Improvement in estimation accuracy of a sound source direction by a frequency domain binaural model with information on listener’s head movement in a conversation

Yoshifumi Chisaki* and Shogo Tanaka†
*Graduate school of Science and Technology, Kumamoto University, Kumamoto, Japan
E-mail: chisaki@cs.kumamoto-u.ac.jp Tel: +81-96-342-3621
†Graduate school of Science and Technology, Kumamoto University, Kumamoto, Japan

Abstract—Estimation of sound source directions and separation of sound sources are implemented on many products widely, and binaural hearing aids are one of its applications. In a conversation using a binaural hearing aid, continuous tracking of sound sources with acoustics signals are sometimes complicated because sound sources move dynamically. In order to make the tracking of sound sources simple, it is considered to be helpful to use non-verbal information in communication. Since a user’s body movement, including a head, corresponds to speakers’ positions, it is possible to estimate communication zone where sound sources locate by a head direction. In this paper, a head movement in a conversation, as non-verbal information in communication, is measured and discussed. A rotational angle of a head movement is estimated by an inertial sensor which has an accelerometer, angular velocity sensor, and a magnetic field sensor with attaching to left ear position. A relationship between a head movement and directions of sound sources during a conversation by multiple speakers is presented. The results indicate a possibility for estimating a communication zone.

I. INTRODUCTION

A binaural signal processing has highly potential to be adopted to many applications, such as binaural hearing aids. An expected function by introducing a binaural model into hearing aids is signal enhancement. The frequency domain binaural model (FDBM) as one of binaural models can estimate a sound source direction and enhance sound source signal based on the estimated direction of sound source[1]. The frequency domain binaural model can also estimate multiple sound sources’ directions simultaneously[2]. However, the method sometimes estimates wrong direction when multiple sound sources exist. This error causes temporal discontinuity of enhanced speech signal because a process is performed frame by frame. Although spatial area for enhancement can be expected to estimate using temporal information based on the estimated sound sources directions, it makes the algorithm complex. In order to improve a performance on estimation of sound source directions, a speaker’s head direction as one of non-verbal information would be useful information.

A spatial area corresponds to a zone created by listeners and speakers in a conversation and a head movement have a crucial relationship in a communication. The head movement is also concerned with a signal processing based on a head related transfer function, e.g. head tracking plays important role in virtual audio display[3]. Since accelerometer and angular velocity sensor become compact with highly-developed technology, a motion sensor can be buried in an ear canal hearing aid itself. Thus, taking into account of head movement estimation has potential to assist to estimate spatial area for signal enhancement.

In the previous paper, zone detection methods are discussed[4]. In this paper, a head movement in a conversation is measured, and analyzed for detection of a communication zone for binaural hearing aids. A head movement is detected by an inertial sensor consists of an angular velocity sensor, an accelerometer and a magnetic field sensor, which are attached to one side of ear positions. Experiments in a conversation is performed with several speakers. According to the results, a detected direction of a speaker’s face does not always focus on listeners’ directions, however, it does not exceed the left and right edges of sound sources’ positions in a horizontal plane.

II. CONVERSATION OF GYRO DATA INTO ROTATIONAL ANGLE

The directions of arrival can be estimated simultaneously by Frequency Domain Binaural Model (FDBM) [1][2]. In order to improve the performance on the directions estimation, the estimation of spatial area including A and B is useful as shown...
in Fig. 1. This section describes a detection method of head movement.

In this paper, an inertial sensor is attached to left ear position. The inertial sensor includes 3-axis accelerometer, angular velocity sensor and magnetic field sensor. In order to make estimation process of spatial area simple, a head movement and a sound source are mapped onto a horizontal plane. The conversion of the sensor output into a rotational angle on a plane is mentioned here. The relationship between axes for sensors and a rotational angle on a plane is mentioned in section II-A. In section II-B, mapping sensors data onto a horizontal plane where sound sources locate is described.

A. Sensor

Inertial sensor TSND121 manufactured by ATR-Promotions is attached to left ear position as shown in Fig. 2. The values $A_x$, $A_y$, $A_z$ by 3-axis accelerometer and $M_x$, $M_y$, $M_z$ by 3-axis magnetic field sensor are obtained in a coordinate system as shown in Fig. 3.

B. Mapping on a rotational angle plane

Fig. 4 shows a relationship of coordinate systems. The upper figure is a head-based coordinate system, and the lower one is a gravity-based coordinate system.

When XYZ axes are rotated with $\alpha^\circ$, $\beta^\circ$, $\gamma^\circ$ against $X'Y'Z'$ axes, $\alpha^\circ$ and $\beta^\circ$ can be obtained by

$$\alpha = \arctan(A_y/A_x)$$

$$\beta = \arctan(A_z/\sqrt{A_y^2 + A_x^2}).$$

Values $M'_x$, $M'_y$, and $M'_z$ on a gravity-based $X'Y'$ plane can be obtained by

$$\begin{bmatrix}
M'_x \\
M'_y \\
M'_z
\end{bmatrix} =
\begin{bmatrix}
\cos \beta & 0 & \sin \beta \\
\sin \alpha \sin \beta & \cos \alpha & -\sin \alpha \cos \beta \\
\cos \alpha \sin \beta & \sin \alpha & \cos \alpha \cos \beta
\end{bmatrix}
\begin{bmatrix}
M_x \\
M_y \\
M_z
\end{bmatrix}.$$

Finally, rotation angle $\gamma$ on a gravity-based $X'Y'$ plane is

$$\gamma = \arctan(M'_y/M'_x).$$

III. RELATIONSHIP BETWEEN A DIRECTION OF HEAD AND THAT OF SOUND SOURCE

A detected angle with a sensor does not match a direction of sound source because rotation angle range of neck is limited and eyes can follow a sound source. In order to confirm the difference between a direction of sound source and that of head, preliminary experiment is performed. A target is set to each angle as shown in Fig. 5, and five male subjects are instructed to face to each direction from $0^\circ$ 50 times. Subjects are also requested not to move the lower part of their body and to keep their head on a gravity-based $X'Y'$ plane as possible as they can.

Results of a preliminary experiment are shown in Table I. These results show that a detected angle with a sensor does not match a target direction. In order to adjust the differences,
TABLE I
THE MEDIAN OF HEAD ROTATIONAL ANGLE AGAINST EACH TARGET ANGLE.

<table>
<thead>
<tr>
<th>Target</th>
<th>-80°</th>
<th>-60°</th>
<th>-40°</th>
<th>-20°</th>
<th>20°</th>
<th>40°</th>
<th>60°</th>
<th>80°</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median</td>
<td>-54.4</td>
<td>-43.7</td>
<td>-28.0</td>
<td>-12.6</td>
<td>9.87</td>
<td>25.1</td>
<td>37.5</td>
<td>49.5</td>
</tr>
</tbody>
</table>

(a) Two other talkers at a box table.

(b) Three other talkers at a box table.

(c) Four other talkers at a round table.

Fig. 7. The experiment arrangement in case of two, three or four other talkers.

a regression line,
\[ \hat{\gamma} = 0.56\gamma, \]
is introduced based on the results. The coefficient of determination \( R^2 \) is 0.99.

IV. UNSTABLENESS OF A HEAD

Since human’s motion always has unstableness even if human intend to fix a head, a target angle cannot be evaluated strictly. In this section, a fluctuation of a head is measured as a preliminary experiment as shown in Fig. 6. A subject is instructed to watch a front target, and a state of a head is recorded into video. According to the results by video analysis, a head moves ±15° in rotational angle. The results are taken into account for evaluation.

V. ZONE ESTIMATION

Communication zone consists of multiple small zones. A horizontal plane are divided into a small zone with a certain angle range which labeled as A-N. Let us assume left and right boundary angles, \( \theta_{Z_l} \) and \( \theta_{Z_r} \), which corresponds to an arbitrary zone \( Z \). The estimated zone \( S(t) \) is obtained by

\[
S(t) = \{ Z | \theta_{Z_l} < \hat{\gamma}(t) < \theta_{Z_r} \} \\
Z \in \{ A, B, ... , N \} \]

VI. EXPERIMENTS

In this section, a head movement in a conversation at a box or a round table are examined.

A. Experimental condition

Fig. 7(a), (b) and (c) show arrangement of other talkers in case of two, three and four, respectively. A subject and other speakers are instructed to keep a conversation three minutes and concentrate on the conversation. The number of subjects is five, and those are male.

The spatial area on a gravity-based \( X'Y' \) plane are divided into zone labeled as A-N. A solid line in figures shows a shape of table.
B. Analysis on a head direction against other talkers

Fig. 8 shows results of a head tracking when the number of other speakers are two as shown in Fig. 7(a). Left and right vertical axes are an estimated zone and the number of samples which estimated in the zone, namely frequency, respectively. Fig. 8(a)-(e) are the results for each subject. A solid line indicates an estimated head direction, and a box with a dotted line shows utterance section. Thus, a solid line overlaps a box means that a face focuses on a person who is speaking. Fig. 7(b) and 7(c) show arrangement of speakers in case of three and four other speakers, respectively. Fig. 9 and Fig. 10 depict results of a head tracking when the number of other speakers are three and four. According to the all results, it can be said that a subject does not always face to other speakers. However, it would be expected to estimate communication zone from the face that subject faces within a range from edge to edge of subjects.

C. Evaluation

An ideal detection of communication zone is to estimate direction for each sound source. However, according to the results on analysis of a head direction estimation, estimation of the direction for each sound source is quit difficult. Therefore, two evaluations are introduced. One is to estimate the directions with a certain margin, the other one is to estimate a range from edge to edge of sound source positions. Performance on estimation is defined as

\[ Q = \frac{n(S(j) \leq T)}{n(j)} \times 100 \text{ [%]} \tag{7} \]

where \( T \) is a correct zone, \( j \) is integer variable for index from 1 up to the number of frames, and function \( n() \) gives the number of elements.

Since a head has unstableness, a correct zone \( T \) is defined as a zone of a sound source direction with an additional margin \( \pm 7.5^\circ \), namely, the zone is \( \pm 15^\circ \) from a center of the zone.

Table II shows the results of evaluation with \( Q \). Average of \( Q \) is 65.2\% in case of two other talkers, and According to the analysis of a head movement for each subject, most of subjects face to a target direction. However, \( Q \) for subject I \( I \) is 54.1\%, and then this indicates individual variation. The results in case of three or four other talkers show increase in \( Q \), however, the similar tendency on individual variation.

Table III shows the results of evaluation by \( Q \) whose evaluation range from edge to edge of sound sources with an additional margin \( \pm 7.5^\circ \). According to the results in three patterns, the results indicate possibility to estimate a communication zone with a certain range.

D. Evaluation in a duration of utterance

In the previous evaluation, the results includes a duration of silence. When a communication zone information is applied to estimation of sound source positions, a head direction in a duration of utterance is especially important. Table IV shows results without silent duration. Evaluation range \( T \) in Eq. 7 is set to a zone of a sound source direction with an additional margin \( \pm 7.5^\circ \), namely, the zone is \( \pm 15^\circ \) from a center of the zone. In case of two other talkers, the highest value of \( Q \) is 72.8\% for subject V and the lowest one is 17.6\% for subject I. Individual variation is also observed in other two patterns.

Table V shows results when evaluation range \( T \) in Eq. 7 is set to all zones of sound source directions with an additional margin \( \pm 7.5^\circ \) because subject faces to a listener when other speakers talk each other. The lowest value of \( Q \) is 79.9\% for subject III in case of two other talkers, however, \( Q \) in all other cases is over 80\%.

VII. Conclusions

This paper performed experiments and analysis of a head tracking for binaural hearing aids. A head movement is estimated with an inertial sensor which is attached to left ear in a conversation with 2-4 other speakers. As the results, a zone where other speakers can be estimated, and the performance on estimation of the zone was over 80\%. The results show possibility to improve performance on estimating of sound source directions by a correction based on a head movement.
information. In future, the proposed zone information detection will be combined into the frequency domain binaural model.

ACKNOWLEDGMENT

Part of this work was supported by JSPS KAKENHI Grant Number 24500147.

REFERENCES


Fig. 8. Subject’s head direction zone when other talkers are in zone F and J.
Fig. 9. Subject’s head direction zone when other talkers are in border of zone E, zone CD and zone FG.

Fig. 10. Subject’s head direction zone when other talkers are in border of zone D, G, I and L.