HEAD ORIENTATION ESTIMATION OF A SPEAKER BY UTILIZING KURTOSIS OF A DOA HISTOGRAM WITH RESTORATION OF DISTANCE EFFECT

Masahito Togami and Yohei Kawaguchi

Central Research Laboratory, Hitachi Ltd.
1-280, Higashi-koigakubo Kokubunji-shi, Tokyo 185-8601, Japan

ABSTRACT

In this paper, we propose a head-orientation estimation method from multichannel acoustic signals. Sharpness of a DOA histogram which is extracted by using the sparseness based DOA estimation method varies depending on the head orientation of a speaker. The proposed method utilizes this phenomenon to estimate the head orientation of the speaker. The proposed method uses more than two microphone arrays. In addition to estimation of the speaker location, the proposed method estimates kurtosis of the DOA histogram of each array. Kurtosis is regarded as a measure of sharpness of a DOA histogram in the proposed method. However, kurtosis also depends on the distance between the speaker and the microphone array (distance effect). The distance effect is experimentally revealed by the regression analysis. The head orientation of a speaker is estimated by the restored kurtosis which is free from the distance effect. Experimental results on a reverberant environment show that the proposed method can estimate the head orientation of a speaker more accurately than a conventional head-orientation estimation method.

Index Terms— head-orientation estimation, sparseness, DOA histogram, kurtosis

1. INTRODUCTION

Many researches have been done for estimation of a sound-source location with multiple microphones [1][2][3][4]. Conventionally, SRP-PHAT is frequently used for a broadband signal such as human speech. Recently, in addition to sound-source localization, head-orientation estimation methods have been also studied [5][6][7][8]. Head-orientation estimation is also important for communication robots. By using head-orientation estimation, a communication robot can know whether a speaker utter toward it or not. Conventional head-orientation estimation approaches assume that a radiation pattern of human speech is not constant around a speaker head. Some conventional methods [5] estimate the head orientation of the speaker by measuring the radiation pattern directly. However, the sensitivity of each microphone is set to be configured precisely to measure the radiation pattern correctly. In practical applications, the configuration of each microphone is not always possible. Instead of the radiation pattern estimation, conventional methods which use a shape of the evaluation function of SRP-PHAT have been studied [7][8]. OGCF is one of the head-orientation estimation methods based on SRP-PHAT. OGCF uses a modified evaluation function of SRP-PHAT. In the evaluation function of OGCF, an evaluation function of each microphone array is combined with a weight function depending on the virtual head orientation of the speaker. When the virtual head orientation is close to a microphone-array direction, the evaluation function of this array at the estimated sound source location is big-valued, the virtual head orientation which gives the maximum evaluation function at the estimated sound source location is selected as the estimate of the head orientation of the speaker. In OGCF, there are two problems for precise head-orientation estimation. One problem is that the effect of the head orientation to the shape of the evaluation function of SRP-PHAT is not so big. Therefore, a high-resolution head-orientation estimation is required. Another problem is that in OGCF, the evaluation function of SRP-PHAT is not restored depending on the distance between a sound source and a microphone array. By the inverse-square-law attenuation of the sound-source power, the peak of a evaluation function of SRP-PHAT with a microphone array which is close to the sound source is sharper than one with a microphone array which is far from the sound source (distance effect). Therefore, restoration for the distance effect is required. To overcome the above two problems, in this paper, we propose a novel head-orientation estimation method by using kurtosis of a Direction-of-Arrival (DOA) histogram with restoration of the distance effect. Previously, one of the authors proposed DOA histogram based sound source localization methods [3][4]. In these methods, sound sources are assumed to be sufficiently sparse, and there is only one source at each time-frequency point. A DOA histogram is made from the DOA estimation result at each time-frequency point, and the peak of the DOA histogram is extracted as the final DOA estimation result. We focus on that sharpness of the DOA histogram is depending on the head orientation of the speaker. In the proposed method, kurtosis is used as a measure of sharpness of a DOA histogram. Kurtosis is also affected by the distance effect. The proposed method cancels the distance effect by utilizing the estimated source location.

2. PROBLEM STATEMENT

In this paper, multiple microphones are divided into multiple sub-microphone arrays. The multichannel-input signal in the $p$-th sub-microphone array is converted into the time-frequency domain by the short-term-Fourier transform. The multichannel-input signal, $x_p(f, \tau)$, is modelled as follows:

$$x_p(f, \tau) = s(f, \tau) a_p(f) + N_p(f, \tau),$$

where $f$ is the frequency index, $\tau$ is the frame index, $a_p(f)$ is the transfer function between the sound source location and the $p$-th sub-microphone array (steering vector), $s(f, \tau)$ is a sound source signal which is assumed to be a speech signal, $N_p(f, \tau)$ is the background noise component, which is uncorrelated with the sound source signal. $a_p(f)$ contains information about the sound-source location and the sound-source orientation. In this paper, a microphone array is on the horizontal space, and the sound source localization on the horizontal space is the problem that is treated in this paper.
2.1. Sparseness based DOA estimation method

A speech signal is known to be a sparse signal in the time-frequency domain. A sparseness based DOA estimation method assumes that there is only one source at each time-frequency point, and the DOA of the \( p \)-th microphone array at each time-frequency point, \( \Theta_p(f, \tau) \), is estimated as follows [4]:

\[
\Theta_p(f, \tau) = \arg\max_{\Theta \in \Omega_\Theta} |a_p, \omega(f)^H x_p(f, \tau)|^2 ,
\]

where \( \Omega_\Theta \) is a set of discrete azimuths from \(-180\) to \(180\) degrees, \( a_p, \omega(f) \) is a virtual steering vector which is calculated by the microphone-array alignment, \( a_p, \omega(f) \) is simulated by the propagation theory from the sound source whose azimuth is \( \Theta \) to the microphone array, and the \( l_1 \)-norm of \( a_p, \omega(f) \) is normalized to be one without loss of generality. The azimuth range from \(-180\) to \(180\) degrees is unequal partitioning. Azimuth difference between arbitrary two adjacent points is proportional to \( \frac{1}{2 \sin(\text{azimuth})} \). Therefore, partitioning is performed so as the distance of the virtual steering vectors between adjacent two points to be equal. When the DOA of the speech source at frame \( \tau \) and frequency \( f \) is \( \Theta \), and \( N(f, \tau) = 0 \), in this case, \( |a_p, \omega(f)^H x_p(f, \tau)| \) is maximized at \( \Theta = \Theta \). Therefore, the DOA of the speech source at \( (f, \tau) \) is correctly estimated by Eq. 2. However, actually, \( N(f, \tau) \neq 0 \), so the estimated DOA is not always the speech-source direction. Therefore, the conventional method calculates a DOA histogram at multiple time-frequency points, and the maximum peak of the DOA histogram is regarded as a speech-source direction.

2.2. Sparseness based sound source localization

By using the DOA estimation result of each microphone array, the sound-source location is estimated by taking a cross point of the DOA estimation results of arbitrary two microphone array. The sound-source location is estimated by peak-searching of the \( x-y \) histogram. The \( x-y \) histogram is composed of sub-\( x-y \) histograms which are made from the DOA estimation results of arbitrary two microphone arrays as follows:

\[
I(x, y) = \sum_{q=1}^{Q} I_q(x, y),
\]

where \( I(x, y) \) is the \( x-y \) histogram, \( Q \) is the number of microphone-array pairs, \( I_q(x, y) \) is the \( q \)-th \( x-y \) histogram. The \( q \)-th microphone-array pair is composed of the \( q_1 \)-th microphone array and the \( q_2 \)-th microphone array. \( I_q(x, y) \) is obtained by conversion of two-dimensional azimuth histogram, \( I_{\text{azimuth}, q}(\Theta_{q_1}, \Theta_{q_2}) \) as follows.

\[
I_q(x, y) = \sum_{i=0}^{1} \sum_{j=0}^{1} w(i, j) I_{\text{azimuth}, q}(\Theta_{\text{NEAR}(q_1, x, y)} + i, \Theta_{\text{NEAR}(q_2, x, y)} + j),
\]

where \( \text{NEAR}(q_1, x, y) \) is the index of the direction which is smaller than \( \Theta(q_1, x, y) \) and is nearest to \( \Theta(q_1, x, y) \), \( \Theta(q_1, x, y) \) is the direction of \( (x, y) \) at the center point of the \( q_1 \)-th microphone array, and \( \text{NEAR}(q_2, x, y) \) is similarly defined. \( \Theta \) in \( \Omega_\Theta \) is sorted in ascending order. Therefore, the sound source localization problem is how to achieve \( I_{\text{azimuth}, q}(\Theta_{q_1}, \Theta_{q_2}) \). In each time-frequency point, If the DOA of the \( q_1 \)-th microphone array is \( \Theta_{q_1} \) and the DOA of the \( q_2 \)-th microphone array is \( \Theta_{q_2} \), then \( I_{\text{azimuth}, q}(\Theta_{q_1}, \Theta_{q_2}) \) is incremented by one.

3. PROPOSED METHOD

3.1. Head-orientation effect to DOA histograms

To achieve the sound-source location and the head orientation of a speaker, sharpness of a DOA histogram is utilized. In Fig. 1, DOA histograms of various head orientations in the same source position are depicted. Four microphone arrays are used to make the DOA histograms. These arrays are the same arrays used in the experiments in this paper. Sharpness of a DOA histogram obviously depends on the head orientation of the speaker. Affected by the radiation pattern of human speech, the direct sound to a microphone array is relatively small-valued, when the head orientation of the speaker is not in front of the microphone array, the ratio between \( |s(f, \tau) a_p(f)| \) and \( |N_p(f, \tau)| \) in Eq. 1 (SNR) is relatively small-valued. The smaller SNR is, the bigger the variance of the DOA estimation result becomes, and sharpness of the DOA histogram becomes flat. To measure sharpness of the DOA histogram, in the proposed method, kurtosis of the DOA histogram is used. Kurtosis of the DOA histogram, \( K \), is defined as follows:

\[
K = \left( \frac{\sum_{i=\text{peak}-1}^{\text{peak}+1} H(i) |i - \text{peak}|^4}{V} \right)^{1/2},
\]

where \( V = \sum_{i=\text{peak}-1}^{\text{peak}+1} H(i) |i - \text{peak}|^2 \), \( H(i) \) is a DOA histogram, \( i \) is the index which indicates a particular azimuth in \( \Omega_\Theta \), \( \text{peak} = \arg\max_{\Omega_\Theta} H(i), \) and \( V \) is the window size for kurtosis calculation. \( \hat{H}(i) \) is normalized as \( \sum_{i=\text{peak}-1}^{\text{peak}+1} \hat{H}(i) = 1 \).

3.2. Restoration of kurtosis

As discussed previously, kurtosis has information about the head orientation of the speaker. However, it is not recommended that kurtosis is used directly for the head orientation of the speaker, because kurtosis also depends on the distance between the microphone array and the sound source. Kurtosis at near-distance is bigger than that at far-distance (distance effect). To normalize the distance effect, the proposed method restores kurtosis by using the estimated
sound-source location. At first, kurtosis is defined as the function of the sound-source location and restored kurtosis at the basis distance as \( K = f(r), \bar{K} = f(r_{\text{basis}}) \), where \( K \) is the restored kurtosis, \( r \) is the distance between the microphone array and the estimated sound source location, \( r_{\text{basis}} \) is the basis distance, \( f(r) \) is a function for normalization of kurtosis. From the inverse-square-law attenuation of the sound source, \( f(r) \) is considered to be a polynomial of \( \frac{r}{r_{\text{basis}}} \). In this paper, we limit \( f(r) \) to \( f(r) = \frac{r}{r_{\text{basis}}} + \beta \) or \( f(r) = \frac{r^2}{r_{\text{basis}}} + \beta \). When \( r \) is infinity, the DOA histogram becomes completely flat, and kurtosis becomes 1.8. Therefore, \( \beta \) is set to be 1.8. To decide \( f(r) \), we measured kurtosis at multiple distances under a reverberant environment. Measurement is performed with four microphone arrays simultaneously, and data of two person are measured. A regression analysis is used to decide \( f(r) \), and \( f(r) \) which minimizes the regression error was \( f(r) = \frac{r}{r_{\text{basis}}} + \beta \) for each person. In Fig. 2, the regression error is depicted for both functions with four distances \( \alpha \) is determined from the observed kurtosis. Therefore, the restored kurtosis \( \bar{K} \), is calculated from the observed kurtosis \( K \) as follows:

\[
\bar{K} = \frac{(K - \beta)r_{\text{basis}} + \beta}{r_{\text{basis}}}.
\] (6)

### 3.3. Head-orientation estimation

From the restored kurtosis, the head orientation of the speaker is determined as the following equation which is similar to OGCF [7]:

\[
\theta_{\text{orient}} = \arg\max_{\theta} \sum_{p=1}^{P} \frac{w(x)}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} - \Delta(p, \langle x, y \rangle_{\text{est}}) \bar{K}_p,
\] (7)

where \( \bar{K}_p \) is the restored kurtosis of the \( p \)-th microphone array, \( \langle x, y \rangle_{\text{est}} \) is the estimated sound source location, \( \Delta(p, \langle x, y \rangle_{\text{est}}) \) is the angle of the vector from the \( p \)-th microphone array location to \( \langle x, y \rangle_{\text{est}} \), \( w(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} \) is defined by the same way implemented in OGCF [7], \( w(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} \), and \( \sigma \) is a parameter related to the radiation pattern of human speech.

### 3.4. Summary of proposed method

The flow of the proposed method is shown in this subsection.

**Initialize**

For each array the DOA histogram \( H_p(i) \) is set to be 0, and for arbitrary pair, \( I_{\text{simath},p}(i, j) \) is set to be 0.

**Processes at each time-frequency point**

The flow at each time-frequency point is shown as follows:

1. DOA of each array, \( \Theta_p(f, \tau) \), is estimated by Eq. 2.

2. For arbitrary pair, \( I_{\text{simath},q}(\Theta_{q1}(f, \tau), \Theta_{q2}(f, \tau)) \) is incremented by one.

3. For each pair, \( H(\Theta_q(f, \tau)) \) is incremented by one.

**Estimation of sound-source location and head orientation**

1. From the estimated two dimensional histogram of arbitrary pair, the \( x - y \) histogram \( J(x, y) \), is obtained by Eq. 3 and Eq. 4. By peak-searching of \( I(x, y) \), the sound-source location of the speaker is estimated as \( \langle x, y \rangle_{\text{est}} \).

2. From the estimated one dimensional histogram of each pair, \( H_p(i) \), kurtosis around the peak of this histogram is calculated as \( K_p \) by Eq. 5.

3. The estimated kurtosis of each pair is restored by using the estimated sound-source location \( \langle x, y \rangle_{\text{est}} \) and Eq. 6.

4. The head orientation of the speaker, \( \theta_{\text{orient}} \), is estimated by the restored kurtosis \( \bar{K}_p \) and Eq. 7.

### 4. EXPERIMENT

The head-orientation estimation performance of the proposed method is evaluated under a reverberant environment. The experimental environment is shown in Fig. 3. The reverberation time was about 300 ms. Evaluation is done with two male speakers. Each speaker stood up at thirteen points. The length of each utterance was about three seconds. There were ambient noise. At each location point, each speaker uttered speech toward each microphone array and each corner which is \( (x, y) = (\pm 1.5, 1.5), (\pm 1.5, 0), (0, \pm 1.5), (1.5, 1.5), (-1.5, 1.5), (-1.5, 0), (0, -1.5), (-1.5, -1.5) \). Therefore, the number of the head orientation patterns were 8 at each point. Therefore, there were total 208 utterances. Four microphone arrays (m1-m4) were used in this experiment. Each microphone array was composed of 6 microphones. The microphone alignment of each microphone array is shown in Fig. 4. The proposed method was compared with OGCF.

5. CONCLUSION

In this paper, we propose a novel head-orientation estimation method by using kurtosis of a DOA histogram with restoration of the distance effect. We focus on that sharpness of a DOA histogram is depending on the head orientation of a speaker. The proposed method estimates the head orientation by kurtosis of a DOA histogram of each microphone array. Furthermore, because kurtosis depends on the distance between a microphone array and a sound source, the proposed method restored kurtosis by using the estimated sound-source location. Experimental results in a real reverberant room show that the proposed method can estimate the head orientation of a speaker more accurately than a conventional method.

6. REFERENCES


