pm 20.4^\circ$ and the most frequent combination of classes of movements involves rotating and tipping movements (0.5∼ 1 kHz) so that they can determine the accurate direction of sound sources. Consequently, we designed our system that can distinguish between sounds from the front and the rear by simply rotating its head at least by 10°. The reason that it can distinguish front from rear sources by rotating 10° is that the error margin for a single moving sound slower than 1 rad/s is about 10°, as shown in the top left part of Figure 7. Figure 9 shows the process to localise sounds over the entire azimuth range for a humanoid robot performing the following steps:

1. The robot detects speech signals classified by GMM. Our VAD requires at least 200 ms in order to discriminate between speech signals and noises. The robot can then detect the period of these signals by using

$$\sum_{i=i-n_a}^{i+n_b} f(i) \ge \text{threshold}, \quad (14)$$

where $f(i)$ is the i-th speech frame classified by Equation (13). If some speech frames exist within the interval

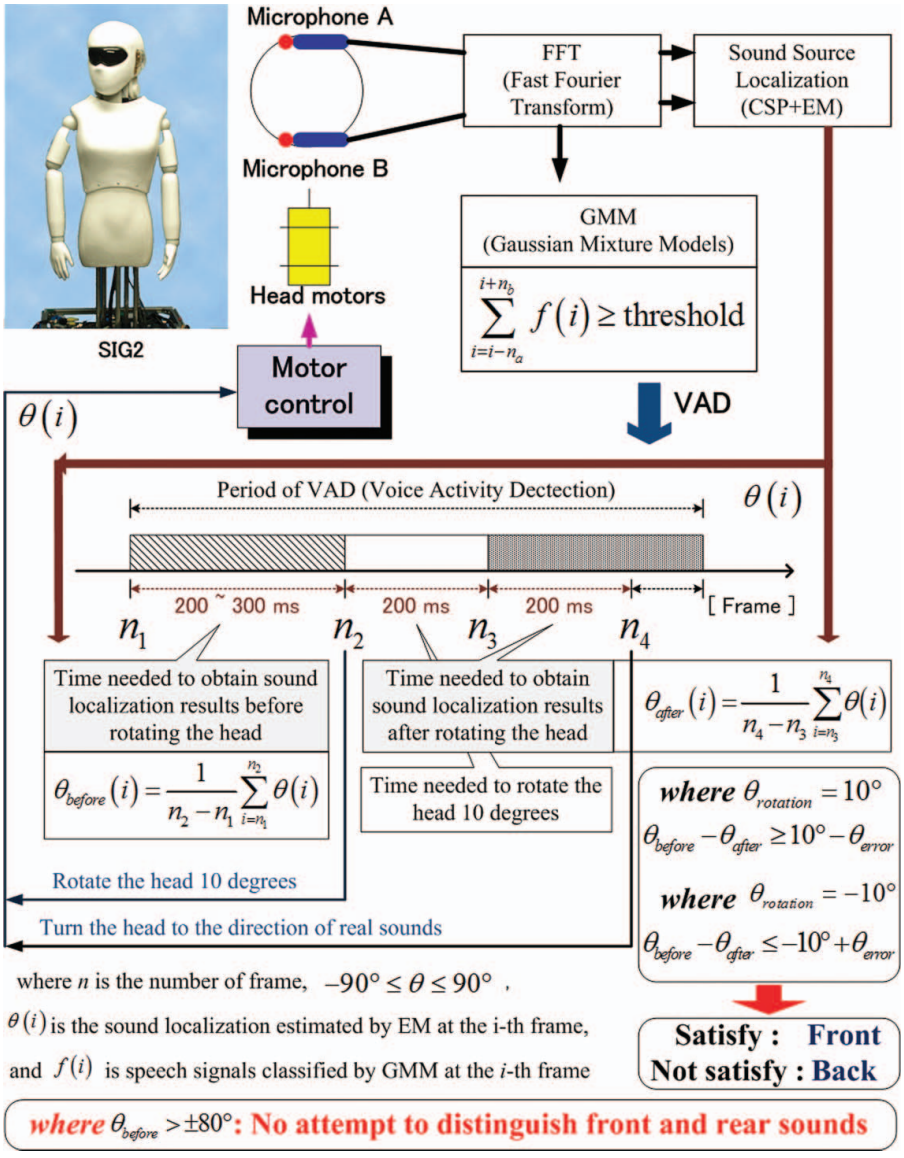

Figure 9. System overview of localising sounds for whole azimuth.

of designated frames from the n_a-th frame to the n_b-th frame, we can decide that the i-th frame is within the interval of the target speech.

2. Before turning its head 10° in the direction of the detected signals, the robot calculates the average of the sound localisation events by using

$$\theta_{before}(i) = \frac{1}{n_2 - n_1} \sum_{i=n_1}^{n_2} \theta(i) \tag{15}$$

where $\theta(i)$ is the estimated sound localisation event of the i-th frame and θ_{before} is the average angle between the n_1-th frame and the n_2-th frame.

3. After turning its head 10°, the robot obtains the average of the sound localisation events between the n_3-th frame and the n_4-th frame for 200 ms by using

$$\theta_{after}(i) = \frac{1}{n_4 - n_3} \sum_{i=n_3}^{n_4} \theta(i). \tag{16}$$

4. Finally, using the difference between the initial average angle calculated by Equation (15) and the final average angle calculated by Equation (16), the robot can localise sounds over the entire azimuth range and turn its head to that direction. In the bottom part of Figure 9, $\theta_{rotation}$ is an angle for rotating a motor and it turns the head 10° or −10° toward the angle of θ_{before}. θ_{error} is the value to compensate the error of calculating θ_{before} and θ_{after}. The system can logically distinguish between front and back localisation if sound signals are continuously generated for longer than 0.6 s as shown in Figure 9. This is because our system has a delay of more than 200 ms for detecting speech signals, 200 ms for rotating the motor by 10° and more than 200 ms for localising sounds after turning its head. Here, we rotated the head motor less than 0.25 rad/s (0.25 rotations per 1 second) in order to avoid the effect of motor noise. We confirmed that the magnitude of our motor noise is less than 55 dB (A) when rotating that less than 0.25 rad/s, at that time, our sound localisation system could work without the disturbing noise generated from the motor. In addition, where the angle region is at more than ± 80°, our system does not try to distinguish front and rear localisations because sounds are coming from the side in these cases as shown in Figure 10. Besides, although our sound localisation system over the entire azimuth range has been evaluated for fixed sounds, it would be able to cope with linearly moving sounds slower than the average walking speed, 1 m/s, of healthy adults. In future work, we will consider the evaluation of our system to distinguish whether the direction of linearly moving sounds is in the front or rear by rotating two microphones.

5. Evaluation of binaural active audition

Figure 10 shows the results of applying our system to the SIG2 robot. In this experiment, the robot distinguished between sounds coming from the front and rear whenever speech signals of 'sig', its name, were generated. The length of the speech signal was about 0.75 s, and speech signals were generated 20 times at each position. The upper left part of Figure 10 shows the success rate for sound localisation by only rotating robot's head at 10°.

In the upper left part of Figure 10, while the success rate for sound localisation in the front area was 96.5%, the one in the rear area was 65.6% due to the sound diffraction created by the artificial auricle used in SIG2 (see the bottom left part of Figure 5.3.4). To avoid a lowering of performance in the rear area, we considered sound source localisation by combining rotating with tipping. We ascertained that the most frequent combination of classes of movements involves rotating and tipping movements when blind people localise a sound (see Table 1). Therefore, to evaluate the influence on tipping movements, we experimentally compared results of our sound source localisation in front region with that of rear region. To do this experiment, after

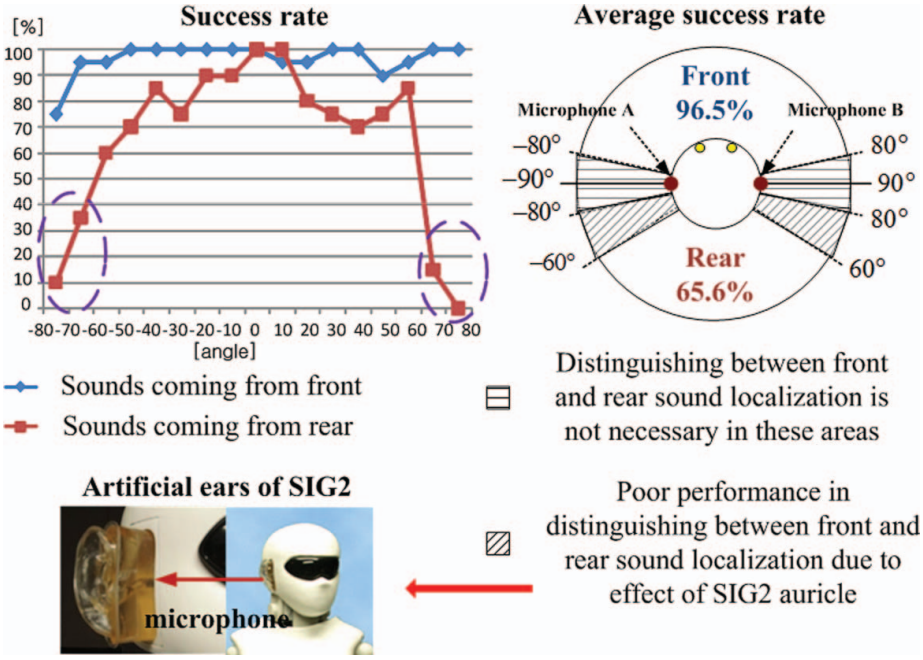

Figure 10. Results of localising sounds for whole azimuth.

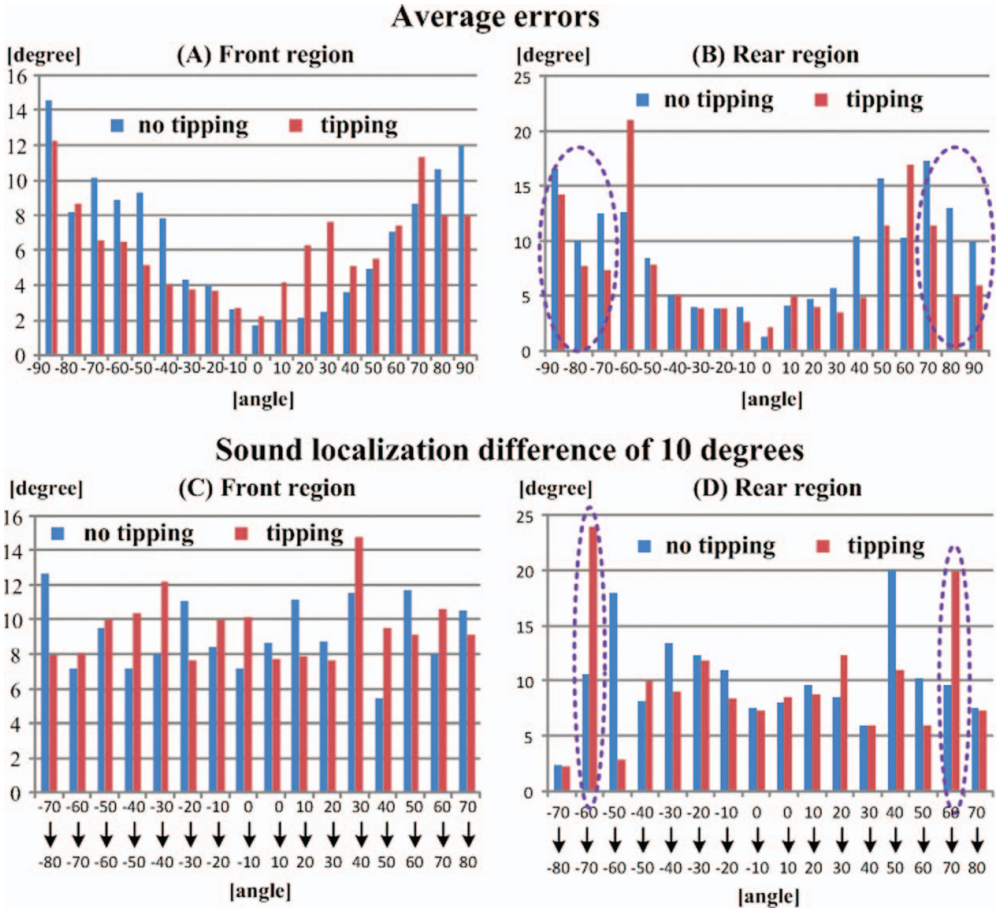

Figure 11. Sound localisation difference between tipping and no tipping.

tipping the robot's head to −30° down, we recorded five commands, 'sig', 'ohayogozaimasu', 'konnichiwa', 'konbanwa' and 'oyasuminasai', which mean 'the name of our robot', 'good morning', 'good afternoon', 'good evening' and 'good night', respectively. They were produced at every 10° over the entire azimuth range, at a distance of 1.5 m from the head of the robot and at a magnitude of 85 dB (A). Since the robot was at the centre of a square room with

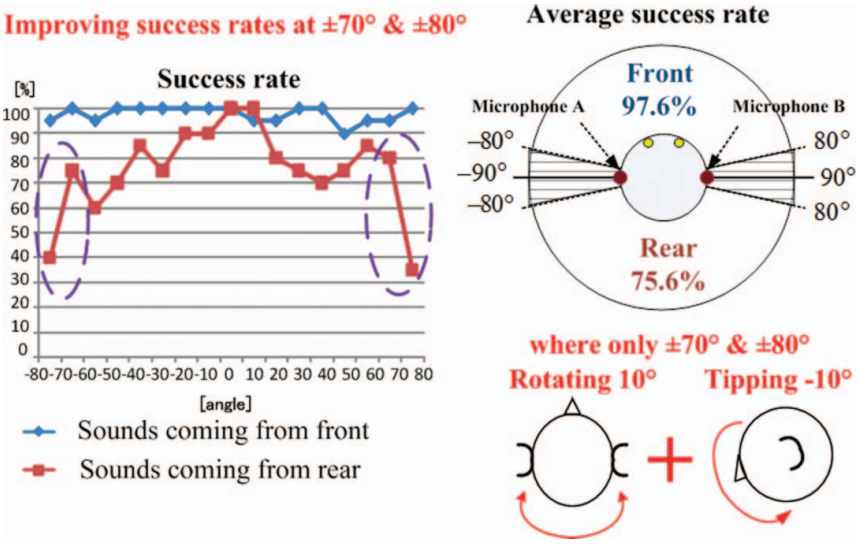

Figure 12. Improvement in localising sounds for whole azimuth with tipping.

sides of about 5 m and background noise of about 55 dB (A), the reverberation effect was neglected. We calculated the average results for each measurement point.

As shown in Figure 11A, while most average errors in the front region were usually about 10° regardless of tipping or no tipping, at ±70° and ± 80° in the rear area, average errors with tipping were less than those with no tipping as shown in Figure 11B. In addition, at sound localisation difference from ±60° to ±70°, the sound localisation difference with tipping was much larger than that with no tipping as shown in Figure 11D. We analysed that localising sounds after tipping the head down and rotating it ±10° is more effective at ±70° and ±80° in the rear area than without tipping due to the artificial auricle of SIG2. Therefore, we applied tipping and rotating movements to disambiguate front–rear confusions at ±70° and ±80° in the rear area which have poor performance as shown in Figure 10. We confirmed that the success rate for sound localisation in the rear area can be increased by about 10 points if combining rotating with tipping movements at only ±70° and ±80° in the rear area (other azimuths used only rotating movements), as shown in Figure 12.

6. Conclusion

We have dealt with problems which still have been left on binaural audition systems for robots.

1. Sound localisation systems for robots should use the minimum number of microphones and no prior information because these are compatible with various devices at low cost and do not usually require the large amount of computing power.
2. Since it is natural that robots move and rotate their bodies and heads, the method should localise moving sounds while coping well with the effects created by moving microphones.
3. Two microphones which are symmetrically installed in a robot's head can't generally distinguish whether a sound source is coming from the front or rear.

For the first goal, we first integrated CSP analysis with an EM algorithm to accurately localise moving sounds without having to measure impulse response data in advance. Second, we used 3D moving sound creation tool called SoundLocus Lite developed by Arinis sound technologies. Therefore, we created moving sounds including accurate track information such as an azimuth and a distance according to a created frame or time. We then evaluated our sound localisation method using created moving sounds instead of recorded sounds while practically moving a speaker. This way enables us to evaluate various sound localisation applications in the same condition. Finally, we solve the problem of distinguishing whether sounds are coming from the front or back by rotating and/or tipping a robot's head equipped with only two microphones.

In future work, sounds in real environments contain reverberation, reflection, resonance or ambient noises. Since these effects are critical for evaluating sound localisation systems, we should take these into account to compare them to real sounds. Also, we only considered rotating and tipping of at least 10° so that the robot could distinguish whether a sound was coming from the front or the rear. In next step, we need to evaluate the effect of sound source localisation according to the three motions of rotating, tipping and pivoting in order to localise multiple and/or moving sources.

Acknowledgements

This research was partially supported by MEXT, Grant-in-Aid for Scientific Research and Global COE program of MEXT, Japan.

References

Bahoura M, Pelletier C. 2004. Respiratory sound classification using cepstral analysis and gaussian mixture models. Proceedings of the IEEE/EMBS International Conference; Sep. 1–5; San Francisco, USA.

Berglund EJ. 2005. Active audition for robots using parameter-less self-organising maps. Ph.D. thesis (October). The University of Queensland, Australia.

Blauert J. 1996. Spatial hearing—the psychophysics of human sound localization. Rev. ed. Cambridge, MA: The MIT Press.

Cheng CI, Wakefield GH. 2001. Introduction to head-related transfer functions (HRTFs): space. J Audio Eng Soc. 49(4):231–248.

Hara I, Asano F, Kawai Y, Kanehiro F, Yamamoto K. 2004. Robust speech interface based on audio and video information fusion for humanoid HRP-2. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS-2004); October; Sendai, Japan. p. 2404–2410.

Huang J, Ohnishi N, Sugie N. 1998. Spatial localization of sound sources: azimuth and elevation estimation. In: Proceedings of IEEE/IMTC International Conference on Instrumentation and Measurement Technology; May; St. Paul, MN, USA. p. 330–333.

Hwang S, Park Y, Park Y. 2005. Sound source localization using HRTF database. In: Proceedings of International Conference on Control, Automation, and Systems (ICCAS2005); June; Busan, South Korea. p. 751–755.

Kim H-D. 2008. Binaural active audition for humanoid robots. Ph.D. thesis (September). Kyoto University, Japan.

Kim H-D, Choi J-S, Kim M. 2007a. Human-robot interaction in real environments by audio-visual integration. Int J Control, Automation, and Systems. 5(1):61–69.

Kim H-D, Komatani K, Ogata T, Okuno HG. 2007b. Real-time auditory and visual talker tracking through integrating EM algorithm and particle filter. IEA/AIE-2007, LNAI 4570; June. Kyoto, Japan: Springer-Verlag. p. 280–290.

Lu L, Zhang H-J, Jiang H. 2002. Content analysis for audio classification and segmentation. IEEE Trans Speech Audio Process. 10(7):504–516.

Moon TK. 1996. The expectation-maximization algorithm. IEEE Signal Process Mag. 13(6):47–60.

Nakadai K, Hidai K-i, Okuno HG, Kitano H. 2002. Real-time speaker localization and speech separation by audio-visual

integration. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS-2002); May; Washington DC, USA. p. 1043–1049.

Nishiura T, Yamada T, Nakamura S, Shikano K. 2000. Localization of multiple sound sources based on a CSP analysis with a microphone array. In: Proceedings of IEEE/ICASSP International Conference on Acoustics, Speech, and Signal Processing; June; Istanbul, Turkey. p. 1053–1056.

Schmidt RO. 1986. Multiple emitter location and signals parameter estimation. IEEE Trans Antennas and Propagation. AP-34:276–280.

Shah JK, Iyer AN, Smolenski BY, Yantormo RE. 2004. Robust voiced/unvoiced classification using novel feature and Gaussian mixture model. Paper presented at: IEEE/ICASSP International Conference on Acoustics, Speech, and Signal Processing; May; Montreal, Canada.

Thurlow WR, Mangels JW, Runge PS. 1967. Head movements during sound localization. Journal of the Acoustical Society of America. 42(2):489–493.

Valin J-M, Yamamoto S, Rouat J, Michaud F, Nakadai K, Okuno HG. 2007. Robust recognition of simultaneous speech by a mobile robot. IEEE Trans Robot. 23(4):742–752.