Study of Adaptive Processing on Body-Conducted Speech Recognition for Support System of Marine Engine Operation*

Ishimitsu, Shunsuke**

In recent years, speech recognition systems have been employed in vehicle instrumentation in a wide variety of environments. Speech recognition plays an important role in the chief engineering systems on ships. In this type of system, speech recognition supports the engine room controls, and lower than 0-dB signal-to-noise ratio (SNR) operability is required. In such a low SNR environment, a noise signal can be misjudged as speech, dramatically decreasing the recognition rate. Therefore, this study focuses on a recognition system that uses body-conducted signals. Since noise is not introduced within body-conducted signals that are conducted in solids, even within sites which are low SNR environments such as engine rooms, it is necessary to construct a system with a high recognition rate. However, when constructing such systems, in order to create models specialized for body-conducted speech, it is necessary to use learning data consisting of sentences that must be read numerous times. Therefore, in the present study we applied a method in which the specific nature of body-conducted speech is reflected within an existing speech recognition system with only small numbers of vocalizations. Because two of the prerequisites for operation within a site such as an engine room where noise exists are both "hands-free" and "eyes-free" operations, we also investigated the effects of making this type of system wireless.

1. Introduction

In recent years, speech recognition systems have been used in a wide variety of environments, including automobile internal systems. Speech recognition plays a major role in the dialogue-type marine engine operation support system [1] currently under investigation. In this system, speech recognition would come from the engine room, which contains the engine apparatus, electric generator, and other equipment, and control support would also be performed within the engine room. Therefore, operations with a 0-dB signal-to-noise ratio (SNR) or less are required. To date, noise has been determined to be a portion of the speech in such low SNR environments, and the speech recognition rates have been remarkably low. This has prevented the introduction of recognition systems, and to the present date, almost no research has been performed on speech recognition systems that operate within low SNR environments. In this study, we investigated a recognition system that uses body-conducted speech, that is, types of speech that are conducted within a physical body rather than speech signals themselves [2]-[4].

Since noise is not introduced within body-conducted signals that are conducted in solids, even within sites such as engine rooms which are low SNR environments, it is necessary to construct a system with a high recognition rate. However, when constructing such systems, in order to create a dictionary specialized for body-conducted speech, learning data consisting of sentences that must be read numerous times is required. Therefore, in the present study, we applied a method in which the specific nature of body-conducted speech is reflected within an existing speech recognition system with only small numbers of vocalizations. Because two of the prerequisites for operating within a site such as an engine room where noise exists are both "hands-free" and "eyes-free" operations, we also investigated the effects of making such a system wireless

2. Dialogue-type marine engine operation support system using body-conducted speech

Since the numbers of sailors have decreased dramatically in recent years, there is a shortage of skilled maritime engineers. Therefore, a database which stores the knowledge used by skilled engineers has been constructed [1].

*Received March 2, 2006

**University of Hyogo

Fig.1 Dialogue-type marine engine operation support system using body-conducted speech.

Fig.2 Sampling location for body-conducted speech.

In this study, this knowledge database is accessed by speech recognition. This system can be used to educate sailors, and make it possible to check the ship's engines.

Figure 1 shows a conceptual diagram of a dialogue-type marine engine operation support system using body-conducted speech. The signals are detected with a body-conducted microphone and then wirelessly transmitted, and commands or questions from the speech-recognition system located in the engine control room are interpreted. After a search is made for a response to these commands or questions, speech recognition results, and confirmation as to whether or not it is best to reflect such commands within the control system are speech-synthesized and then outputted to a monitor. The speech-synthesized sounds are replayed in an ear protector/speaker unit, and while continuing communications, work can be performed as safety is continuously confirmed.

The present research is concerned with the development of the body-conducted speech recognition portion of this system. In this portion of the study, a system was created based on a recognition engine that is itself based on an

HMM (Hidden Markov Model) incidental to a database [5]. With this system, multivariate normal distribution is used as the output probability density function, and a mean vector μ that takes an n -dimensional vector as the frame unit of speech features quantities and a covariance matrix Σ are used; these are expressed as follows: [6].

$$b(o, \mu, \Sigma) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}(o-\mu)^T \Sigma^{-1}(o-\mu)} \quad (1)$$

As for the HMM parameters, they are shown using the two parameters of this output probability and the state transition probability. To update these parameters using conventional methods, utterances repeated 10 to 20 times, at the very least, would be required. To perform learning with only several utterances, we focused only on the relearning of the mean vector μ within the output probability, and thus created a user-friendly system for performing adaptive processing.

Our goal is to introduce the system we are investigating into the "Oshima-maru" training ship of the Oshima National College of Maritime Technology. The Oshima-maru is a 226 ton training vessel, with 56 regular crewpersons. Its main engine is a diesel engine with a capacity of 1300 PS x 370 rpm. Situated within the engine room, in addition to the main engine, are two electric generators. During ordinary sailing operations, the noise level within the engine room is 98 dB SPL (sound pressure level), while at anchor it is 94 dB SPL.

3. Adaptive processing for body-conducted speech

3.1 Sampling location for body-conducted speech

Figure 2 shows candidate locations for the body-conducted speech during this experiment. The upper left part of the upper lip was selected as the signal sampling location [7]. Amplitudes of the body-conducted speech at the zygomatic arch and pharynx are 10 to 20 dB lower than the body-conducted speech at the upper left part of the upper lip. Based on the frequency characteristics, we believed that recognition of the body-conducted signal would be difficult using an acoustic model built using acoustic speech signals; however, by using the upper left part of the upper lip, the highest clarity point, we thought that it would be possible to recognize the body-conducted speech with the acoustic model built from acoustic speech by using adaptive signal processing or speaker adaptation.

To investigate the effectiveness of a body-conducted signal model, we also examined the characteristics of feature vectors. Although the upper left part of the upper lip has the

closest proximity to voice characteristics, it is not close enough to capture all of the characteristics of speech. Because of this, we concluded that it would be difficult to build a body-conducted speech model solely with a voice model.

However, we concluded that it might be possible to build a body-conducted speech recognition system by building the model at the upper left part of the upper lip, and by optimizing speech-conducted speech signals based on a voice model.

3.2 Adaptive processing for body-conducted speech

In order to use body-conducted speech with a speech recognition system, models specialized for body-conducted speech are required. Within the construction of such systems, in order to create models specialized for body-conducted speech, learning data consisting of sentences that must be read numerous times are needed. With this system, multivariate normal distribution is used as an output probability density function, and a mean vector μ that takes an n-dimensional vector as the frame unit of speech featuring quantities and a covariance matrix Σ are used; these are expressed as equation (1).

The HMM parameters are shown using the two parameters of this output probability and the state transition probability.

The mean vectors μ , the covariance matrix Σ within the

output probability $b(o, \mu, \Sigma)$ and the state transition probability have to be determined for new HMM. To update these parameters using conventional methods, utterances repeated 10 to 20 times, at the very least, would be required using the Baum-Welch algorithm [8]. Because the utterances must be repeated numerous times, a convergence calculation of an algorithm to determine all parameters simultaneously must be used. Therefore, in the present study, we applied a method in which the specific nature of body-conducted speech is reflected within an existing speech recognition system with only small numbers of vocalizations. To perform learning with only several utterances, we focused only on the relearning of the mean vector μ within the output probability. The value of the covariance matrix Σ and the state transition probability are set as the initial value of an unspecified speaker HMM. By making it concentrate only on the presumption of a mean vector μ , parameter estimation using a small number of vocalizations was made possible [9].

Table 1 Speech recording equipment

Recorder	TEAC RD-200T
Microphone	Ono Sokki MI-1431
Microphone amplifier	Ono Sokki SR-2200
Accelerator	Ono Sokki NP-2110
Accelerator amplifier	Ono Sokki PS-602

Table 2 Experimental conditions

Valuation method	Three set utterance of 100 words
Vocabulary	JEITA 100 locality names
Microphone position	From the mouth to about 20cm
Accelerator position	The upper left part of the upper lip

Table 3 The result of preliminary testing

	anchorage		Cruising	
	Speech	Body	Speech	Body
Anechoic room	45%	14%	2%	45%
Anechoic room + noise	64%	10%	0%	49%
Cabin	35%	9%	1%	42%
Cabin + noise	62%	4%	0%	48%

4. Recognition experiments

4.1 Selection of the optimal initial model

The speech recording equipment used is shown in Table 1, and the experimental conditions are shown in Table 2. For system evaluation, we used speech as extracted in the following four environments:

- Speech within a room interior under silence
- Body-conducted speech within a room interior under silence
- Speech within the engine room of the Oshima-maru as the ship was running on Nav/Full
- Body-conducted speech within the engine room of the Oshima-maru as the ship was running on Nav/Full

The noise within the engine room of the Oshima-maru as the ship was running was 98 dB SPL, and the S/N ratio when a microphone was used was -25 dB. This data

consisted of 100 terms read by a male aged 20, and the terms were read three times in each environment. As for the body-conducted speech, extractions were taken from the upper left portions of the upper lip [3],[4],[8], the effectiveness of which was confirmed in previous research. The initial dictionary model to be used for learning was a model for an unspecified speaker created by adding noise to speech extracted within an anechoic room. This model for an unspecified speaker was selected through preliminary testing. The results of preliminary testing are shown in Table 3. The model which added the noise to the speech recorded in the anechoic room had the highest recognition rate. 64% recognition was achieved for the speech, and 49% for the body-conducted speech. These percentages were not sufficient for practical use. Therefore, adaptation processing was applied to the construction of a new model.

4.2 The Effect of Adaptation Processing

The recognition test results in the cases where adaptive processing was performed for room-interior speech and engine-room interior speech are shown in Table 4, Figures 3 and 4. The underlined portions show the results of the tests performed in each stated environment. From these results, it can be observed that in tests of recognition and signal adaptation via speech within the machine room, there was almost no operation whatsoever. That result is shown in Figure 3, and is thought to have been due to the fact that, as the engine room noise was louder than the speech sounds, extraction of the speech features failed. Conversely, with room interior speech, signal adaptation was performed, and in the case where the environments for performing signal adaptation and recognition were equivalent, an improvement of the recognition rate of 27.66 % was achieved, as shown in Figure 4. In this case, there was also a 12.99 % improvement of the recognition rate for body-conducted speech within the room interior; however, that recognition rate was around 20 %, and thus would be unable to withstand practical use. Nevertheless, based on these results, we found that using this method enabled recognition rates exceeding 90 % with just one iteration of the learning samples.

The results of cases where adaptive processing was performed for room-interior body-conducted speech and engine-room interior body-conducted speech are shown in Table 5, Figures 5 and 6. Similarly to the case where adaptive processing was performed using speech, when the environment where adaptive-processing and the environment where recognition was performed were equivalent, high recognition rates of around 90 % were

Table 4 Result of adaptation processing with speech (%)

Valuation	Candidate for adaptation		
	Room	Engine Room	No adaptation
Speech(Room)	<u>90.66</u>	1.33	63.00
Body(Room)	22.66	1.33	9.67
Speech(Engine)	1.00	<u>1.50</u>	0.67
Body(Engine)	46.50	1.50	45.00

Fig.3 Signal adaptation with speech (cruising).

Fig.4 Signal adaptation with speech (room).

Table 5 Result of adaptation processing with body-conducted speech (%)

Valuation	Candidate for adaptation		
	Room	Engine Room	No adaptation
Speech(Room)	40.67	46.17	63.00
Body(Room)	<u>86.83</u>	26.83	9.67
Speech(Engine)	1.50	1.00	0.67
Body(Engine)	49.00	<u>95.50</u>	45.00

obtained, as shown in Figure 5. In Figure 6 It can be observed that signal adaptation using engine-room interior body-conducted speech and the recognition results were 95 % and above, with 50 % and above improvements, and that we had thus attained the level needed for practical usage.

Fig.5 Signal adaptation with body-conducted speech (room).

Fig.6 Signal adaptation with body-conducted speech (cruising).

4.3 Investigation of the effects of making the system wireless

We next investigated the effects of making the system wireless. The specification of the wireless system used for this experiment is shown in Table 7. This data consisted of 100 terms read by a male aged 20 (a different man than the person who read the terms in the earlier experiment), and for each one environment, the terms were uttered three times. For the body-conducted speech, the man who read the terms wore a helmet, and the terms were thus extracted from the top of his head (calvaria) portion. The effectiveness of this position has been confirmed in the documentation [4]. The initial dictionary model to be used for learning was, as in the previous tests, a model for an unspecified speaker. Here, the noise was white noise generated by a speaker, and was set so as to be in the vicinity of 0 dB SNR. The results for this experiment are shown in Table 8. From these results we concluded that if adaptive processing is performed when wired, the recognition rate becomes high, and thus the usefulness is confirmed. However, for speech transmitted wirelessly, the recognition rate was lower. The reason for

Table 6 Specifications for a wireless system

Manufacturer	MOTOROLA
Part number	GL2000
Frequency	154.45-154.61MHz
Transmitting output	1W/5W

Table 7 Results of a wireless vs. a cable system (%)

Conditions			No adaptation	adaptation
Cable	Quiet	speech	53.33	98.33
		body	43.66	97.00
wireless	Quiet	speech	3.33	77.00
		body	5.00	79.33
wireless	Noisy	speech	1.60	57.66
		body	2.00	62.00

Fig.7 /hachinohe/ with cable.

Fig.8 /hachinohe/ with wireless.

this is thought to be because, when the wireless type system was used, the noise was in the same frequency bandwidth as speech. The spectrogram analysis result of the speech /hachinohe/ using the cable and wireless is shown in Figures 7 and 8, respectively. From these figures it can be seen that

although the speech remained less than 4000 Hz, noise overlap on the whole zone can be observed.

This result suggests that it is necessary to test another method, such as wireless LAN instead of a walkie-talkie.

5. Conclusions

We investigated a body-conducted speech recognition system for the establishment of a usable dialogue-type marine engine operation support system that is robust in noise, even in a low SNR environment such as an engine room. By bringing body-conducted speech close to audio quality, we were able to examine ways to raise the rate of speech recognition. We introduced an adaptive processing method and confirmed the effectiveness of adaptive processing via small repetitions of utterances. Concretely, in an environment of 98-dB SPL, within only one utterance of the learning data, improvements of 50 % or above of the recognition rates were achieved, and recognition rates of 95 % or higher were attained. Based on these results, it was confirmed that this method will be effective.

In a wireless version of the system, the results showed a worsening of recognition rates due to noise in the speech bandwidth; even when adaptive processing was performed, a sufficient recognition rate could not be obtained. It is necessary to test a wireless system within the actual environment of the Oshima-maru, and it is also necessary to investigate other wireless methods.

Acknowledgements

The authors would like to express their profoundest gratitude for the advice and suggestions so kindly provided by Dr. Kuniyuki Matsushita, Director of Engines, Yuge National College of Maritime Technology, and by Mr. Toshikazu Yoshimi of Pioneer Corporation. We would also like to thank Mr. Masashi Nakayama, Mr. Yasuki Murakami, and the students at the Oshima National College of Maritime Technology who cooperated in the experiment. The work described was supported by a Grant-in-Aid for Scientific Research 16560707.

References

- [1] "Support system using oral communication and simulator for marine engine operation", Matsushita, K. and Nagao, K., Journal of Jap. Inst. Mar. Eng. Vol.36, No.6, pp.34~42. (2001)
- [2] "Study for constructing a recognition system using the bone

conduction speech", Ishimitsu, S., Kitakaze, H., Tsuchibushi, Y., Takata, Y., Ishikawa, T., Saito Y., Yanagawa H. and Fukushima M., Proc. Autumn Meet. Acoust. Soc. Jpn. pp.203~204 (2001).

- [3] "Study for bone-conducted speech recognition system under noisy environment", Haramoto, T. and Ishimitsu, S., Proc. 31st gra. Stu. Mech. Soc. Jpn., pp.152 (2001).
- [4] "Improvement of the speech sound quality of the vibration pick up microphone for speech recognition under noisy environment", Saito, Y., Yanagawa, H., Ishimitsu, S., Kamura K. and Fukushima M., Proc. Autumn Meet. Acoust. Soc. Jpn. I, pp.691~692 (2001).
- [5] "Continuous speech corpus for research", Itabashi S., Japan Information Processing Development Center (1991).
- [6] "Speech recognition using the probability model", Nakagawa S., IEICE (1988).
- [7] "Study of Body-Conducted Speech Recognition for Support of Maritime Engine Operation", Ishimitsu, S., Nakayama M. and Murakami, Y, Journal of Jap. Inst. Mar. Eng. Vol.39, No.4, pp.35~40. (2001)
- [8] "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains", Baum, L.E., Petrie, T., Soules, G. and Weiss, N., Ann. Math. Stat., 41, 1, pp.164~171.(1970)
- [9] "Method of modifying feature parameter for speech recognition", Ishimitsu, S. and Fujita, I. United States Patent 6,381,572. (1998).

Author

· Ishimitsu, Shunsuke
graduated in production system engineering from Toyohashi University of Technology, Toyohashi, Japan in 1990. He is an Associate Professor at the department of

Mechanical Engineering, University of Hyogo. His current research interests include adaptive algorithms, time-frequency analysis and noise-robust speech recognition systems.