Chapter 2

Evaluation of an Active Microphone with a Parabolic Reflection Board for Monaural Sound-Source-Direction Estimation

Tetsuya Takiguchi, Ryoichi Takashima and
Yasuo Ariki

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/56045

1. Introduction

For human-human or human-computer interaction, the talker's direction and location are important cues that determine who is talking. This information can be helpful, especially in multi-user conversation scenarios such as a meeting system, robotic communication, and so on. There have been studies for understanding of a conversation scene based on the talker localization approach (e.g., [1, 2]). An approach using the turn-taking information obtained from DOA (Direction-of-Arrival) estimation results for the discrimination of system requests or users' conversations has also been proposed [3].

Many systems using microphone arrays have been tried in order to localize sound sources. Conventional techniques, such as MUSIC (MUltiple SIgnal Classification), CSP (Cross-power Spectrum Phase), and so on (e.g., [4–9]), use simultaneous phase information from microphone arrays to estimate the direction of the arriving signal. There have also been studies on binaural source localization based on interaural differences, such as interaural level difference and interaural time difference (e.g., [10, 11]). Sound source localization techniques focusing on the auditory system have also been described in [12, 13].

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., [14–17]). In our previous work [18], 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 chapter 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 chapter, 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. We call this microphone with the function of the rotation an active microphone.

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 chapter, 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. We compared our proposed method with the signal-power-based method, and as the methods for obtaining the directivity of the microphone, we compared the use of the parabolic reflection board with the use of a shotgun microphone. Its effectiveness is confirmed by sound-source-direction estimation experiments in a room environment.

2. Active microphone

2.1. Parabolic reflection board

In this chapter, 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. The parabolic reflector has been used for estimation of the direction of arrival in the radar field [19]. As shown in Figure 1, under the assumptions associated with plane waves, 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-power-based estimation of sound source direction

In [20], a simple signal-power-based method using a parabolic reflection board has been described. The use of parabolic reflection can increase the power gain of the signal arriving from directly in front of the parabolic board. In that method, the microphone with a parabolic

http://dx.doi.org/10.5772/56045

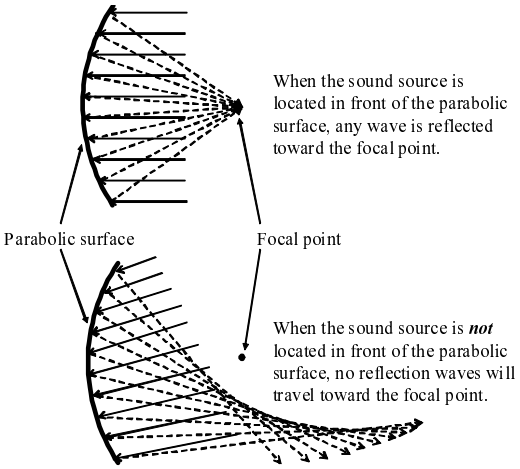

Figure 1. Concept of parabolic reflection

reflection board rotates, and calculates the power of the observed signal for each angle of the parabolic reflection board. Then, the direction having maximum power was selected as the sound source direction:

$$\hat{i} = \underset{i}{\operatorname{argmax}} \sum_{n} \sum_{\omega} \log |O_i(\omega; n)|^2. \tag{1}$$

Here, $O(\omega; n)$ is the ω-th frequency bins of short-term linear spectrum at the frame n. i is the angle of the parabolic reflection board (microphone).

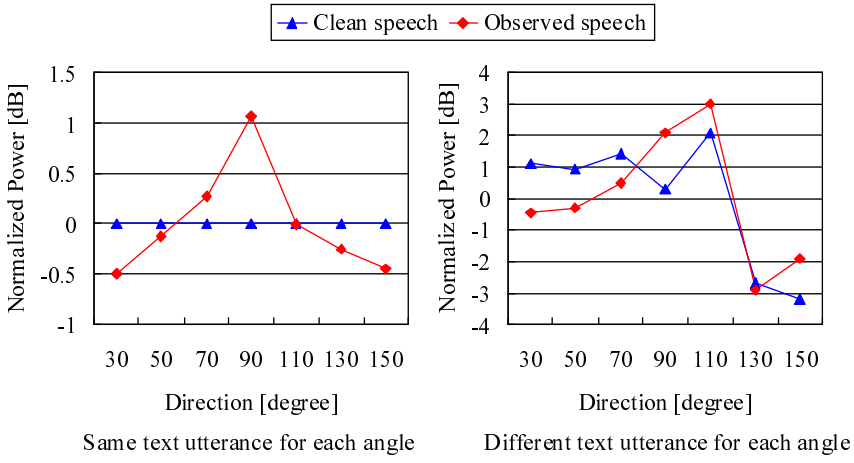

Figure 2. Power of a clean speech segment and the speech segment observed by the microphone with a parabolic reflection board for each angle. The power was normalized so that the mean values of all directions was 0 dB.

Its effectiveness has been confirmed on white noise signals. However, the signal-power-based method was not effective for finding the direction of a talking person. This is because the power of the uttered speech signals varies for all directions of the parabolic reflection board. Figure 2 shows the power of a clean speech segment and the observed speech segment for each angle of the parabolic reflection board. The size of the recording room was about 6.3 m $\times$ 7.2 m (width $\times$ depth). The target sound source was located at 90 degrees and 2 m from the microphone. The diameter of the parabolic reflection board was 24 cm, and the distance to the focal point was 9 cm. The speech signal was sampled at 12 kHz, and windowed with a 32-msec Hamming window every 8 msec. The power was normalized so that the mean values of all directions were 0 dB. In the left portion of Figure 2, the text utterance is the same for each angle of the parabolic reflection board, and in the right portion, the text utterance is different for each angle. As shown in this figure, the power of the observed speech was most enhanced by the parabolic reflection board at 90 degrees (target direction). However, when the utterance text differs at each angle of the parabolic reflection board, the power of observed speech at 90 degrees did not have the maximum power since the power of input speech for another direction had a higher power than that for 90 degrees. For this case, the signal-power-based fails to estimate the direction of the sound source correctly.

In this chapter, in order to estimate the direction of the sound source correctly when the power of the uttered speech signals varies for all direction of the parabolic reflection board, the acoustic transfer function is used instead of the power. Since the acoustic transfer function does not depend on the uttered clean speech, the use of the acoustic transfer function can estimate the direction of the sound source without the influence of the varying power of the uttered speech signals.

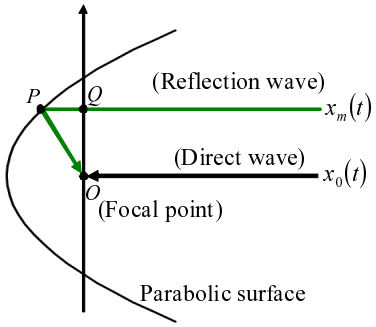

Figure 3. Observed signal at the focal point, where the input signal is coming from directly in front of the parabolic surface

2.3. Signal observed using parabolic reflection

Next, we consider the signal observed using parabolic reflection [20]. As shown in Figure 3, when the sound source is located directly in front of the parabolic surface and there are no background noise and no directivity of the sound source, the observed signal at the focal point at discrete 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):

http://dx.doi.org/10.5772/56045

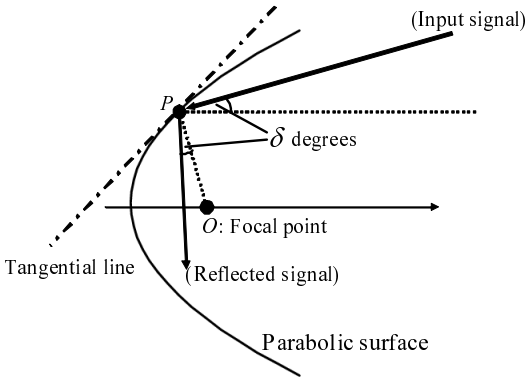

Figure 4. Observed signal at the focal point, where the input signal is coming from δ degrees

$$o(t) = x_p(t) + \sum_{m=1}^{M} x_m(t) \tag{2}$$

where $o(t)$, x_p and x_m ($m = 1, \cdots, M$) are observed sound, direct sound and reflection sound, respectively. Based on the properties of a parabola, the time difference to the focal point between the direct and reflection waves is constant without depending on m. Therefore, (2) can be written as

$$o(t) = s(t) * h_p(t) + \sum_{m=1}^{M} s(t-\tau) * h_m(t) \tag{3}$$

where $s(t)$ and τ are clean speech and the time difference, respectively. h_p is the acoustic transfer function of a direct wave and h_m is that of a reflection wave. By applying the short-term Fourier transform, the observed spectrum at frame n is given by

$$\begin{aligned} &O(\omega; n) \\ &\approx S(\omega; n) \cdot (H_p(\omega; n) + e^{-j2\pi\omega\tau} \cdot \sum_{m=1}^{M} H_m(\omega; n)) \\ &= S(\omega; n) \cdot (H_p(\omega; n) + H_r(\omega; n)). \end{aligned} \tag{4}$$

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, as shown in Figure 4, when the input signal is coming from δ degrees (not coming from directly in front of the parabolic surface), the direction of the reflected signal at the parabolic surface is off δ degrees from PO. Therefore, 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:

$$O(\omega; n) \approx S(\omega; n) \cdot H_p(\omega; n). \tag{5}$$

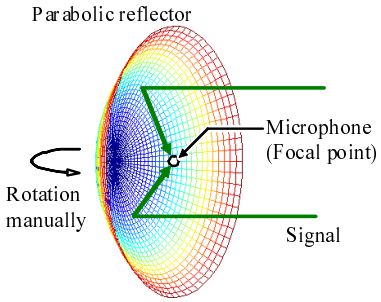

Figure 5. Active microphone with parabolic reflection

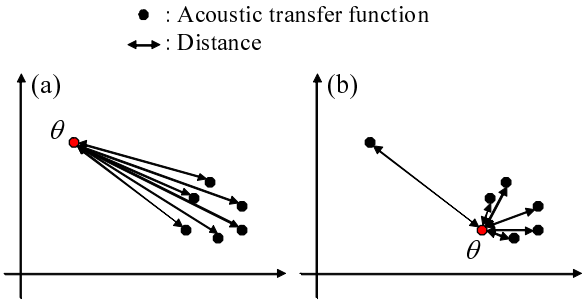

Figure 6. Acoustic transfer function in a feature space for each angle of the active microphone. (a) The case that the direction θ has the acoustic transfer function which is the farthest from those of other directions. (b) The case that the acoustic transfer function of θ is similar to most of the acoustic transfer functions of other directions.

2.4. Estimation of sound source direction

As shown in Figure 5, 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 chapter. Then, from equations (4) and (5), the spectrum of the signal observed at a microphone angle θ can be expressed as

$$\begin{aligned} O_\theta(\omega;n) &\approx S_\theta(\omega;n)\cdot H_\theta(\omega;n) \\ H_\theta(\omega;n) &= \begin{cases} H_p(\omega;n)+H_r(\omega;n) & (\theta=\hat{\theta}) \\ H_p(\omega;n) & (\theta\neq\hat{\theta}) \end{cases} \end{aligned} \tag{6}$$

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 direction. 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 acoustic transfer function is estimated at each discrete direction θ, and the sound source direction can be estimated by selecting the direction whose the acoustic transfer

http://dx.doi.org/10.5772/56045

function is the farthest from the acoustic transfer functions of other directions. The sum of the mutual distances is used to find such a direction. For each discrete direction θ, the Euclidean distances from the acoustic transfer function of θ to those of other directions are measured. As shown in Figure 6, when the direction θ has the acoustic transfer function which is the farthest from those of other directions, the sum of the distances becomes larger than those of other directions. Hence, the sound source direction is estimated by selecting the direction having the maximum sum of the distances from the acoustic transfer function of the direction to those of other directions:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \sum_{\theta'} (\bar{H}_{\theta} - \bar{H}_{\theta'})^2 \tag{7}$$

where θ' is all directions of the microphone except θ, and $\bar{H}$ is the expectation of H in regard to the time frame. 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_{θ} from observed speech signals.

3. Estimation of the acoustic transfer function

In our previous work [18], 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 chapter, 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 reverberant speech signal, $o(t)$, in a room environment is generally considered to be the convolution of clean speech and the acoustic transfer function $o(t) = \sum_{l=0}^{L-1} s(t-l)h(l)$, where $s(t)$, $h(l)$ and L are a clean speech signal, an acoustic transfer function (room impulse response) from the sound source to the microphone, and the length of the acoustic transfer function, respectively.

In recent studies for robust speech recognition and speech dereverberation, the reverberant speech in the STFT (Short-Term Fourier Transform) domain is often modeled so that each frequency bin of the reverberant speech is represented by the convolution of the frame sequences of clean speech and the acoustic transfer function [21, 22].

$$O(\omega; n) = \sum_{l'=0}^{L'-1} S(\omega; n-l') \cdot H(\omega; l') \tag{8}$$

Here $O(\omega;n)$, $S(\omega;n)$, and $H(\omega;n)$ are the ω-th frequency bins of short-term linear spectra of the frame n. L' is the length of the acoustic transfer function in the STFT domain. However, that modeling is complex for estimating the frame sequence of the acoustic transfer function, and it is difficult to deal with the estimated components of the acoustic transfer function for this talker localization task. Therefore, in this chapter, we employ a simpler modeling of the reverberant speech, which is approximately represented as the product of clean speech and the acoustic transfer function.

$$O(\omega;n) \approx S(\omega;n) \cdot H(\omega;n) \tag{9}$$

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:

$$O_{cep}(d;n) \approx S_{cep}(d;n) + H_{cep}(d;n) \tag{10}$$

where d is the cepstral index. 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 (10), if O and S are observed, H can be obtained by

$$H_{cep}(d;n) \approx O_{cep}(d;n) - S_{cep}(d;n). \tag{11}$$

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 (11) is estimated in an ML manner [23] by using the expectation maximization (EM) algorithm, which maximizes the likelihood of the observed speech:

$$\hat{H} = \underset{H}{\operatorname{argmax}} \Pr(O|H, \lambda_S). \tag{12}$$

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.

$$\lambda_S = \{w_k, N(\mu_k^{(S)}, \sigma_k^{(S)^2})\}, \quad \sum_k w_k = 1 \tag{13}$$

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. These parameters are estimated using the EM algorithm using a clean speech database.

http://dx.doi.org/10.5772/56045

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.

$$\begin{aligned} Q(\hat{H}|H) &= E[\log \Pr(O, c|\hat{H}, \lambda_S)|H, \lambda_S] \\ &= \sum_c \frac{\Pr(O, c|H, \lambda_S)}{\Pr(O|H, \lambda_S)} \cdot \log \Pr(O, c|\hat{H}, \lambda_S) \end{aligned} \tag{14}$$

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

$$\Pr(O, c|\hat{H}, \lambda_S) = \prod_n w_{c(n)} \Pr(O(n)|c(n), \hat{H}, \lambda_S) \tag{15}$$

where w is the mixture weight and O_n is the cepstrum at the n-th frame. 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 (15) can be written as

$$\Pr(O, c|\hat{H}, \lambda_S) = \prod_n w_{c(n)} \cdot N(O(n); \mu_k^{(S)} + \hat{H}(n),\ \Sigma_k^{(S)}) \tag{16}$$

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

$$\begin{aligned} Q(\hat{H}|H) = &\sum_k \sum_n \Pr(O(n), c(n) = k|H, \lambda_S) \log w_k \\ &+ \sum_k \sum_n \Pr(O(n), c(n) = k|H, \lambda_S) \\ &\cdot \log N(O(n); \mu_k^{(S)} + \hat{H}(n), \Sigma_k^{(S)}). \end{aligned} \tag{17}$$

Here $\mu_k^{(S)}$ and $\Sigma_k^{(S)}$ 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.

$$\begin{aligned} Q(\hat{H}|H) = -\sum_k \sum_n \gamma_k(n) \sum_{d=1}^{D} \Bigg\{ &\frac{1}{2} \log(2\pi)^D \sigma_{k,d}^{(S)^2} \\ &+ \frac{(O(d; n) - \mu_{k,d}^{(S)} - \hat{H}(d; n))^2}{2\sigma_{k,d}^{(S)^2}} \Bigg\} \end{aligned} \tag{18}$$

$$\gamma_k(n) = \Pr(O(n), k|H, \lambda_S) \tag{19}$$

Here $O(n)$ is the cepstrum at the n-th frame for observed speech data. D is the dimension of the $O(n)$, and $\mu_{k,d}^{(S)}$ and $\sigma_{k,d}^{(S)^2}$ 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

$$\hat{H}(d;n) = \frac{\sum_k \gamma_k(n) \frac{O(d;n) - \mu_{k,d}^{(S)}}{\sigma_{k,d}^{(S)^2}}}{\sum_k \frac{\gamma_k(n)}{\sigma_{k,d}^{(S)^2}}}. \tag{20}$$

Therefore, the frame sequence of the acoustic transfer function $\hat{H}_\theta(d;n)$ is estimated from the signal $O_\theta(d;n)$ observed at direction θ in the cepstral domain using equation (20).

We obtain $\hat{H}_\theta(d;n)$ at a discrete direction. Next, the d-th dimension of the mean vector $\bar{H}_\theta(d)$ is obtained by averaging $\hat{H}_\theta(d;n)$ per frame n.

$$\bar{H}_\theta(d) = \sum_n \hat{H}_\theta(d;n) \tag{21}$$

In a similar way, we obtain the mean vector $\bar{H}_\theta(d)$ at all discrete directions, and the sound source direction is estimated using equation (7) using the cepstral vector $\bar{H}_\theta$. In this chapter, the angle of the parabolic reflection microphone was changed manually from 30 degrees to 150 degrees in increments of 20 degrees.

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 5 was used for the experiments. The diameter was 24 cm, and the distance to the focal point was 9 cm. The microphone located at the focal point was 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 lengths: 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 $\times$ 7.2 m (width $\times$ depth). Figure 7 shows the environment of the experiment.

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. For estimation of the acoustic transfer function from the observed speech

http://dx.doi.org/10.5772/56045

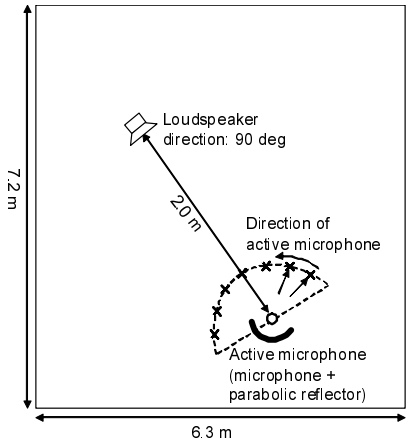

Figure 7. Experimental conditions

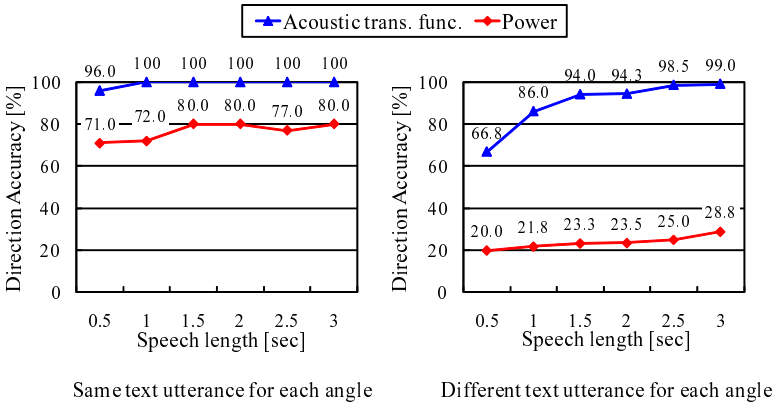

Figure 8. Performance of an active microphone with a parabolic reflection board

signal, 16-order MFCCs (Mel-Frequency Cepstral Coefficients) were used as feature vectors of the clean speech and estimated acoustic transfer function. Then, the 1st and 2nd orders of MFCCs of the estimated acoustic transfer function were used for estimating the sound source direction using equation (7). The test data was spoken by the same female who recorded the training data. The text utterances, however, were different.

4.2. Experiment results

Figure 8 shows the direction accuracy performance using the acoustic transfer function estimated at various speech lengths. The performance is compared to the power-based technique. The left figure shows the accuracy for the same text utterance at each angle of the active microphone, and the right figure shows the accuracy for a different text utterance at each angle of the active microphone. The test data for the same text utterance consisted of 100 segments. The test data for the different utterance consisted of 600, 300, 200, 150,

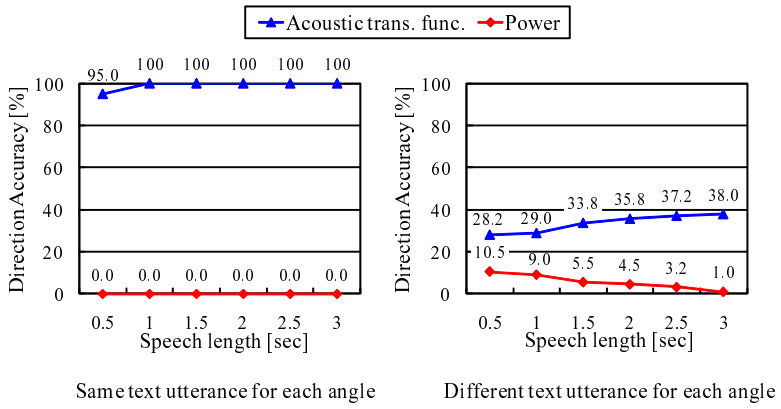

Figure 9. Performance of a shotgun microphone without a parabolic reflection board

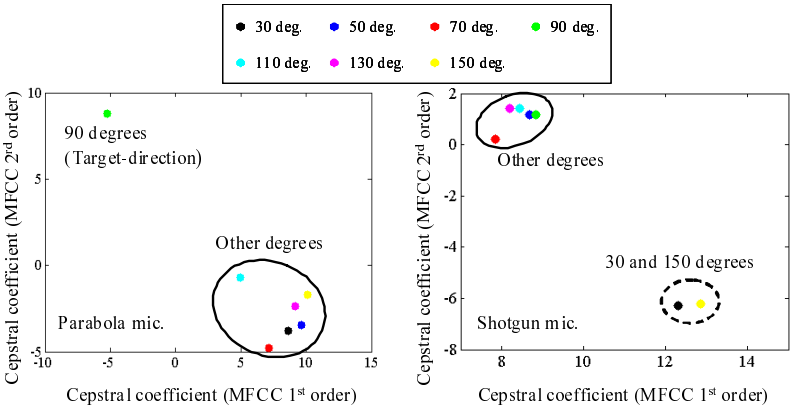

Figure 10. Mean values of the acoustic transfer functions for the microphone with a parabolic reflection board (left) and the shotgun microphone (right)

120, and 100 segments, where one segment has a time length of 0.5, 1.0, 1.5, 2.0, 2.5, and 3.0 seconds, respectively. The test for the same text utterance was conducted 100 times, and that for the different text utterance was conducted 600 times by changing the combination of the text utterances for each direction.

As shown in the left figure, the performance for both the techniques based on the power and the acoustic transfer function is high. However, the possibility of there being an identical text utterance at each angle of the active microphone will be very small in a real environment. In the right portion of Figure 8, 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 different text utterances. This is because the new method uses the information of the acoustic transfer function, which depends on the direction of the

http://dx.doi.org/10.5772/56045

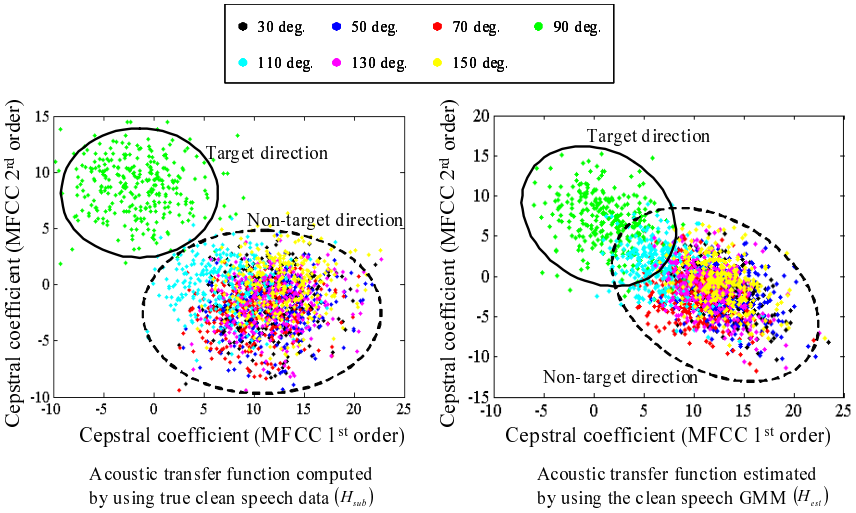

Figure 11. Acoustic transfer function computed by using true clean speech data (left) and that estimated by the proposed method using only the statistics of clean speech GMM (right) at each angle in the cepstral domain

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 reason for this is that the statistics for the observed speech are not readily available if there are not enough samples are used to estimate the acoustic transfer function.

Figure 9 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 comes to be that at 30 or 150 degrees in equation (7). The mean values of all acoustic transfer functions for a parabolic reflection board and the shotgun microphone are plotted in Figure 10, where the acoustic transfer function is computed by (11) using true clean speech signal $S_{cep}(d;n)$ and the total number of frames is 36,600. Then the mean values are computed. As shown in the right potion of Figure 10, 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. As shown in the left potion of Figure 10, on the other hand, the acoustic transfer function that is farthest from all the other acoustic transfer functions is that at 90 degrees to the target direction.

Figure 11 shows the plot of acoustic transfer function for 300 segments of observed speech for the case of the active microphone. In the left potion of Figure 11, the acoustic transfer function H_{sub} was computed by (11) using true clean speech signal $S_{cep}(d;n)$. On the other hand, in the right potion of Figure 11, the acoustic transfer function H_{est} was estimated by

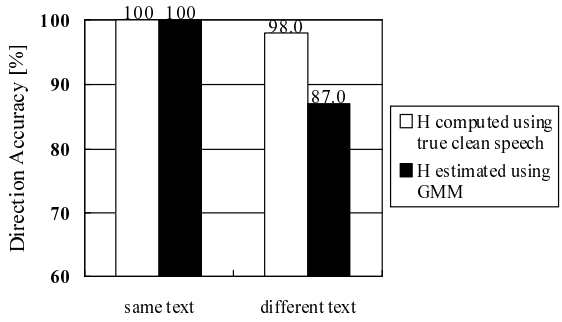

Figure 12. Comparison of true clean speech data and clean speech model

(20) using only the statistics of clean speech GMM. As shown in the left potion of Figure 11, 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 the right potion of Figure 11, though the distribution 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 12 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 chapter 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 with a parabolic reflection board. 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.

It is difficult for this method to estimate the directions of multiple sound sources because it is difficult to estimate the acoustic transfer functions of multiple sound sources. Also, the background noise and the reverberation may cause the measurement error of the acoustic transfer function. We will evaluate the performance of the proposed system in noisy environments and various room environments. In addition, we intend to investigate the performance of the proposed system when the directivity and the orientation of the sound source changes, and to test the performance of the system in a speaker-independent speech model.

http://dx.doi.org/10.5772/56045

Author details

Ryoichi Takashima, Tetsuya Takiguchi and Yasuo Ariki

Graduate School of System Informatics, Kobe University, Kobe, Japan

References

[1] X. Anguera, C. Wooters, and J. Hernando. Acoustic beamforming for speaker diarization of meetings. IEEE Trans. Audio, Speech and Language Processing 2007; 15 2011–2022.

[2] S. Araki, M. Fujimoto, K. Ishizuka, H. Sawada, and S. Makino. A DOA based speaker diarization system for real meetings. proceedings of the Hands-free Speech Communication and Microphone Arrays, HSCMA 2008, 29–32 May 2008, Trento, Italy.

[3] T. Yamagata, A. Sako, T. Takiguchi, and Y. Ariki. System request detection in conversation based on acoustic and speaker alternation features. proceedings of the Interspeech 2007, 2789–2792 August 2007, Antwerp, Belgium.

[4] D. Johnson and D. Dudgeon. Array Signal Processing. New Jersey: Prentice Hall; 1996.

[5] C. Knapp and G. Carter. The generalized correlation method for estimation of time delay. IEEE Trans. on Acoustics, Speech and Signal Processing 1976; 24 320–327.

[6] M. Omologo and P. Svaizer. Acoustic event localization in noisy and reverberant environment using csp analysis. proceedings of the International Conference on Acoustics, Speech, and Signal Processing, ICASSP 1996, 921–924 May 1996, Atlanta, Georgia.

[7] F. Asano, H. Asoh, and T. Matsui. Sound source localization and separation in near field. IEICE Trans. on Fundamentals of Electronics, Communications and Computer Sciences 2000; E83-A 2286–2294.

[8] Y. Denda, T. Nishiura, and Y. Yamashita. Robust talker direction estimation based on weighted CSP analysis and maximum likelihood estimation. IEICE Trans. on Information and Systems 2000; E89-D 1050–1057.

[9] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein. Robust localization in reverberant rooms. In: Microphone Arrays: Signal Processing Techniques and Applications. Berlin, Grmany: Springer; 2001. p157–180.

[10] F. Keyrouz, Y. Naous, and K. Diepold. A new method for binaural 3-D localization based on HRTFs. proceedings of the International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2006, vol.5 341–344 May 2006, Toulouse, France.

[11] M. Takimoto, T. Nishino, and K. Takeda. Estimation of a talker and listener's positions in a car using binaural signals. proceedings of the 4th Joint Meeting ASA and ASJ, ASA/ASJ06, 3216 November 2006, Honolulu, Hawaii.

[12] O. Ichikawa, T. Takiguchi, and M. Nishimura. Sound source localization using a pinna-based profile fitting method. proceedings of the International Workshop on

Acoustic Echo and Noise Control, IWAENC 2003, 263–266 September 2003, Kyoto, Japan.

[13] N. Ono, Y. Zaitsu, T. Nomiyama, A. Kimachi, and S. Ando. Biomimicry sound source localization with fishbone. IEEJ Trans. Sensors and Micromachines 2001; 121-E(6) 313–319.

[14] T. Kristjansson, H. Attias, and J. Hershey. Single microphone source separation using high resolution signal reconstruction. proceedings of the International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2004, 817–820 May 2004, Quebec, Canada.

[15] B. Raj, M. V. S. Shashanka, and P. Smaragdis. Latent dirichlet decomposition for single channel speaker separation. proceedings of the International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2006, vol.5 May 2006, Toulouse, France.

[16] G.-J. Jang, T.-W. Lee, and Y.-H. Oh. A subspace approach to single channel signal separation using maximum likelihood weighting filters. proceedings of the International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2003, 45–48 Apr 2003, Hong Kong, China.

[17] T. Nakatani and B.-H. Juang. Speech dereverberation based on probabilistic models of source and room acoustics. proceedings of the International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2006, I–821–I–824 May 2006, Toulouse, France.

[18] R. Takashima, T. Takiguchi, and Y. Ariki. Single-channel talker localization based on discrimination of acoustic transfer functions. In: P. Strumillo (ed.) Advances in Sound Localization. Rijeka: InTech; 2011. p39–54.

[19] B. Saka and A. Kaderli. Direction of arrival estimation and adaptive nulling in array-fed reflectors. proceedings of the Electrotechnical Conference, 274–277 1998.

[20] T. Takiguchi, R. Takashima, and Y. Ariki. Active microphone with parabolic reflection board for estimation of sound source direction. proceedings of the Hands-free Speech Communication and Microphone Arrays, HSCMA 2008, 65–68 May 2008, Trento, Italy.

[21] A. Sehr, R. Maas, and W. Kellermann. Reverberation model-based decoding in the logmelspec domain for robust distant-talking speech recognition. IEEE Trans. on Audio Speech and Language Processing 2010; 18 1676–1691.

[22] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang. Blind speech dereverberation with multi-channel linear prediction based on short time Fourier transform representation. proceedings of the International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2008, 85–88 March 2008, Las Vegas, Nevada.

[23] B.-H. Juang. Maximum likelihood estimation for mixture multivariate stochastic observations of Markov chains. AT&T Technical Journal 1985; 64 1235–1250.