Detection of Speaker Direction based on the On-and-Off Microphone Combination for Entertainment Robots

Takeshi Kawabata[1], Masashi Fujiwara[1], and Takanori Shibutani[1]

Kwansei Gakuin University, 2-1 Gakuen, Sanda City, 669-1337, Japan,
kawabata@ksc.kwansei.ac.jp,
WWW home page: http://ist.ksc.kwansei.ac.jp/~kawabata/

Abstract. An important function of entertainment robots is voice communication with humans. For realizing them, accurate speech recognition and a speaker-direction detection mechanism are necessary. The direct-noise problem is serious in such speech processing. The microphone attached to the robot body receives not only human voices but also motor and mechanical noises directly. The direct noises are often larger than distance voices and fatally degrade the speech recognition rate. Even if the microphone close to the user (”on-mic”) is used for speech recognition, the body microphones (”off-mic”) are still necessary for detecting the speaker direction under the severe condition with direct noises. This paper describes a new method for detecting the speaker direction based on the on-and-off microphone combination. The system searches for the spectral elements of ”on-mic” voice in the other ”off-mic” channels. The segregated power ratio or the time delay between the ”off-mic” channels is used for detecting the speaker direction. Experiments show that the proposed method effectively improves the direction detection accuracy during the robot moves.

1 Introduction

Recent mechatronics technologies realized the autonomous robots which work together with our human beings. In near future, house-keeping robots may cook breakfast, wash clothes, and clean rooms. Also entertainment robots may sing, dance, gesticulate, and chat with us. However an important function of such robots is voice communication with humans, it is still difficult now. The serious problem in speech recognition is direct noises. The microphone attached to the robot body receives not only human voices but also motor and mechanical noises directly. Direct motor noises are often larger than incoming voices and fatally degrade the speech recognition rate.

An ”on-mic” approach is promising to avoid this problem. That means the microphone located close to the user. For example, a mobile-phone like voice commander or a head-set microphone can receive the user’s voice without any direct noises.

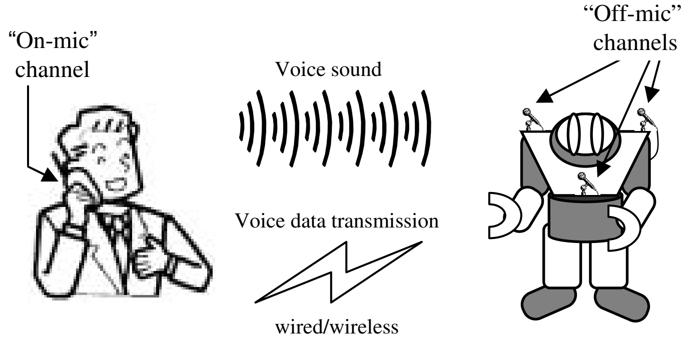


Fig. 1. On-mic and Off-mic channels

Even if the robot has the "on-mic" input channel, body microphones (i.e. "off-mic" channels) are still necessary because these signals are used for detecting the speaker direction. When the command "Come here!" was recognized, the robot has to detect the direction of the speaker and turn its body before walking.

This paper describes a new method for detecting the speaker direction based on the on-and-off microphone combination. The system searches for the spectral elements of "on-mic" voice from the other "off-mic" channels. The segregated power ratio or the time delay between the "off-mic" channels is used for detecting the speaker direction (Fig. 1). Experiments show that the proposed method effectively improves the direction detection accuracy during the robot moves.

2 Detection of Speaker Direction

2.1 Localization Queues

At the beginning of this section, we have to mention about the excellent direction detection mechanism of the human auditory system. We humans can easily find the direction of incoming sounds using two ears. The main queues of sound localization are the interaural level difference (ILD) and the interaural time difference (ITD) [1, 2].

Notice that these queues are essential to construct the machine auditory system. Several sophisticated researches have been carried [3, 4] for adding the auditory function to computers. The system enables a computer to localize incoming sounds, to segregate them into speech/noise, and to recognize the speech as a command. It works well even under the office-level noise environment.

In the case of real robots, the serious problem occurs derived from their direct motor and mechanical noises. The direct noises are often larger than incoming sounds and destroy the direction detection queues.

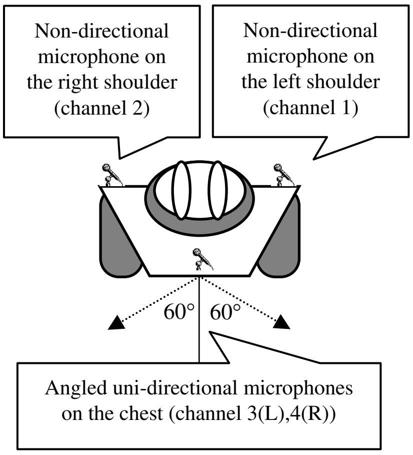

Non-directional microphone on the right shoulder (channel 2)
Non-directional microphone on the left shoulder (channel 1)
60°
60°
Angled uni-directional microphones on the chest (channel 3(L),4(R))

Fig. 2. Microphones attached to the robot body

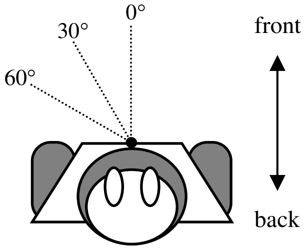
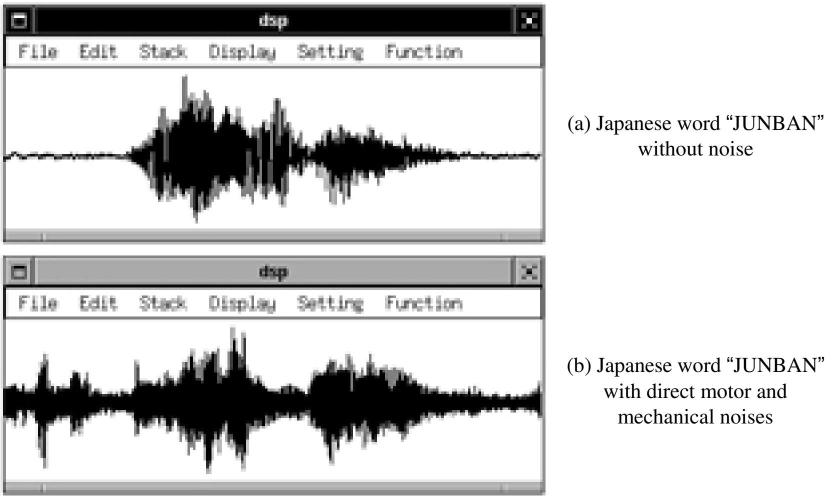


Fig. 3. Incident angles of voice data

2.2 Traditional Methods and Problems

Figure 2 shows the positions of microphones attached to the robot body in our experiments. Non-directional microphones are attached to the left and right shoulders of the robot (channel 1 and 2). Angled uni-directional microphones (i.e. one-point stereo microphone) are attached to the robot chest (channel 3 and 4). And we use a head-set microphone as an "on-mic" channel for capturing clear voice (channel 5).

Phonetically balanced 50 words are pronounced by a speaker from three incident angles ($0^o, 30^o, 60^o$) (Fig. 3). The training data set is pronounced under no noise condition. And the test data set is pronounced during the robot drives its arms. All data are recorded through the four "off-mic" channels and the one "on-mic" channel. Figure 4 (a) shows a speech waveform example of Japanese word "Junban" without noise. Figure 4 (b) shows a speech waveform of the same word with the motor and mechanical noises. The average S/N ratio of the test data is 10dB.


dsp
File Edit Stack Display Setting Function
dsp
File Edit Stack Display Setting Function

(a) Japanese word “JUNBAN” without noise

(b) Japanese word “JUNBAN” with direct motor and mechanical noises

Fig. 4. Speech samples with/without motor and mechanical noises

ILD-based Method A simple level-based direction detection mechanism is shown in Fig. 5. The system calculates the log powers of both (3 and 4) channel signals and their difference. The calibration unit was tuned by training data in advance. Table 1 shows the performance of the simple direction detection mechanism based on the level-based method. A direction detection result is judged to be correct only when the incident angle and the recognized angle are identical. Under the condition without noise, such a simple mechanism still determines the sound directions with 82 % accuracy. However, the performance was drastically degraded under the motor and mechanical noises.

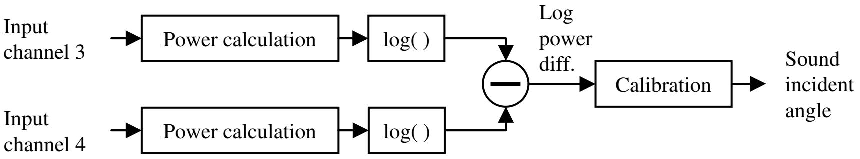


Fig. 5. A simple direction detection mechanism based on the level difference between two channels (ch3: uni-directional microphone on the robot chest(L), ch4: uni-directional microphone on the robot chest(R))

Table 1. Direction detection scores by the simple level-based method

Noise condition	Direction detection score(%)
without noise	82.0
motor and mechanical noises	34.7

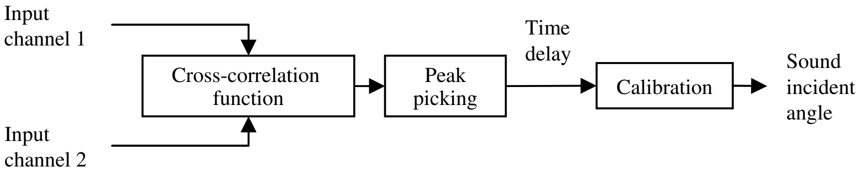


Fig. 6. A simple direction detection mechanism based on the time difference between two channels (ch1: non-directional microphone on the robot shoulder(L), ch2: non-directional microphone on the robot shoulder(R))

ITD-based Method A simple time-based direction detection mechanism is shown in Fig. 6. The system calculates the cross-correlation function between channel 1 and channel 2, and determines their time delay by searching for the peak. The calibration unit was tuned by training data in advance. Table 2 shows the performance of the simple direction detection mechanism based on the time-based method. Under the condition without noise, this mechanism determines the sound directions with 72 % accuracy. This score is little bit worse than the level-based method. On the contrary, under the severe condition with direct motor and mechanical noises, the time-based method achieved better scores than the level-based method. This result indicates that the time difference feature is more robust for direction detection against the direct noises than the level difference feature.

2.3 New Method using the On-and-Off Microphone Combination

The motor and mechanical noises propagate through the robot body and vibrate the microphones directly. The strong noise hides the interaural difference derived

Table 2. Direction detection scores by the simple time-based method

Noise condition	Direction detection score(%)
without noise	72.0
motor and mechanical noises	63.3

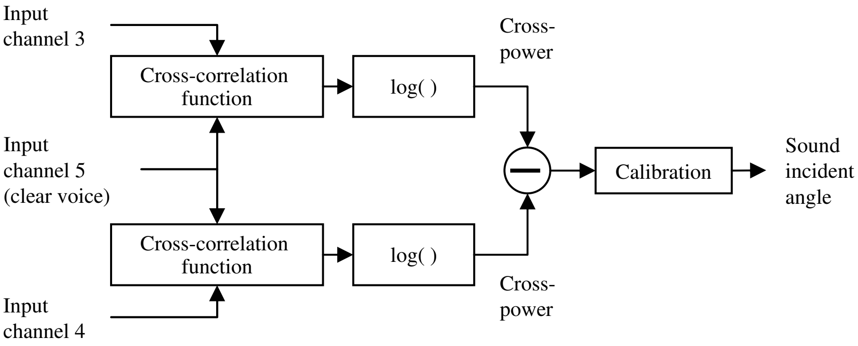


Fig. 7. New direction detection mechanism based on the level difference and on-and-off microphone combination (ch3: uni-directional microphone on the robot chest(L), ch4: uni-directional microphone on the robot chest(R), ch5: uni-directional microphone close to the user)

Table 3. Direction detection scores by the simple and new level-based method under the direct motor and mechanical noises

Method	Direction detection score(%)
simple level-based	34.7
on-and-off mic combination	59.3

from incoming speech. This is the reason why the simple level-based direction detection method does not work well during the robot moves.

The framework shown in Fig. 1 has an "on-mic" channel which is the microphone located close to the user. A mobile-phone like voice commander or a head-set microphone can receive the user's clear voice for accurate speech recognition. Also this clear voice can be used for detecting the speech direction.

Modified ILD-based Method Figure 5 shows the signal flow diagram of a new level-based direction detection mechanism. In spite of the power calculation, the system calculates the cross-correlation function between the "on-mic" channel and two "off-mic" channels. This means that the system searches for the spectral elements of the "on-mic" voice in the other "off-mic" signals. Only the log power elements derived from the voice signals are picked up and compared. The calibration unit was tuned by training data in advance. Table 3 shows the performance of the modified direction detection mechanism based on the level-based method. About the 1/3 of direction detection errors are reduced by this approach.

Modified ITD-based Method Figure 6 shows the signal flow diagram of a new time-based direction detection mechanism. The system calculates the

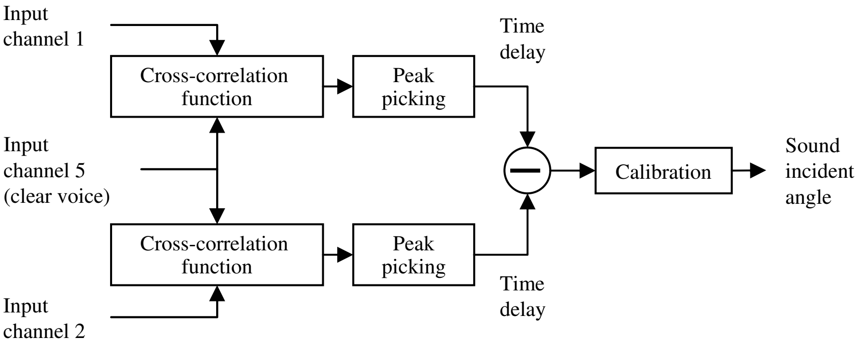


Fig. 8. New direction detection mechanism based on the time difference and on-and-off microphone combination (ch1: non-directional microphone on the robot shoulder(L), ch2: non-directional microphone on the robot shoulder(R), ch5: uni-directional microphone close to the user)

Table 4. Direction detection scores by the simple and new time-based method under the direct motor and mechanical noises

Method	Direction detection score(%)
simple time-based	63.3
on-and-off mic combination	76.0

cross-correlation function between the ”on-mic” channel and two ”off-mic” channels similarly to the modified level-based method. Peak picking for the cross-correlation functions determines the time delays between the ”on-mic” channel and two ”off-mic” channels. The difference of them indicates the time delay of voice signal elements between ”off-mic” channels (channel 1 and channel 2). The calibration unit was tuned by training data in advance. Table 4 shows the performance of the modified direction detection mechanism based on the time-based method. The best direction detection score 76.0 % is achieved by this method under the motor and mechanical noise condition.

3 Summary

The new method for detecting the speaker direction based on the on-and-off microphone combination was proposed in this paper. For realizing the voice communication functions on entertainment robots, accurate speech recognition and a speaker-direction detection mechanism are necessary. The microphones attached to the robot body (”off-mic”) are suitable for detecting the speaker direction. However, they are often disturbed by the direct motor and mechanical noises. The microphone located close to the user (”on-mic”) is suitable for accurate speech recognition because it can receive the user’s voice without any

Table 5. Summary of direction detection scores

Method	Noise condition	Ch1	Ch2	Ch3	Ch4	Ch5	Score(%)
simple level-based	without noise	-	-	o	o	-	82.0
simple level-based	moter and mech.	-	-	o	o	-	34.7
proposed level-based	moter and mech.	-	-	o	o	o	59.3
simple time-based	without noise	o	o	-	-	-	72.0
simple time-based	moter and mech.	o	o	-	-	-	63.3
proposed time-based	moter and mech.	o	o	-	-	o	76.0

direct noises. However, it cannot determine the speaker direction. This paper improved the traditional direction detection method by using the on-and-off microphone combination. The system searches for the spectral elements of "on-mic" voice in the other "off-mic" channels. The segregated power ratio or the time delay between the "off-mic" channels is used for detecting the speaker direction. Experiments show that the proposed method effectively improves the direction detection accuracy. The best direction detection score 76.0 % was achieved under the motor and mechanical noise condition.

The physiological or technological researches in this field are progressing year by year [5–8]. We would like to apply new knowledge into our system in future works.

References

1. Jeffress, L. A.: A place theory of sound localization. J. Comp. Physiol. Psychol. **41** (1948) 35–39
2. Blauert, J.: Spatial hearing: The psychophysics of human sound localization. (Revised ed.). MIT Press (1997).
3. Nakatani, T., Okuno, H.: Harmonic Sound Stream Segregation Using Localization and Its Application to Speech Stream Segregation. Speech Communcations **27, 3-4,** Elsevier (1999) 209–222
4. Aoki, M., Okamoto, M., Aoki, S., Matsui, H., Sakurai T., Kaneda, Y.: Sound source segregation based on estimating incident angle of each frequency component of input signals acquired by multiple microphones. Acoustical Science and Technology **22, 2** (2001) 149–157
5. Huang, J., Ohnishi, N., Guo, X., Sugie, N.: Echo avoidance in a computational model of the precedence. Speech Communication **27** (1999) 223–233
6. Renevey, P., Vetter, R., Kraus, J.: Robust speech recognition using missing feature theory and vector quantization. Proc. Eurospeech-2001 **2** (2001) 1107-1110
7. Nakadai, K., Matusura, D., Okuno, H., Kitano, H.: Applying Scattering Theory to Robot Audition System. Proc. IROS-2003 (2003) 1147–1152
8. Furukawa, S., Maki, K., Kashino, M., Riquimaroux, H.: Dependence of the interaural phase difference sensitivities of inferior collicular neurons on a preceding tone and its implications in neural population coding. J. Neurophysiol. in press (2005)