Estimation of speaker and listener positions in a car using binaural signals

Madoka Takimoto¹, Takanori Nishino^{2,*}, Hiroyuki Hoshino^{1,3} and Kazuya Takeda¹

¹Graduate School of Information Science, Nagoya University,
Furo-cho, Chikusa-ku, Nagoya, 464-8603 Japan

²Center for Information Media Studies, Nagoya University,
Furo-cho, Chikusa-ku, Nagoya, 464-8603 Japan

³Toyota Central R&D Labs., Inc.,
41-1 Yokomichi, Nagakute, Nagakute-cho, Aichi, 480-1192 Japan

(Received 23 April 2007, Accepted for publication 27 June 2007)

Keywords: Binaural signal, Gaussian mixture model, Interaural level difference,
Detection of sound source direction, In-car environment

PACS number: 43.60.Jn [doi:10.1250/ast.29.110]

1. Introduction

The detection of sound source direction is a very important technique and is in wide use in fields such as speech enhancement, sound recording, and security systems. Robot hearing is an important subject related to the detection of sound source direction [1,2]. It is necessary for technology to achieve the same performance as human beings in various environments. To date there have been many studies based on microphone arrays; these reports describe methods that employ many microphones to obtain high detection performance. However, reducing the number of microphones could contribute to lowering costs and facilitating maintenance.

In this paper, speaker and listener positions are estimated by a binaural signal between them in noisy conditions. Several methods of detecting sound source direction using binaural signals have been proposed. In our research, the method using the cepstrum of an interaural level difference (ILD) [3] was employed. In that study, experiments for estimating sound source direction were conducted in eight reverberation conditions, and the results showed that the cepstrum of ILD was a useful feature parameter. However, evaluating not only user positions but also user situations and environments is an important problem for robot hearing. Our experiments were conducted for an in-car environment [4], which can be considered a complex acoustic environment, because it is a limited space that contains many acoustic materials with various shapes. Moreover, a speaker talks behind a listener or without looking at a listener in a car. Although speaker and listener positions do not change, different transfer functions exist between them. Therefore, the purpose of our investigation is to specify the speaker and listener positions within a car using binaural signals. Positional patterns that involve motion of the speaker's head are also considered. The given information includes a statistical model of the ILD cepstrum for all positional patterns. A Gaussian mixture model for each positional pattern is generated. The models are evaluated by specifying the positional pattern in six binaural signal-to-noise ratio conditions.

2. Measurement

Binaural room impulse responses (BRIR) were measured in a car (TOYOTA, SPRINTER TRUENO). In our measurements, the speaker and listener sat in two of the four car seats. Two kinds of head and torso simulators were used as speaker (B&K, 4128) and listener (G.R.A.S. Sound & Vibration, KEMAR Manikin). A swept sine signal [5] with a duration of 2.73 s was transduced by a loudspeaker (B&K, 4135) installed in the manikin. Microphones (SONY, ECM-77B) were positioned at the entrance of both ear canals to block them. The background noise level was 37.2 dB(A). The sampling frequency was 48 kHz.

Eighteen patterns of acoustic transfer functions were measured, involving those that have the same positional pattern but in which the speaker faces a different direction. Table 1 shows the positional conditions. In this table, an angle denotes the rotation of the speaker manikin's head; a positive angle is the clockwise rotation, and a negative angle the counterclockwise rotation. Figure 1 shows the BRIR between the seat in back of the driver and the driver's seat. The listener manikin sat in the seat in back of the driver, and the speaker manikin sat in the driver's seat ($\{4, 1, 0^\circ\}$ condition in Table 1). At this BRIR, attenuation is fast, and the amplitude of a reflection wave is greater than that of a direct wave.

Two kinds of road noises were also recorded using the listener manikin. One of them was recorded in a moving car and the other during idling when the engine ran at 800 rpm. In our study, these noises were called binaural noise. The binaural noise was recorded at each seat. However, the noise in the driver's seat was not recorded in a moving car. The sampling frequency was 48 kHz.

3. Experiment

3.1. Experimental conditions [3]

Our proposed method employs the ILD envelopes to represent a rough tendency. To obtain the ILD envelopes, we calculate parameters, which are similar to a cepstrum, with the following procedure.

- (1) The signal that arrives at the left ear is denoted as $s_L[t]$, and that for the right ear is $s_R[t]$. Both signals are

*e-mail: nishino@media.nagoya-u.ac.jp

Table 1 Positional patterns (1: seat next to the driver, 2: driver's seat, 3: rear seat on the left, 4: seat in back of the driver). Angle is rotation of speaker manikin's head.

{Listener, Speaker, Angle} =		
{1, 2, 0°}, {1, 3, 0°}, {1, 4, 0°},		
{2, 1, 0°}, {2, 3, 0°}, {2, 4, 0°},		
{3, 1, 0°}, {3, 2, 0°}, {3, 4, 0°},		
{4, 1, 0°}, {4, 2, 0°}, {4, 3, 0°},		
{1, 2, -45°}, {2, 1, 45°}, {3, 1, 45°},		
{3, 2, -45°}, {4, 1, 45°}, {4, 2, -45°}

Fig. 1 Right ear's binaural room impulse response between seat in back of driver (listener) and driver's seat (speaker).

truncated respectively with the hamming window whose frame length is l and frame shift is l_s .

- (2) The ILD is calculated using Eq. (1). However, when one of the signals has a lower absolute value of amplitude than the threshold, calculation of the ILD is not conducted:

$$|S_{LR}(f)| = \frac{|S_L(f)|}{|S_R(f)|}, \quad (1)$$

where $S_L(f)$ is the magnitude response of the left ear's signal, $S_R(f)$ is that of the right ear, and f denotes frequency. In our experiments, the threshold is 0.005 (from the result of a preliminary experiment).

- (3) The Fourier transform is applied to the logarithm of the ILD, and the feature parameter $c[n]$ is obtained using Eq. (2). In our study, the ILD cepstrum is denoted by $c[n]$. Thus, the ILD envelope is obtained by the lower-order ILD cepstrum.

$$c[n] = \frac{1}{N} \sum_{k=0}^{N-1} 10 \log_{10} |S_{LR}(f)| e^{j2\pi kn/N}, \quad (2)$$

$(n = 0, \dots, N).$

If a BRIR consists of the convolution of the room impulse response and the HRTF, the ILD cepstrum can be represented as the sum of components resulting from the room impulse response and the HRTF.

The speaker and listener positions were estimated by using a Gaussian mixture model. The training procedure is

described as follows.

- (1) The distribution of the ILD cepstrum is approximated with the Gaussian distribution. The statistical models for every positional pattern can be represented by

$$\lambda_p = \{w_m, \mu_m, \Sigma_m | m = 1, 2, \dots, M\}. \quad (3)$$

- (2) The EM (expectation maximization) algorithm gives the weight for each distribution w_m , the mean μ_m , and the covariance matrix Σ_m . Then, the estimation model λ_p is trained for every positional pattern. That is to say, the statistical models of the ILD cepstrum for all positional patterns were given. Our method uses the diagonal covariance matrix.

Below is the procedure for estimating the positional pattern with the Gaussian model.

- (1) The ILD cepstrum of the input signals is calculated.
- (2) The posterior probabilities between the statistical value, such as the mean and variance, and every trained Gaussian model are calculated. The model that gives the maximum probability is considered to be the positional pattern.

The frame length was 128, frame shift was 32, and the number of mixtures was one. ILD cepstra from 0 to 15-th were used.

In the experiment, a human speech-like (HSL) noise [6] was used as the training signal. HSL noise is a kind of bubble noise generated by superposing a lot of speech signals, and characteristics of HSL noise can be controlled by changing the number of superpositions. We used HSL noise with 24 superpositions, signal durations of which were 2s. The training binaural signals were obtained by convolving the BRIR and the HSL noises.

Speech signals were used as the evaluation signal. Since six subjects (three males and three females) read out 22 sentences each, the resultant evaluation signals were 132.

The evaluation binaural signals were obtained by convolving the BRIR and speech signals. A binaural signal-to-noise ratio was defined as the following equation to mix binaural speech signals and the binaural noise signals:

$$\text{SNR}_{\text{binaural}} = 20 \log_{10} \frac{\sqrt{\sum s_L[t]^2 + \sum s_R[t]^2}}{\sqrt{\sum n_L[t]^2 + \sum n_R[t]^2}} \text{ [dB]}, \quad (4)$$

where s_L and s_R are binaural speech signals, and n_L and n_R are binaural noise signals. The $\text{SNR}_{\text{binaural}}$ was calculated except silence. The experiments were conducted under six binaural SNR conditions: -5, 0, 5, 10, 15, and ∞ dB. The ∞ dB denotes that no noise is mixed.

In the case of the idling condition, the number of evaluation signals was 2,376 (132 speech signals multiplied by 18 positional patterns). In the case of the moving car 1,848 (132 speech signals multiplied by 14 positional patterns).

3.2. Results

Figure 2 shows the results when the binaural noise in idling was used, and Fig. 3 shows in the conditions of the moving car noise. The correct answer was when both manikins' positions were estimated correctly. In other words, the answer was correct when an appropriate model was selected.

Fig. 2 Results when binaural noise in idling condition was used.

4. Conclusion

We examined the estimation of speaker and listener positions using binaural signals. Binaural room impulse responses were measured for 18 positional patterns, and evaluation binaural signals were obtained by convolving these impulse responses and speech signals. Experiments were conducted in six SNR conditions for two kinds of noises. In the results, we obtained a greater than 95% correct rate in 5-dB binaural SNR. It was clarified that speaker and listener positions are detected by binaural signals and that our method works well in noisy conditions. Therefore, our evaluated method is robust to situations of the sound source and the influence of noise. Future works will focus on improving performance in hard noisy environments, using interaural time differences and examining acoustic environments in cars.

References

- [1] K. Nakadai, H. Okuno and H. Tsujino, "Improvement of robot audition system by scattering theory," *JSAI Tech. Rep.*, SIG-

Fig. 3 Results when binaural noise in moving car condition was used.

Challenge-0318-6, pp. 33–38 (2003).

- [2] H. Nakashima, Y. Kawabuchi, K. Hagiwara, Y. Chisaki and T. Usagawa, "A specific word recognition system by using FDBM for humanoid," *Proc. Spring Meet. Acoust. Soc. Jpn.*, pp. 643–644 (2004).
- [3] T. Nishino, N. Inoue, K. Itou and K. Takeda, "Estimation of sound source direction based on Gaussian model using envelope of interaural level difference," *J. Acoust. Soc. Jpn. (J)*, **63**, 3–12 (2007).
- [4] M. Takimoto, T. Nishino and K. Takeda, "Estimation of a talker and listener's positions in a car using binaural signals," *J. Acoust. Soc. Am.*, **120**, 3216 (2006).
- [5] N. Aoshima, "Computer-generated pulse signal applied for sound measurement," *J. Acoust. Soc. Am.*, **69**, 1484–1488 (1981).
- [6] S. Kajita, D. Kobayashi, K. Takeda and F. Itakura, "Analysis of speech features included in human speech-like noise," *J. Acoust. Soc. Jpn. (J)*, **53**, 337–345 (1997).