Monaural sound-source-direction estimation using the acoustic transfer function of a parabolic reflection board

Ryoichi Takashimaa
Graduate School of Engineering, Kobe University, Kobe, 657-8501, Japan

Tetsuya Takiguchi and Yasuo Ariki
Organization of Advanced Science and Technology, Kobe University, Kobe, 657-8501, Japan

(Received 6 June 2009; revised 4 December 2009; accepted 4 December 2009)

This paper presents a sound-source-direction estimation method using only a single microphone with a parabolic reflection board. 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 parabolic reflector and its associated microphone rotate together and observe the speech at each angle. The acoustic transfer function is estimated from the observed speech using the statistics of clean speech signals. Its effectiveness has been confirmed by monaural sound-source-direction estimation experiments in a room environment. © 2010 Acoustical Society of America. [DOI: 10.1121/1.3278603]

PACS number(s): 43.60.Jn [WMC]

I. INTRODUCTION

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., Refs. 1 and 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 Refs. 3 and 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., Refs. 5–8). 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 Gaussian mixture model (GMM) 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.

II. ACTIVE MICROPHONE

A. 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 para-
The parabolic reflector has been used for estimation of the direction of arrival in the radar field. As shown in Fig. 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 deg (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.

**B. Signal observed using parabolic reflection**

Next, we consider the signal observed using parabolic reflection. In Ref. 1, 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 Fig. 2, 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):

\[
o(t) = x_p(t) + \sum_{m=1}^{M} x_m(t),
\]

where \( o(t) \), \( x_p \), and \( x_m \) \((m=1, \ldots, M)\) 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, Eq. (1) can be written as

\[
o(t) = s(t) * h_p(t) + \sum_{m=1}^{M} s(t-\tau) * h_m(t),
\]

where \( s(t) \) and \( \tau \) 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

\[
O(\omega;n) = S(\omega;n) \cdot \left( H_p(\omega;n) + e^{-j2\pi\omega\tau} \sum_{m=1}^{M} H_m(\omega;n) \right)
\]

\[
= S(\omega;n) \cdot (H_p(\omega;n) + H(\omega;n)).
\]

Here \( H_p \) is the acoustic transfer function of the direct sound that is not influenced by parabolic reflection. \( H \) 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:

\[
O(\omega;n) = S(\omega;n) \cdot H_p(\omega;n).
\]

**C. Estimation of sound source direction**

As shown in Fig. 3, 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 Eqs. (3) and (4), the spectrum of the signal observed at a microphone angle \( \theta \) can be expressed as

\[
O_\theta(\omega;n) = S_\theta(\omega;n) \cdot H_p(\omega;n)
\]
\[ H_d(\omega; n) = \begin{cases} H_p(\omega; n) + H_r(\omega; n) & (\theta = \theta') \\ H_p(\omega; n) & (\theta \neq \theta') \end{cases} \] (5)

where \( S_d \) and \( H_d \) are spectra of clean speech and acoustic transfer function at the angle \( \theta \), and \( \tilde{\theta} \) is the sound source direction. Assuming that \( H_p \) is nearly constant for each angle, when the active microphone does not face the sound source, the value of \( H_d \) will be almost the same for every non-target direction. On the other hand, the only condition under which \( H_d \) 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 \( \theta \), and the sound source direction can be estimated by selecting the direction whose acoustic transfer function is the farthest from the acoustic transfer functions of other directions:

\[ \hat{\theta} = \arg \max_{\theta'} \sum_{\theta} (\tilde{H}_d - H_{\theta})^2, \] (6)

where \( \theta' \) is all directions of the microphone except \( \theta \), and \( \tilde{H} \) 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 Sec. III, we will describe how to estimate \( H \) from observed speech signals.

### III. 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.

#### A. 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:

\[ O_{\text{cep}}(d; n) = S_{\text{cep}}(d; n) + H_{\text{cep}}(d; n), \] (7)

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

\[ H_{\text{cep}}(d; n) = O_{\text{cep}}(d; n) - S_{\text{cep}}(d; n). \] (8)

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

#### B. Maximum-likelihood-based parameter estimation

The sequence of the acoustic transfer function in Eq. (8) is estimated in a ML manner\(^{12}\) by using the expectation maximization (EM) algorithm, which maximizes the likelihood of the observed speech:

\[ \hat{H} = \arg \max_H \Pr(O|H, \lambda_S). \] (9)

Here, \( \lambda \) 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.

\[ \lambda_S = \{ w_k, N(\mu_k, \sigma_k^2) \}, \sum_k w_k = 1, \] (10)

where \( w_k \), \( \mu_k \), and \( \sigma_k^2 \) 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.

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:

\[ Q(\hat{H}|H) = E[\log \Pr(O,c|\hat{H},\lambda_S)|H,\lambda_S] \]

\[ = \sum_c \frac{\Pr(O,c|H,\lambda_S)}{\Pr(O|H,\lambda_S)} \cdot \log \Pr(O,c|\hat{H},\lambda_S). \] (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

\[ \Pr(O,c|\hat{H},\lambda_S) = \prod_{n=0}^n w_{n,c} \cdot \Pr(O_n|\hat{H},\lambda_S), \] (12)

where \( w \) is the mixture weight and \( O_n,c \) 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 $\lambda_o$ is derived by adding the acoustic transfer function. Therefore, Eq. (12) can be written as

$$\Pr(O,c|\hat{H},\lambda_S) = \prod_{n^{(v)}} w_{c,n^{(v)}} \cdot N(O_{n^{(v)}},\mu^{(S)}_{n^{(v)}} + \hat{H}_{n^{(v)}},\Sigma^{(S)}_{n^{(v)}}),$$

(13)

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

$$Q(\hat{H}|H) = \sum_{k} \sum_{n^{(v)}} \Pr(O_{n^{(v)}},c_{n^{(v)}} = k|\lambda_S) \log w_k$$

$$+ \sum_{k} \sum_{n^{(v)}} \Pr(O_{n^{(v)}},c_{n^{(v)}} = k|\lambda_S) \log N(O_{n^{(v)}},\mu^{(S)}_k)$$

$$+ \hat{H}_{n^{(v)}},\Sigma^{(S)}_{n^{(v)}}.$$  

(14)

Here $\mu^{(S)}_k$ and $\Sigma^{(S)}_{n^{(v)}}$ 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 $H$.

$$Q(\hat{H}|H) = -\sum_{k} \sum_{n} \gamma_k(n) \sum_{d=1}^{D} \left\{ \frac{1}{2} \log(2\pi)^d \sigma^{(S)}_{k,d}^2$$

$$+ \frac{(O(d;n) - \mu^{(S)}_k - \hat{H}(d;n))^2}{2\sigma^{(S)}_{k,d}^2} \right\},$$

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

(15)

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

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

(17)

Therefore, the frame sequence of the acoustic transfer function $\hat{H}(d;n)$ is estimated from the signal $O(d;n)$ observed at direction $\theta$ in the cepstral domain using Eq. (17).

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

$$\bar{H}(d) = \sum_n \hat{H}(d;n).$$

(18)

In a similar way, we obtain the mean vector $\bar{H}(d)$ at all discrete directions, and the sound source direction is estimated using Eq. (6) using the cepstral vector $H_\theta$. In this paper, the angle of the parabolic reflection microphone was changed manually from 30 to 150 deg in increments of 20 deg.

IV. EXPERIMENT

A. Experiment conditions

The direction estimation experiment was carried out in a real room environment. The parabolic reflection microphone shown in Fig. 3 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 deg and 2 m from the microphone. The angle of the parabolic reflection microphone was changed manually from 30 to 150 deg in increments of 20 deg. 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 s. The size of the recording room was about 6.3 m $\times$ 7.2 m (width $\times$ depth). Figure 4 shows the environment of the experiment.

The speech signal was sampled at 12 kHz and windowed with a 32-ms Hamming window every 8 ms. 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 second-order mel-frequency cepstral coefficients were used as feature vectors. The test data were spoken by the same female who recorded the training data. The test utterances, however, were different.

B. Experiment results

Figure 5 shows the direction accuracy performance using the acoustic transfer function estimated at various speech lengths, 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. The test data for the same text utterance consisted of 100 segments. The test
same text utterance for each angle (active mic. with a parabolic reflection board)

Different text utterance for each angle (active mic. with a parabolic reflection board)

FIG. 5. Performance of an active microphone with a parabolic reflection board.

data for the different utterance consisted of 600, 300, 200, 150, 120, and 100 segments, where one segment has time lengths of 0.5, 1.0, 1.5, 2.0, 2.5, and 3.0 s, 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 top 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 bottom portion of Fig. 5, 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. Figure 6 shows the power of a clean speech segment that did not estimate the direction of arrival correctly and the observed speech segment, normalized so that the mean value of all directions was 0 dB.

FIG. 6. Power of a clean speech segment that did not estimate the direction of arrival correctly and the speech segment observed by the active microphone with a parabolic reflection board for each angle. The power was normalized so that the mean value of all directions was 0 dB.

microphone, the power of observed speech at 90 deg did not have the maximum power since the power of input speech for another direction had a higher power than that for 90 deg. On the other hand, the performance of the new method based on the acoustic transfer function was 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 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 used to estimate the acoustic transfer function.

Figure 7 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 deg in Eq. (6). The mean values of all acoustic transfer functions for the shotgun microphone and a parabolic reflection board are plotted in Figs. 8 and 9, respectively, where the acoustic transfer function is computed by Eq. (8) using true clean speech signal $S_{\text{true}}(d;n)$ and the total number of frames is 36 600. Then the mean values are computed. As shown in
FIG. 7. Performance of a shotgun microphone without a parabolic reflection board.

Fig. 8, we can see that the acoustic transfer function that is farthest from all the other acoustic transfer functions is that at 30 or 150 deg. As shown in Fig. 9, on the other hand, the acoustic transfer function that is farthest from all the other acoustic transfer functions is that at 90 deg to the target direction.

Figures 10 and 11 show the plot of acoustic transfer function for 300 segments of observed speech for the case of the active microphone. In Fig. 10, the acoustic transfer function $H_{sub}$ was computed by Eq. (8) using true clean speech signal $S_{cep}(d,n)$. On the other hand, in Fig. 11, the acoustic transfer function $H_{est}$ was estimated by Eq. (17) using only the statistics of clean speech GMM. As shown in Fig. 10, 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 Fig. 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 to 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 $H_{\text{est}}$ decreased. This is because the value of $H_{\text{est}}$ was influenced to some extent by the phoneme sequence of clean speech.

V. 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 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. 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.