Monaural Sound-Source-Direction Estimation Using the Acoustic Transfer Function of an Active Microphone

Ryoichi Takashima **Tetsuya Takiguchi** **Yasuo Ariki**
Department of Computer Science and Systems Engineering
Kobe University, Japan
takashima@me.cs.scitec.kobe-u.ac.jp takigu@kobe-u.ac.jp ariki@kobe-u.ac.jp

Abstract – *This paper introduces an active microphone concept that achieves a good combination of active-operation and signal processing, where a new sound-source-direction estimation method using only a single microphone with a parabolic reflection board is proposed. A simple signal-power-based method using a parabolic antenna has been proposed in the radar field. But the signal-power-based method is not effective for finding the direction of a talking person due to the varying power of the uttered speech signals. In this paper, the sound-source-direction estimation method focuses on the acoustic transfer function instead of the signal power. The use of the parabolic reflection board leads to a difference in the acoustic transfer functions of the target direction and the non-target directions, where the active microphone rotates and observes the speech at each angle. The acoustic transfer function is estimated from the observed speech using the statistics of clean speech signals. Its effectiveness is confirmed by monaural sound-source-direction estimation experiments in a room environment.*

Keywords: Direction of arrival estimation, acoustic reflection, microphones.

1 Introduction

Many systems using microphone arrays have been tried in order to localize sound sources. Conventional techniques, such as MUSIC, CSP, and so on (e.g., [1, 2]), use simultaneous phase information from microphone arrays to estimate the direction of the arriving signal. Also, sound source localization techniques focusing on the auditory system have been described in [3, 4].

Single-microphone source separation is one of the most challenging scenarios in the field of signal processing, and some techniques have been described (e.g., [5, 6, 7, 8]). In our previous work [9], we discussed a sound source localization method using only a single microphone. In that report, the acoustic transfer function was estimated from observed (reverberant) speech using the statistics of clean speech signals without using texts of the user's utterance, where a GMM (Gaussian Mixture Model) was used to model the features of the clean speech. This estimation is performed in the cepstral domain employing a maximum-likelihood-based approach. This is possible because the cepstral parameters are an effective representation for retaining useful clean speech information. The experiment results of our talker-localization showed its effectiveness. However, the previous method required the measurement of speech for each room environment in advance. Therefore, this paper presents a new method that uses parabolic reflection that is able to estimate the sound source direction without any need for such prior measurements.

In this paper, we introduce the concept of an active microphone that achieves a good combination of active-operation and signal processing. The active microphone has a parabolic reflection board, which is extremely simple in construction. The reflector and its associated microphone rotate together, perform signal processing, and seek to locate the direction of the sound source.

A simple signal-power-based method using a parabolic antenna has been proposed in the radar field. But the signal-power-based method is not effective for finding the direction of a person talking in a room environment. One of the reasons is that the power of the speech signal varies for all directions of the parabolic antenna, since a person does not utter the same power (word) for all directions of the parabolic antenna. Therefore, in this paper, our new sound-source-direction estimation method focuses on the acoustic transfer function instead of the signal power. The use of the parabolic reflection board results in a difference in the acoustic transfer functions of the target direction and the non-target directions, where the active microphone with the parabolic reflection board rotates and observes the speech at each angle. The sound source direction is detected by comparing the acoustic transfer

functions observed at each angle, which are estimated from the observed speech using the statistics of clean speech signals. Its effectiveness is confirmed by sound-source-direction estimation experiments in a room environment.

2 Active microphone

2.1 Parabolic reflection board

In this paper, an active microphone with a parabolic reflection board is introduced for estimation of sound source direction, where the reflection board has the shape of a parabolic surface. Under the assumption of the plane wave, any line (wave) parallel to the axis of the parabolic surface is reflected toward the focal point. On the other hand, if the sound source is not located at 90 degrees (in front of the parabolic surface), no reflection wave will travel toward the focal point. Therefore, the use of the parabolic reflection board will be able to give us the difference in the acoustic transfer function between the target direction and the non-target directions.

2.2 Signal observed using parabolic reflection

Next, we consider the signal observed using parabolic reflection [11]. In [11], a simple signal-power-based method using a parabolic reflection board has been described. Its effectiveness has been confirmed on white noise signals, but the signal-power-based method was not effective for finding the direction of a talking person due to the varying power of the uttered speech signals.

As shown in Figure 1(a), when the sound source is located directly in front of the parabolic surface and there is no background noise, the observed signal at the focal point at time t can be expressed by the addition of the waves arriving at the focal point directly (direct wave) and those arriving at the focal point after being reflected by the parabolic surface (reflection waves):

$$o(t) = x_0(t) + \sum_{m=1}^{M} x_m(t) \tag{1}$$

where $o(t)$, x_0 and x_m $(m = 1, \cdots, M)$ are observed sound, direct sound and reflection sound, respectively. Based on the property of a parabola, the time difference to the focal point between the direct and reflection waves is constant without depending on m. Therefore, (1) can be written as

$$o(t) = s(t) * h_0(t) + \sum_{m=1}^{M} s(t-\tau) * h_m(t) \tag{2}$$

where $s(t)$ and τ are clean speech and the time difference, respectively. h_0 is the acoustic transfer function of a direct wave and h_m is that of a reflection wave.

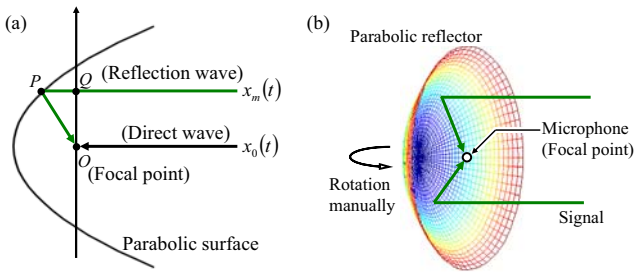

Figure 1: (a) Observed signal at the focal point, where the input signal is coming from directly in front of the parabolic surface. (b) Active microphone with parabolic reflection.

By applying the short-term Fourier transform, the observed spectrum at frame n is given by

$$\begin{aligned} & O(\omega; n) \\ & \approx S(\omega; n) \cdot (H_0(\omega; n) + e^{-j2\pi\omega\tau} \cdot \sum_{m=1}^{M} H_m(\omega; n)) \\ & = S(\omega; n) \cdot (H_p(\omega; n) + H_r(\omega; n)). \end{aligned} \tag{3}$$

Here H_p is the acoustic transfer function of the direct sound that is not influenced by parabolic reflection. H_r is the acoustic transfer function resulting from parabolic reflection.

On the other hand, when the sound source is not located in front of the parabolic surface, parabolic reflection does not influence the acoustic transfer function since no reflection waves will travel toward the focal point:

$$O(\omega; n) \approx S(\omega; n) \cdot H_0(\omega; n) = S(\omega; n) \cdot H_p(\omega; n). \tag{4}$$

2.3 Estimation of sound source direction

As shown in Figure 1(b), a new active microphone with a parabolic reflection board was constructed with the microphone located at the focal point. In order to obtain the signal observed at each angle, the angle of the microphone was changed manually in research carried out for this paper. Then, from equations (3) and (4), the spectrum of the signal observed at a microphone angle θ can be expressed as

$$\begin{aligned} O_\theta(\omega; n) & \approx S_\theta(\omega; n) \cdot H_\theta(\omega; n) \\ H_\theta(\omega; n) & = \begin{cases} H_p(\omega; n) + H_r(\omega; n) & (\theta = \hat{\theta}) \\ H_p(\omega; n) & (\theta \neq \hat{\theta}) \end{cases} \end{aligned} \tag{5}$$

where S_θ and H_θ are spectra of clean speech and acoustic transfer function at the angle θ and $\hat{\theta}$ is the sound source direction. Assuming H_p is nearly constant for each angle, when the active microphone does not face the sound source, the value of H_θ will be almost the

same for every non-target directions. On the other hand, the only condition under which H_θ will have a different value from that obtained at the other angles is when the active microphone faces the sound source.

Therefore, the sound source direction can be estimated by selecting the direction whose the acoustic transfer function is the farthest from the acoustic transfer functions of other directions:

$$\hat{\theta} = \underset{i}{\operatorname{argmax}} \sum_j (\bar{H}_i - \bar{H}_j)^2 \quad (6)$$

where i and j are the angle of microphone, and $\bar{H}$ is the expectation of H. Actually, in this research, the cepstrum of acoustic transfer function is used to calculate this equation. In the next section, we will describe how to estimate H_i from observed speech signals.

3 Estimation of the acoustic transfer function

In our previous work [9], we proposed a method to estimate the acoustic transfer function from the reverberant speech (any utterance) using the clean-speech acoustic model, where a GMM is used to model the feature of the clean speech. The clean speech GMM enables us to estimate the acoustic transfer function from the observed speech without needing to have texts of the user's utterance (text-independent estimation). However, because an active microphone with parabolic reflection board was not used, the previous method required the measurement of speech for each room environment in advance in order to be able to determine the direction of a talking person. In this paper, we can estimate the sound source direction without any need for prior measurements by information fusion of an active microphone and an estimation of an acoustic transfer function.

3.1 Cepstrum representation of reverberant speech

The observed signal (reverberant speech), $o(t)$, in a room environment is generally considered as the convolution of clean speech and acoustic transfer function. The spectral analysis of the acoustic modeling is generally carried out using short-term windowing. Therefore, the observed spectrum is approximately represented by $O(\omega; n) \approx S(\omega; n) \cdot H(\omega; n)$, where the length of the acoustic transfer function may be greater than that of the window. Here $O(\omega; n)$, $S(\omega; n)$, and $H(\omega; n)$ are the short-term linear spectra in the analysis window n.

Cepstral parameters are an effective representation for retaining useful speech information in speech recognition. Therefore, we use the cepstrum for acoustic modeling necessary to estimate the acoustic transfer function. The cepstrum of the observed signal is given by the inverse Fourier transform of the log spectrum:

$$O_{cep}(d; n) \approx S_{cep}(d; n) + H_{cep}(d; n) \quad (7)$$

where O_{cep}, S_{cep}, and H_{cep} are cepstra for the observed signal, clean speech signal, and acoustic transfer function, respectively. As shown in equation (7), if O and S are observed, H can be obtained by

$$H_{cep}(d; n) \approx O_{cep}(d; n) - S_{cep}(d; n). \quad (8)$$

However S cannot be observed actually. Therefore H is estimated by maximizing the likelihood (ML) of observed speech using clean-speech GMM.

3.2 Maximum-likelihood-based parameter estimation

The sequence of the acoustic transfer function in (8) is estimated in an ML manner [10] by using the expectation maximization (EM) algorithm, which maximizes the likelihood of the observed speech:

$$\hat{H} = \underset{H}{\operatorname{argmax}} \Pr(O|H, \lambda_S). \quad (9)$$

Here, λ denotes the set of GMM parameters of the clean speech, while the suffix S represents the clean speech in the cepstral domain. The GMM of clean speech consists of a mixture of Gaussian distributions.

$$\lambda_S = \{w_k, N(\mu_k^{(S)}, \sigma_k^{(S)^2})\}, \quad \textstyle\sum_k w_k = 1 \quad (10)$$

where w_k, μ_k and σ_k^2 are the weight coefficient, mean vector and variance vector (diagonal covariance matrix) of the k-th mixture component, respectively. Those parameters are estimated by EM (Expectation-Maximization) algorithm using a clean speech database.

The estimation of the acoustic transfer function in each frame is performed in a maximum likelihood fashion by using the EM algorithm. The EM algorithm is a two-step iterative procedure. In the first step, called the expectation step, the following auxiliary function Q is computed.

$$\begin{aligned} & Q(\hat{H}|H) \\ &= E[\log \Pr(O, c|\hat{H}, \lambda_S)|H, \lambda_S] \\ &= \sum_c \frac{\Pr(O, c|H, \lambda_S)}{\Pr(O|H, \lambda_S)} \cdot \log \Pr(O, c|\hat{H}, \lambda_S) \end{aligned} \quad (11)$$

Here c represents the unobserved mixture component labels corresponding to the observation sequence O.

The joint probability of observing sequences O and c can be calculated as

$$\Pr(O, c|\hat{H}, \lambda_S) = \prod_{n^{(v)}} w_{c_{n^{(v)}}} \Pr(O_{n^{(v)}}|\hat{H}, \lambda_S) \quad (12)$$

where w is the mixture weight and $O_{n^{(v)}}$ is the cepstrum at the n-th frame for the v-th training data (observation

data). Since we consider the acoustic transfer function as additive noise in the cepstral domain, the mean to mixture k in the model λ_O is derived by adding the acoustic transfer function. Therefore, equation (12) can be written as

$$\begin{aligned}&\Pr(O, c|\hat{H}, \lambda_S)\\ &= \prod_{n^{(v)}} w_{c_{n^{(v)}}} \cdot N(O_{n^{(v)}}; \mu^{(S)}_{k_{n^{(v)}}} + \hat{H}_{n^{(v)}},\ \Sigma^{(S)}_{k_{n^{(v)}}}) \quad (13)\end{aligned}$$

where $N(O; \mu, \Sigma)$ denotes the multivariate Gaussian distribution. It is straightforward to derive that

$$\begin{aligned}&Q(\hat{H}|H)\\ &= \sum_k \sum_{n^{(v)}} \Pr(O_{n^{(v)}}, c_{n^{(v)}} = k|\lambda_S) \log w_k\\ &\quad + \sum_k \sum_{n^{(v)}} \Pr(O_{n^{(v)}}, c_{n^{(v)}} = k|\lambda_S)\\ &\quad \cdot \log N(O_{n^{(v)}}; \mu^{(S)}_k + \hat{H}_{n^{(v)}}, \Sigma^{(S)}_k) \quad (14)\end{aligned}$$

Here $\mu^{(S)}_k$ and $\Sigma^{(S)}_k$ are the k-th mean vector and the (diagonal) covariance matrix in the clean speech GMM, respectively. It is possible to train those parameters by using a clean speech database. Next, we focus only on the term involving H.

$$\begin{aligned}Q(\hat{H}|H) \ = \ -\sum_k \sum_n \gamma_k(n) \sum_{d=1}^{D} \Bigg\{ &\frac{1}{2} \log(2\pi)^D \sigma^{(S)^2}_{k,d}\\ &+ \frac{(O(d;n) - \mu^{(S)}_{k,d} - \hat{H}(d;n))^2}{2\sigma^{(S)^2}_{k,d}} \Bigg\} \quad (15)\end{aligned}$$

$$\gamma_k(n) = \Pr(O(n), k|\lambda_S) \quad (16)$$

Here $O(n)$ is the cepstrum at the n-th frame for observed speech data. D is the dimension of the $O(n)$, and $\mu^{(S)}_{k,d}$ and $\sigma^{(S)^2}_{k,d}$ are the d-th mean value and the d-th diagonal variance value of the k-th component in the clean speech GMM, respectively.

The maximization step (M-step) in the EM algorithm becomes "max $Q(\hat{H}|H)$". The re-estimation formula can, therefore, be derived, knowing that $\partial Q(\hat{H}|H)/\partial \hat{H} = 0$ as

$$\hat{H}(d;n) = \frac{\sum_k \gamma_k(n) \frac{O(d;n) - \mu^{(S)}_{k,d}}{\sigma^{(S)^2}_{k,d}}}{\sum_k \frac{\gamma_k(n)}{\sigma^{(S)^2}_{k,d}}}. \quad (17)$$

Therefore, the sound source direction is estimated by equation (6) using cepstral vector $\hat{H}(d;n)$ to calculate $\bar{H}_i$ or $\bar{H}_j$.

4 Experiment

4.1 Experiment conditions

The direction estimation experiment was carried out in a real room environment. The parabolic reflection microphone shown in Figure 1 was used for the experiments. The diameter was 24 cm, and the distance of the focal point was 9 cm. The microphone located at the focal point is an omnidirectional type (SONY ECM-77B). The target sound source was located at 90 degrees and 2 m from the microphone. The angle of the parabolic reflection microphone was changed manually from 30 degrees to 150 degrees in increments of 20 degrees. Then the acoustic transfer function of the target signal at each angle was estimated for the following speech length: 0.5, 1.0, 1.5, 2.0, 2.5, and 3.0 seconds. The size of the recording room was about 6.3 m × 7.2 m (width × depth).

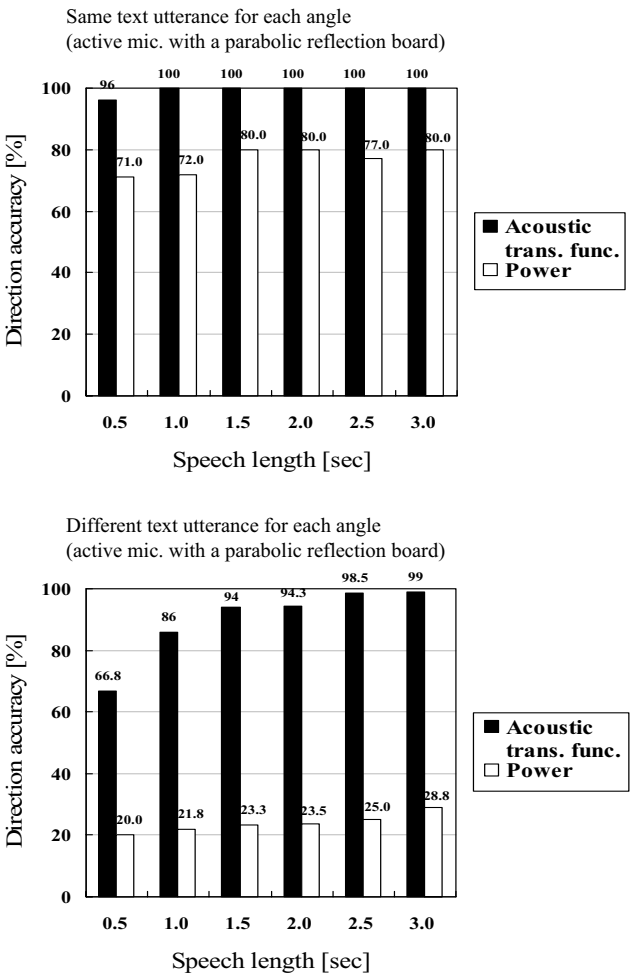

Figure 2: Performance of an active microphone with a parabolic reflection board

The speech signal was sampled at 12 kHz, and windowed with a 32-msec Hamming window every 8 msec. The clean speech GMM was trained by using 50 sentences (spoken by a female) in the ASJ Japanese speech database. The trained GMM has 64 Gaussian mixture components. Then 2nd-order MFCCs (Mel-Frequency Cepstral Coefficients) were used as feature vectors.

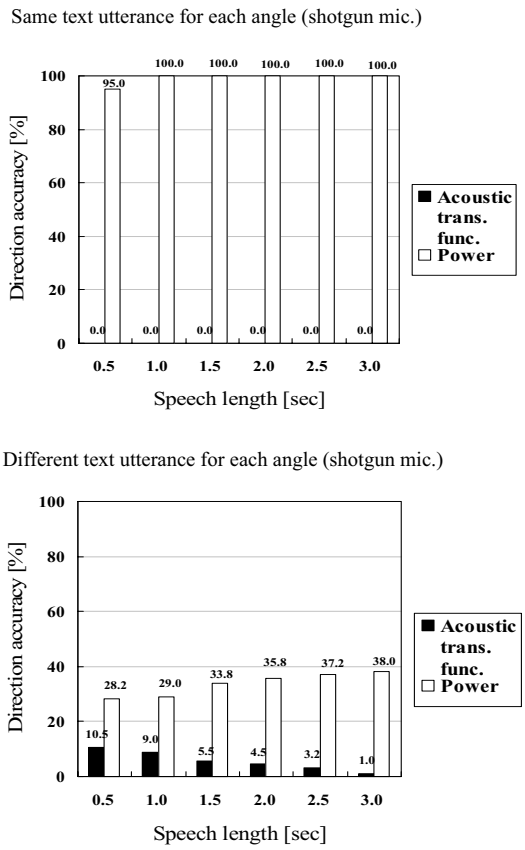

Figure 3: Performance of a shotgun microphone without a parabolic reflection board

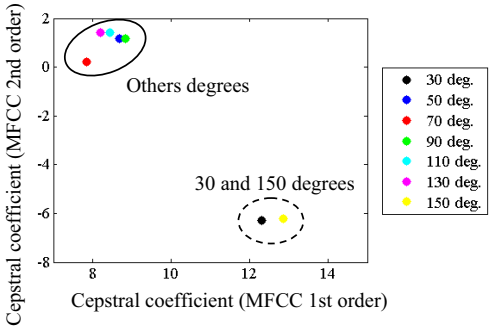

Figure 4: Mean values of the acoustic transfer functions for the shotgun microphone

4.2 Experiment results

Figure 2 shows the performance of the direction accuracy using the acoustic transfer function estimated in various speech length, and the performance is compared to the power-based technique. The top figure shows the accuracy for the same text utterance at each angle of the active microphone, and the bottom figure shows the accuracy for a different text utterance at each angle of the active microphone. As shown in the top figure, the performance for both the techniques based on the power and the acoustic transfer function is high. But the possibility of the same text utterance at each angle of the active microphone will be very small in a real environment.

In the bottom portion of Figure 2, we can see that the performance of the power-based technique degrades drastically when the utterance text differs at each angle of the active microphone, because the power of the speech signal varies for all directions of the active microphone. On the other hand, the performance of the new method based on the acoustic transfer function is high, even for the different text utterance. This is because the new method uses the information of the acoustic transfer function, which depends on the direction of the active microphone only and does not depend on the utterance text. Also, we can see that the shorter the speech length for each angle is, the more the direction accuracy decreases. One of the reasons is that the statistics for the observed speech is not readily available if not enough samples are used to estimate the acoustic transfer function.

Figure 3 shows the performance of a shotgun microphone (SONY ECM-674) without a parabolic reflection board. The power-based method can provide good performance for the same text utterance at each angle of the shotgun microphone due to the directivity of the shotgun microphone, but the performance degrades when the utterance text differs at each angle of the shotgun microphone. On the other hand, the performance of the new method based on the acoustic transfer function is even lower. The directivity of the shotgun microphone changes drastically as the sound-source direction changes from the front direction to the side directions of the shotgun microphone, and as a result, the acoustic transfer function that is farthest from all the other acoustic transfer functions becomes to be that at 30 or 150 degrees in equation (6). The mean values of all acoustic transfer functions are plotted in Figure 4, where the acoustic transfer function is computed by (8) using true clean speech signal, $S_{cep}(d;n)$, and then the mean values are computed. As shown in Figure 4, we can see that the acoustic transfer function that is farthest from all the other acoustic transfer functions is that at 30 or 150 degrees.

Figure 5 and Figure 6 show the plot of acoustic transfer function for 300 segments of observed speech for the case of the active microphone. In Figure 5, the acoustic transfer function H_{sub} was computed by (8) using true clean speech signal, $S_{cep}(d;n)$. On the other hand, in Figure 6, the acoustic transfer function H_{est} was estimated by (17) using only the statistics of clean speech GMM. As shown in Figure 5, when the active microphone does not face the sound source, H_{sub} is distributed in almost the same place. And H_{sub} of the sound source direction is distributed away from the H_{sub} of other directions. In Figure 6, though the dis-

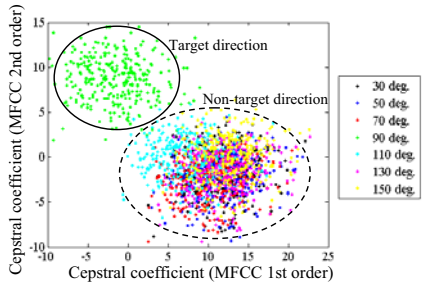

Figure 5: Acoustic transfer function computed by using true clean speech data at each angle in the cepstral domain

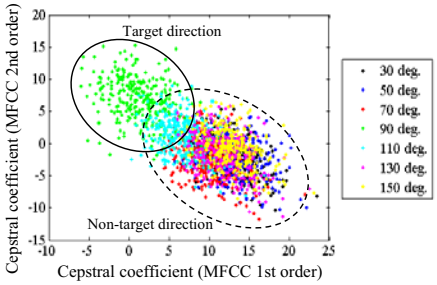

Figure 6: Acoustic transfer function estimated by the proposed method using only the statistics of clean speech GMM at each angle in the cepstral domain

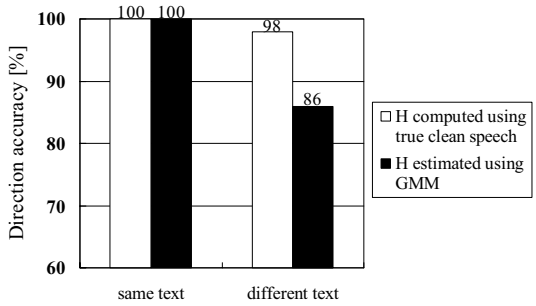

Figure 7: Comparison of true clean speech data and clean speech model

tribution of the estimated H_{est} may have some slight variations, it can be said that the distribution of H_{est} is similar that of H_{sub}.

Figure 7 shows the difference in the direction accuracy between the use of H_{sub} (the true clean speech data) and H_{est} (the statistics of clean speech model: GMM). As shown in this figure, when the utterances for each angle consist of the same text, the direction accuracy was 100%. However, when the texts of utterances for each angle are different, the direction accuracy obtained using H_{est} decreased. This is because the value of H_{est} was influenced to some extent by the phoneme sequence of clean speech.

5 Conclusions

This paper has introduced the concept of an active microphone that achieves a good combination of active-operation and signal processing, and described a sound-source-direction estimation method using a single microphone. The experiment results in a room environment confirmed that the acoustic transfer function influenced by parabolic reflection can clarify the difference between the target direction and the non-target direction. In future work, more research will be needed in regard to different utterances and direction estimation in short intervals. Also, we intend to investigate the performance of the proposed system in noisy environments, such as with multiple sound sources and when the orientation of the speaker's head changes, and to test the performance of the system in a speaker-independent speech model.

References

[1] D. Johnson and D. Dudgeon, "Array Signal Processing," Prentice Hall, 1996.

[2] M. Omologo and P. Svaizer, "Acoustic Event Localization in Noisy and Reverberant Environment Using CSP Analysis," *Proc. ICASSP*, pp. 921-924, 1996.

[3] O. Ichikawa, T. Takiguchi and M. Nishimura, "Sound Source Localization using a Pinna-Based Profile Fitting Method," *Proc. Int. Workshop on Acoustic Echo and Noise Control*, pp. 263-266, September 2003.

[4] N. Ono, Y. Zaitsu, T. Nomiyama, A. Kimachi and S. Ando, "Biomimicry Sound Source Localization with Fishbone," *IEEJ Trans. Sensors and Micromachines*, 121-E, no. 6, pp. 313-319, 2001.

[5] T. Kristjansson, H. Attias and J. Hershey, "Single Microphone Source Separation Using High Resolution Signal Reconstruction," *Proc. ICASSP*, pp. 817-820, 2004.

[6] B. Raj, M. V. S. Shashanka and P. Smaragdis, "Latent Direchlet Decomposition for Single Channel Speaker Separation," *Proc. ICASSP*, pp. 821-824, 2006.

[7] G.-J. Jang. T.-W. Lee and Y.-H. Oh, "A Subspace Approach to Single Channel Signal Separation Using Maximum Likelihood Weighting Filters," *Proc. ICASSP*, pp. 45-48, 2003.

[8] T. Nakatani, B.-H. Juang, "Speech Dereverberation Based on Probabilistic Models of Source and Room Acoustics," *Proc. ICASSP*, pp. I-821-I-824, 2006.

[9] T. Takiguchi, Y. Sumida and Y. Ariki, "Estimation of Room Acoustic Transfer Function Using Speech Model," *IEEE SSP Workshop*, pp. 336-340, 2007.

[10] B.-H. Juang, "Maximum-likelihood estimation of mixture multivariate stochastic observations of Markov chains," AT&T Tech. J., Vol. 64, No. 6, pp. 1235-1249, 1985.

[11] T. Takiguchi, R. Takashima, and Y. Ariki, "Active Microphone with Parabolic Reflection Board for Estimation of Sound Source Direction," *Proc. Joint workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA2008)*, pp. 65-68, 2008.