We are IntechOpen,
the world's leading publisher of Open Access books
Built by scientists, for scientists

6,900
Open access books available

185,000
International authors and editors

200M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

Construction of a Noise-Robust Body-Conducted Speech Recognition System

Shunsuke Ishimitsu
Hiroshima City University
Japan

1. Introduction

In recent years, speech recognition systems have been used in a wide variety of environments, including internal automobile systems. Speech recognition plays a major role in a dialogue-type marine engine operation support system (Matsushita & Nagao,2001) currently under investigation. In this system, speech recognition would come from the engine room, which contains the engine apparatus, electric generator, and other equipment. Control support would also be performed within the engine room, which means that operations with a 0-dB signal-to-noise ratio (SNR) or less are required. Noise has been determined to be a portion of speech in such low SNR environments, and speech recognition rates have been remarkably low. This has prevented the introduction of recognition systems, and up till now, almost no research has been performed on speech recognition systems that operate in low SNR environments. In this chapter, we investigate 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 (Ishimitsu et al. 2001).

Since noise is not introduced into 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 speech recognition rate. However, when constructing such systems, learning data consisting of sentences that must be read a number of times is required for creation of a dictionary specialized for body-conducted speech. 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 a small number 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 number of Japanese sailors has 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 (Matsushita & Nagao,2001).

In this study, this knowledge database is accessed by speech recognition. The 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. A search is made for a response to these commands or questions speech recognition results and confirmation on the suitability of entering such commnads into the control system is made. Commands suitable for entry into the control system are speech-synthesized and output to a monitor. The speech-synthesized sounds are replayed in an ear protector/speaker unit, and while continuing communication, work can be performed while 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 a Hidden Markov Model (HMM) incidental to a database (Itabashi, 1991). 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 feature quantities and a covariance matrix Σ are used; these are expressed as follows: (Baum,1970)
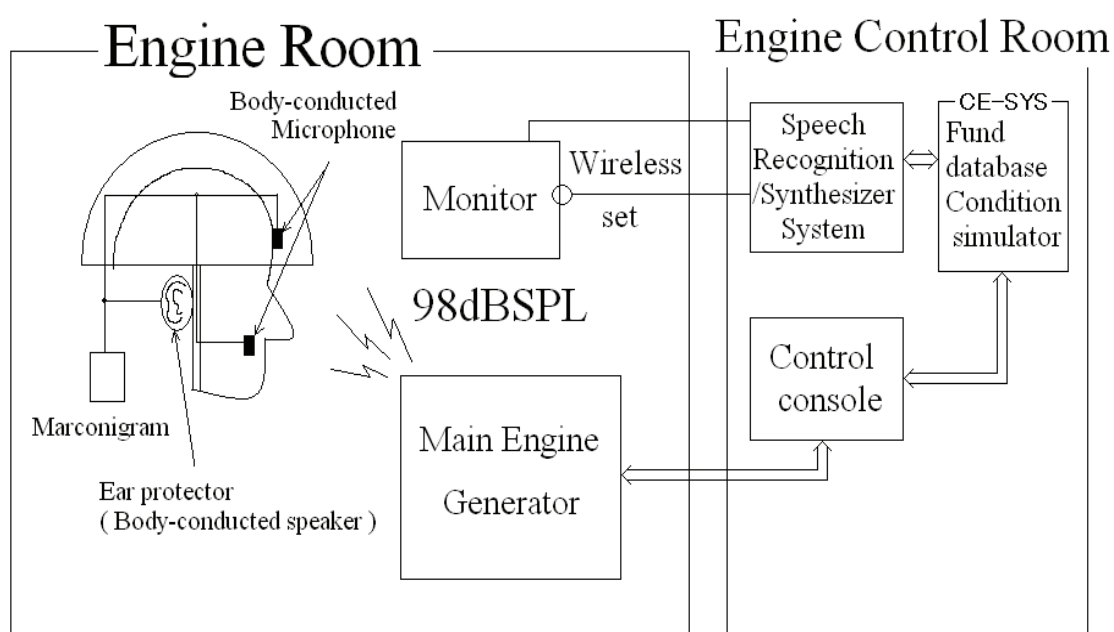

Fig. 1. Dialogue-type marine engine operation support system using body-conducted speech.

$$b(o,\mu,\Sigma) = \frac{1}{(2\pi)^{n/2} \mid \Sigma \mid^{1/2}} e^{-\frac{1}{2}(o-\mu)^t \Sigma^{-1}(o-\mu)} \tag{1}$$

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

3. Investigation into identifying sampling locations for body-conducted speech

3.1 Investigation through frequency characteristics

Figure 2 shows candidate locations for body-conducted speech during this experiment. Three locations - the lower part of the pharynx, the upper left part of the upper lip and the front part of the zygomatic arch - were selected as signal sampling locations. The lower part of the pharynx is an effective location for extracting the fundamental frequency of a voice and is often selected by electroglottograph (EGG). Since the front part of the zygomatic arch is where a ship's chief engineer has his helmet strapped to his chin, it is a meaningful location for sound-transmitting equipment. The upper left part of the upper lip is the location that was chosen by Pioneer Co., Ltd. for application of a telecommunication system in a noisy environment. This location is confirmed to have very high voice clarity (Saito et al., 2001). Figure 3 indicates the amplitude characteristics of body-conducted speech signals at each location, and also shows the difference between a body-conducted signal on the upper lip and the voice when a 20-year-old male reads "Denshikyo Chimei 100" (this is the Japan Electronics and Information Technology Industries Association (JEITA) Data Base selection of 100 locality names). Tiny accelerometers were mounted on the above-mentioned locations with medical tape. Figure 3 indicates that the amplitudes of body-conducted speech at the zygomatic arch and the pharynx are 10-20 dB lower than body-conducted speech at the upper left part of the upper lip.
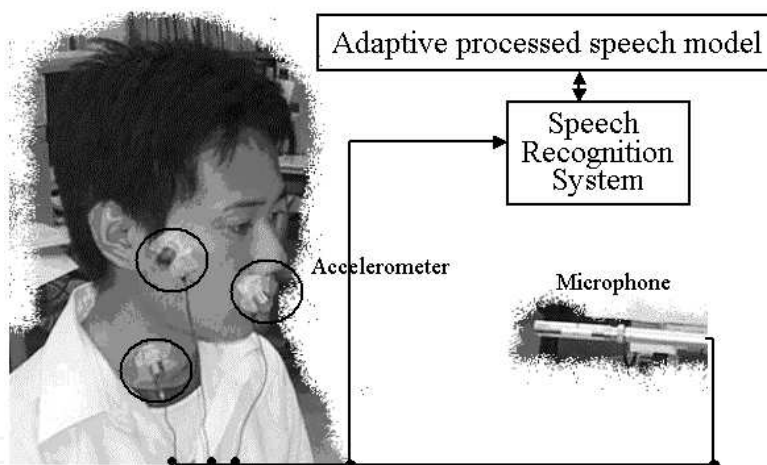

Fig. 2. Sampling location for body-conducted speech.

The clarity of vibration signals from body-conducted speech was poorer using signals from all sites except the upper left part of the upper lip in the listening experiment. Some consonant sounds that were not captured at other locations were extracted at the upper left part of the upper lip. However, compared to the speech signals shown in Figure 4, the amplitude characteristics at the upper left part of the upper lip appear to be about 10 dB lower than those of the voice.

Based on frequency characteristics, we believe that recognition of a body-conducted signal will be difficult utilizing an acoustic model built using acoustic speech signals. However, by using the upper left part of the upper lip, the site with the highest clarity signals, we think it will be possible to recognize body-conducted speech with an acoustic model built from acoustic speech using adaptive signal processing or speaker adaptation.
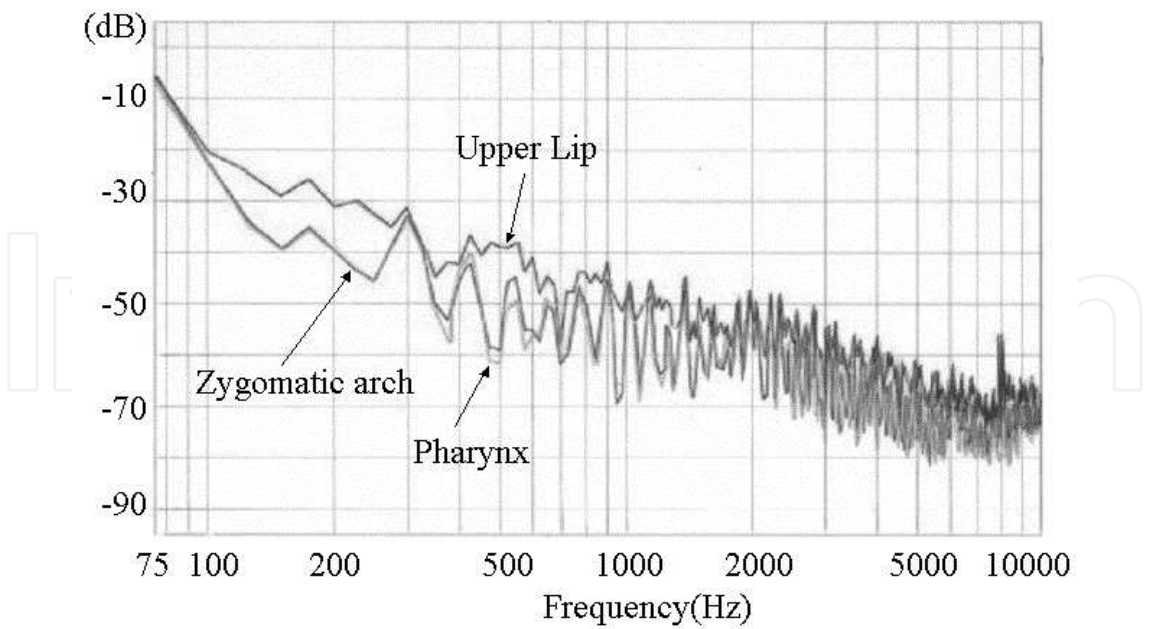
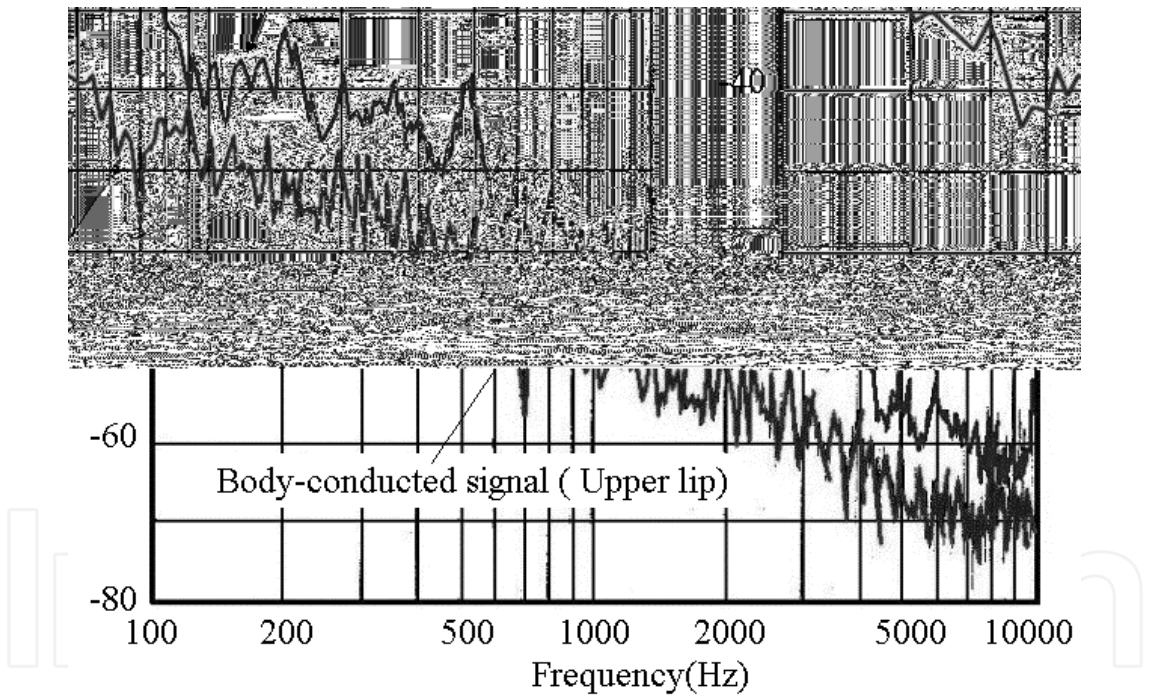Fig. 3. Frequency characteristics of body-conducted speech.

Fig. 4. Frequency characteristics of body-conducted speech and speech.

3.2 Comparison by recognition parameters

To investigate the effectiveness of a body-conducted signal model, we examined the characteristics of feature vectors. We are using LPC (Linear Predictive Coding) mel-cepstrum as the feature vectors to build an HMM. This system is widely used for parameters of speech recognition (Baum,1970). The first to the thirteenth coefficients were used as the feature vectors. The analysis conditions were: 12 kHz sampling, analysis frame length 22 msec, frame period 7 msec, analysis window hamming window.

In this study, we examined a word recognition system. To investigate the possibility of building a body-conducted speech recognition system with a speech model without building an entirely new body-conducted speech model, we compared sampling locations for body-conducted speech parameters at each location, and parameter differences amongst words. Figure 5 shows the difference on mel-cepstrum between speech and body-conducted speech at all frame averages. Body-conducted speech concentrates energy at low frequencies so that it converges on energy at lower orders like the lower part of the pharynx and the zygomatic arch, while the mel-cepstrum of signals from the upper left part of the upper lip shows a resemblance to the mel-cepstrum of speech. They have robust values at the seventh, ninth and eleventh orders and exhibit the outward form of the frequency property unevenly.
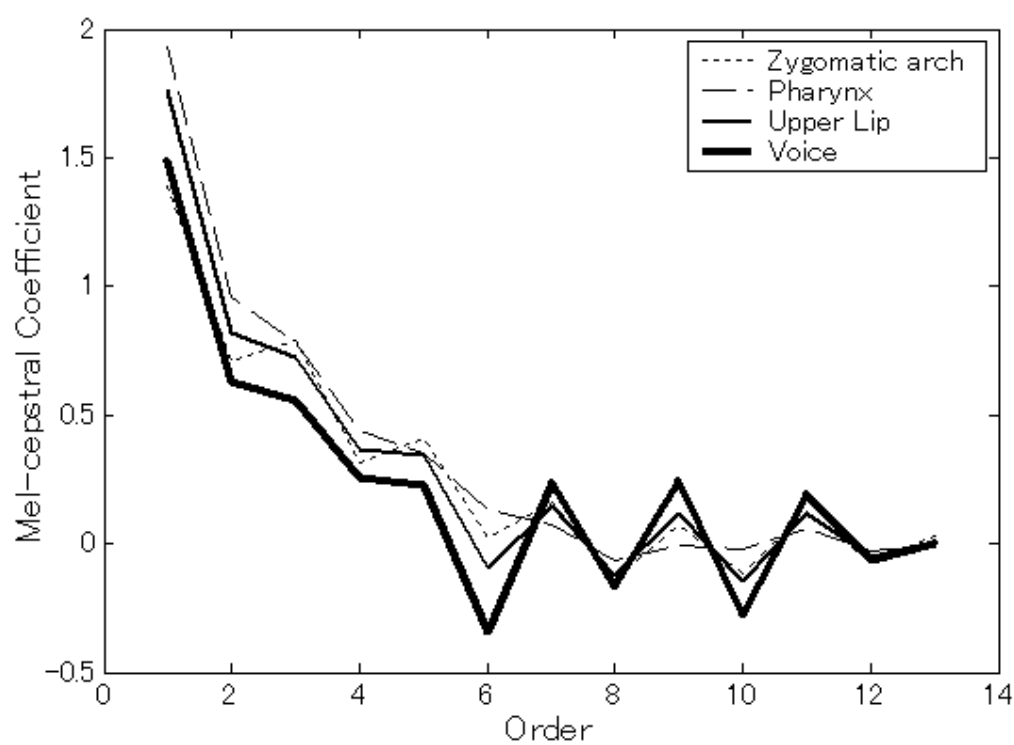

Fig. 5. Mel-cepstrum difference between speech and body-conducted speech.

Although the upper left part of the upper lip has the closest proximity to voice characteristics, it does not capture all of the characteristics of speech. This caused us to conclude that it is difficult to build a body-conducted speech model solely with a voice model.

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

4. Recognition experiments

4.1 Selection of the optimal model

The experimental conditions are shown in Table 1. For system evaluation, we used speech extracted in the following four environments:

- Speech within an otherwise silent room
- Body-conducted speech within an otherwise silent room
- Speech within the engine room of the Oshima-maru while the ship was running
- Body-conducted speech within the engine room of the Oshima-maru while the ship was running

Noise within the engine room of the Oshima-maru when the ship was running was 98 dB SPL (Sound Pressure Level), and the SNR 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.

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

Table 1. Experimental conditions

	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%

Table 2. The result of preliminary testing

Extractions from the upper left part of the upper lip were used for the body-conducted speech since the effectiveness of these signals was confirmed in previous research (Ishimitsu et al, 2001, Haramoto et al, 2001). the effectiveness of which has been 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 result of preliminary testing is shown in Table 2.

4.2 Examination of the body-conducted speech recognition system using a voice model with pretreatment

We have shown that noise-robust speech recognition is possible using body-conducted speech which spreads through the inside of the body. However, the rate of speech recognition for body-conducted speech under the same calm conditions is slightly poorer than the rate of recognition for acoustic speech.

As a result, it was determined desirable to use a dictionary that had not been through an adaptation processing to the environment with a speaker. To that end, we examined how body-conducted speech quality could be improved to that of acoustic speech quality as the next step in our experiments. Specifically, the transfer function between speech and body-conducted speech was computed with adaptation signal processing and a cross-spectral method with the aim of raising the quality of body-conducted speech to that of speech by collapsing the body-conducted speech input during the transfer function. By using this filtering as a pretreatment, we hoped to improve the articulation score and recognition rate of body-conducted speech.
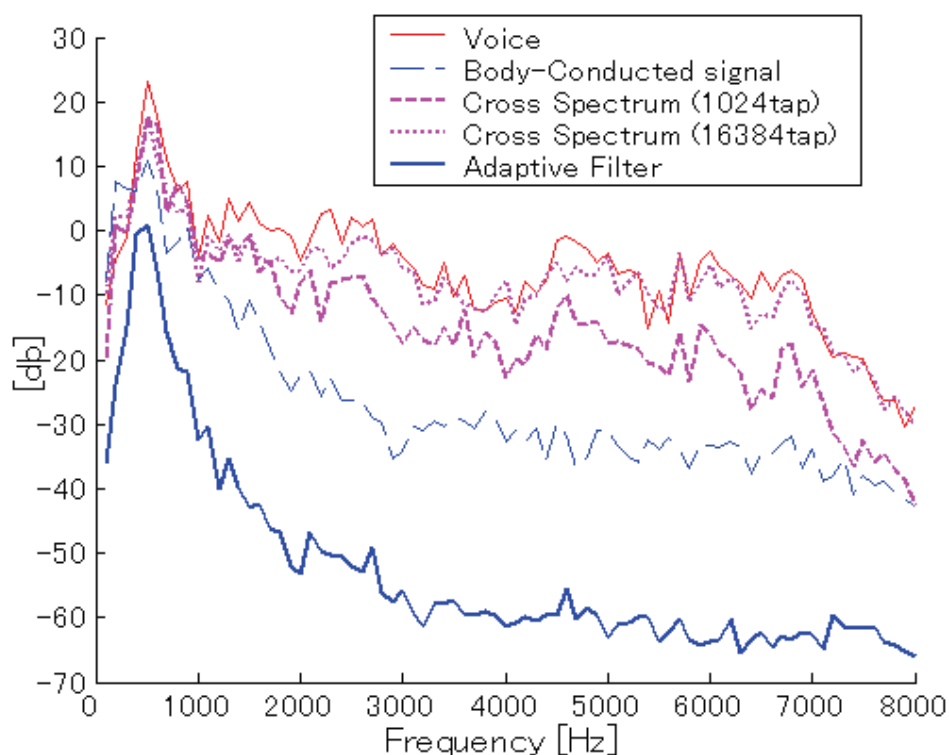

Fig. 6. The frequency characteristics of each method.

First, ten words were selected from the list of 100 place names, and then we analyzed the results using an adaptation filter and a cross-spectral method. The adaptation filter length was set to 1024, and the convergence coefficient was set to 0.01. In the cross-spectral method, the filter length was set to 1024 and 16384.

The frequency characteristics of a speech sound, a body-conducted speech sound, and the results generated by the use of an adaptation filter and a cross-spectral method are shown in Fig. 6. The characteristics of the pretreatment filter when each technique was used are shown in Fig. 7. This pretreatment filter was calculated with an adaptation algorithm using a reverse filter. The characteristics best approached the sound the speech when cross-spectral compensation was applied, and the transfer function took the form of a highpass filter. With an adaptation filter, a LMS algorithm was not able to lapse into a partial solution and the optimal solution could not be calculated this time.

Next, we describe the articulation score on reproduction; applying this pretreatment filter to body-conducted speech. With the adaptation filter, the processing result became a blurred sound. Although it seldom faded in the cross-spectral method, an echo occurred. When the

adaptation filter was applied to body-conducted speech, the results were closer when the filter length approached 16834 than when the filter length approached 1024. However, the echo also became stronger. For this reason (as a result of the speech recognition experiment by the free speech recognition software Julius) we were not able to check the predominance difference. In addition, adaptation to a speaker and environment were not taken into account in this application.
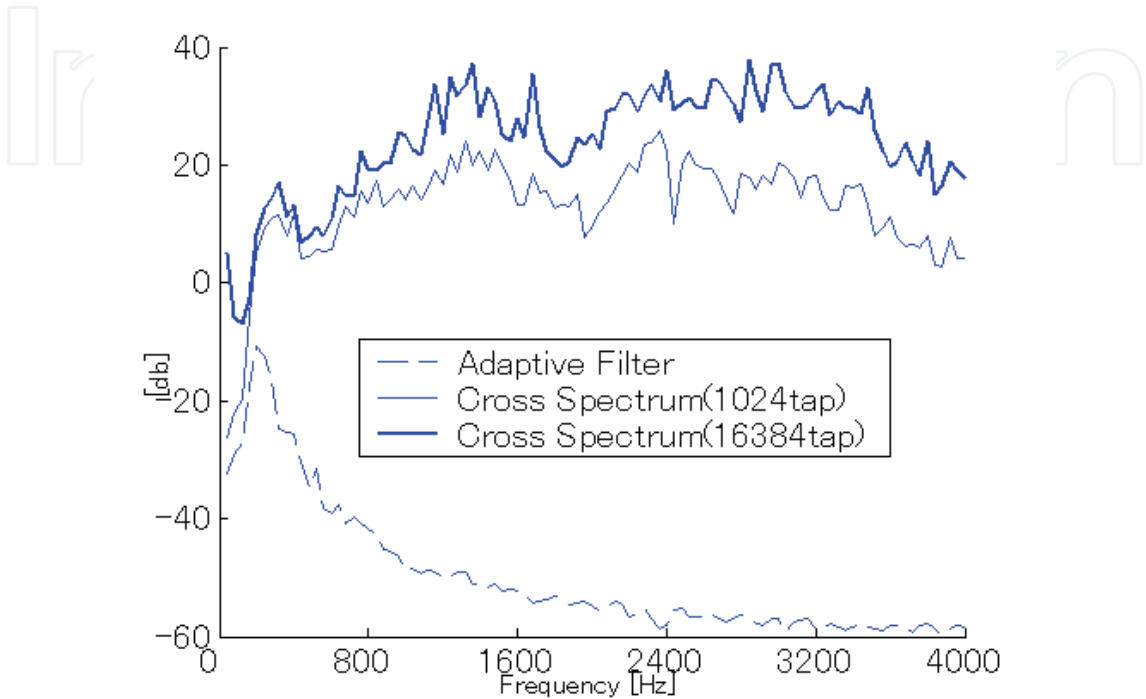

Fig. 7. The characteristics of the pretreatment filter.

Speech	Body	Filtered body
71%	57%	61%

Table 3. The result of preliminary testing with Julius

The highpass filter was designed based on the transfer function calculated with the cross-spectral method. The 16 tap FIR filter (Finite Impulse Response Filter) has a short filter length of a grade that does not generate the echo realized by the highpass filter. The recognition rate of filtered body-conducted speech is shown in Table 3. The recognition rate of body-conducted speech improved by 4% after filtering. Although the sound quality was clear and easy to hear when speech was filtered with the highpass filter, the noise in the high-frequency region was emphasized. For this reason, we concluded that the effect of the filter on the speech recognition rate was inadequate.

4.3 The effect of adaptation processing

The speech recognition test results in the cases where adaptive processing (Ishimitsu & Fujita, 1998) was performed for room interior speech and engine-room interior speech are shown in Table 4, and in Figures 8 and 9. The underlined portions show the results of the tests performed in each stated environment. 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 8, and it is thought that extraction of speech features failed because the engine room noise was louder than the speech sounds. Conversely, with room interior speech, signal adaptation was achieved. When environments for performing signal adaptation and recognition were equivalent, an improvement in the recognition rate of 27.66% was achieved, as shown in Figure 9. There was also a 12.99% improvement in the recognition rate for body-conducted speech within the room interior. However, since that recognition rate was around 20% it 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, and in Figures 10 and 11. Similar 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 obtained, as shown in Figure 10. In Figure 11. It can be observed that signal adaptation using engine-room interior body-conducted speech and speech recognition results were 95% and above, with 50% and above improvements, and that we had attained the level needed for practical usage.

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

Table 4. Result of adaptation processing with speech (%)
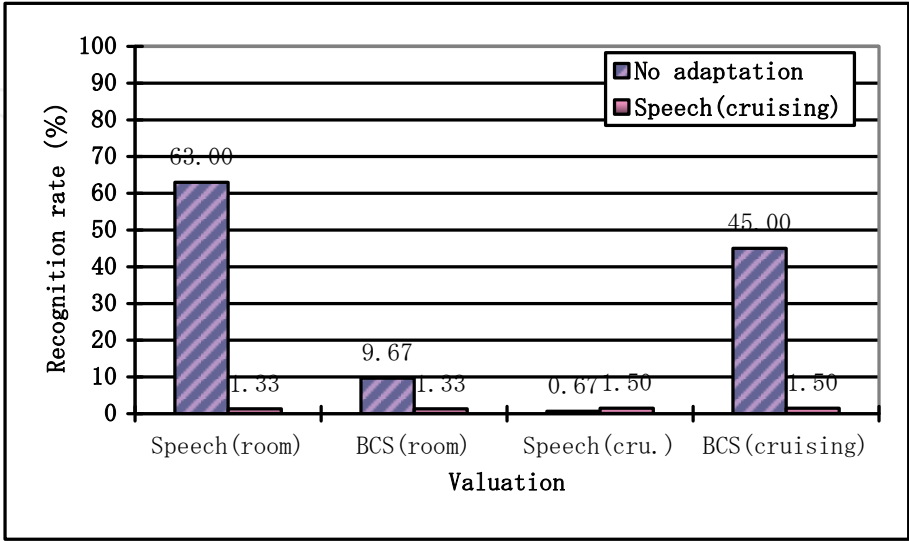
Fig. 8. Signal adaptation with speech (crusing).

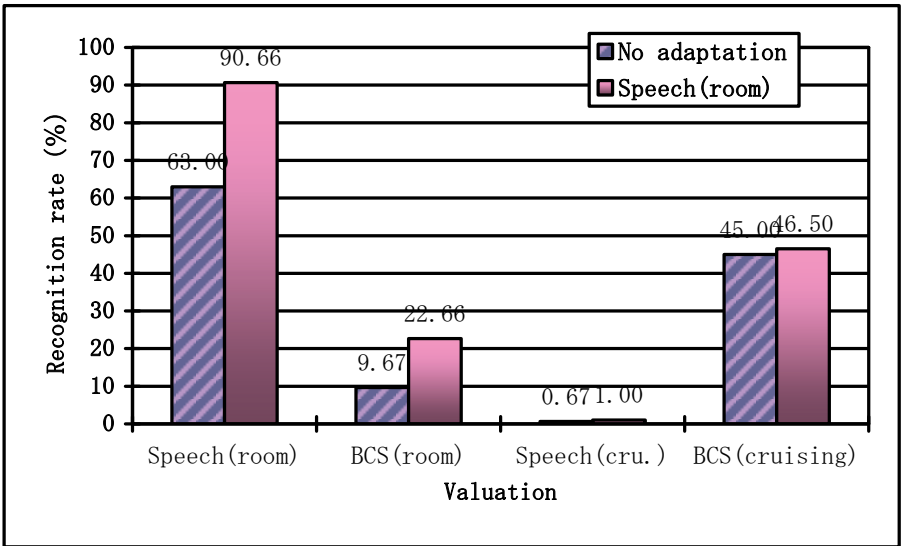| Valuation	Candidate for adaptation		
	Room	Engine Room	No adaptation
Speech(Room)	40.67	46.17	63.00
Body(Room)	**86.83**	26.83	9.67
Speech(Engine)	1.50	1.00	0.67
Body(Engine)	49.00	**95.50**	45.00

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

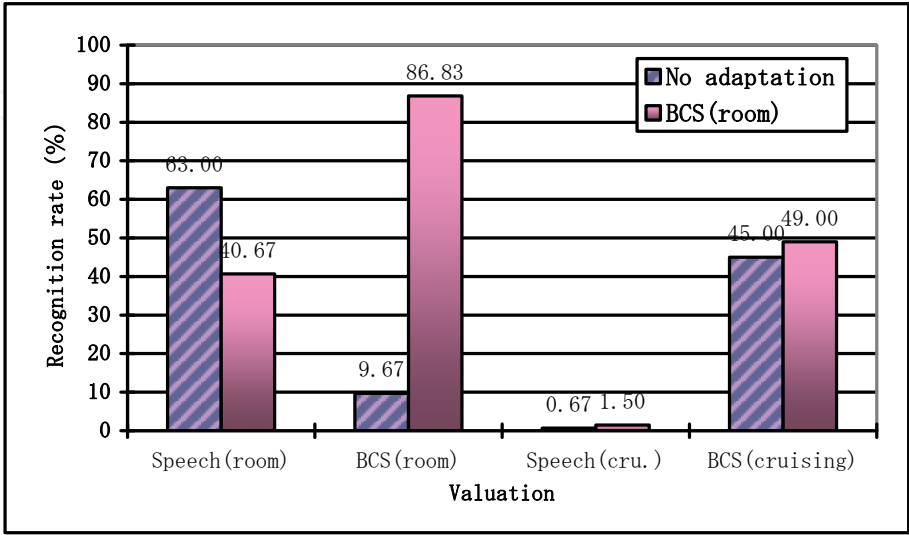Fig. 9. Signal adaptation with speech (room).

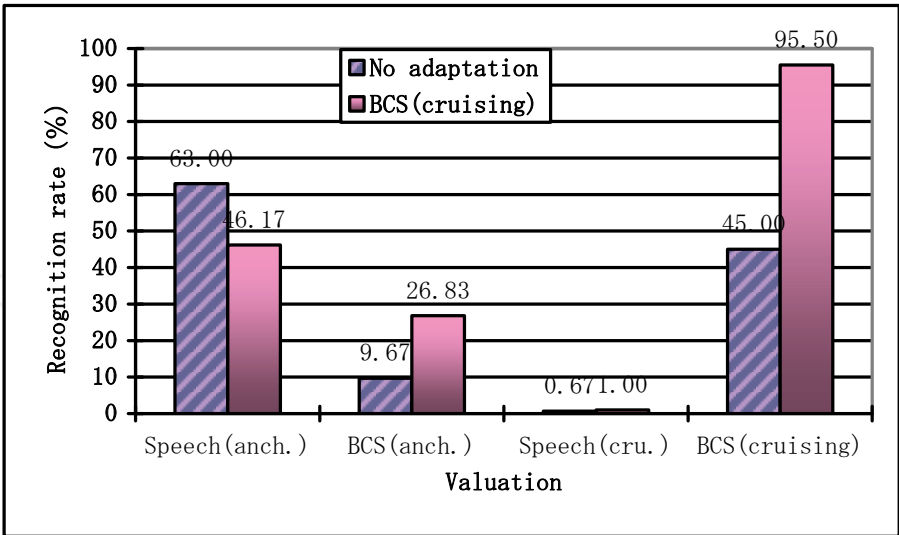Fig. 10. Signal adaptation with body-conducted speech (room).

Fig. 11. Signal adaptation with body-conducted speech (crusing).

4.4 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 6. This data consisted of 100 terms read by a male aged 20 (a different man from the person who read the terms in the earlier experiment), and the terms were spoken three times in each one environment. The man who read the terms wore a helmet during analysis of body-conducted speech and the terms were extracted from the top of his head (calvaria). The effectiveness of this position has been confirmed in documentation (Saito et al. 2001). 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 in the vicinity of 0 dB SNR. The results for this experiment are shown in Table 7. 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. This is thought to be becuase when the wireless type system was used, the noise was in the same frequency bandwidth as speech. The spectrogram analysis results of speech /hachinohe/ using cable and wireless are shown in Figures 12 and 13, respectively. From these figures it can be seen that although speech remained at less than 4000 Hz, noise overlap on the whole zone could be observed. This result suggests that it is necessary to test another method such as wireless LAN instead of a walkie-talkie.

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

Table 6. Specifications for a wireless system

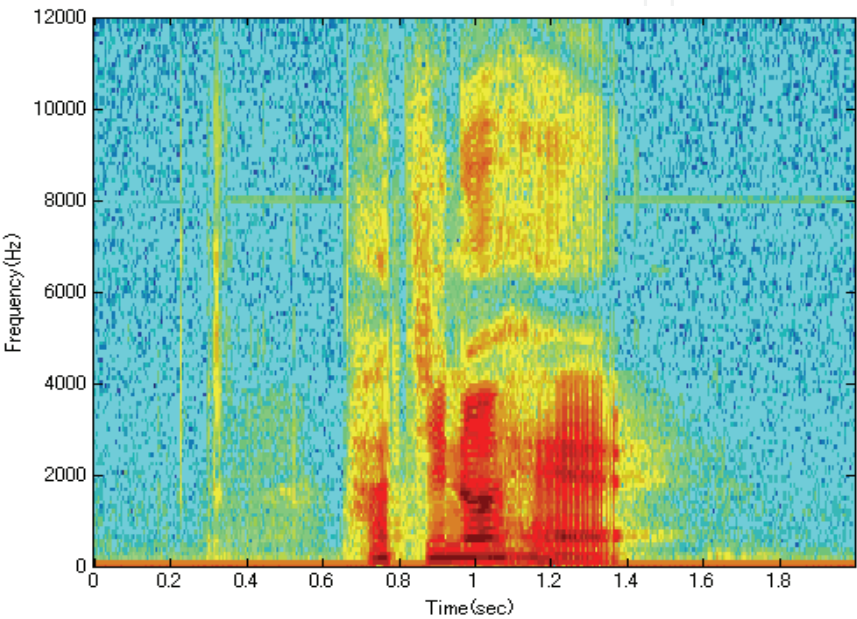| Conditions			No adaptation	adaptation
Cable	Quiet	speech	53.33	98.33
		body	43.66	97.00
wireless	Quiet	speech	3.33	77.00
		body	5.00	79.33
wireless	Noisy	speech	1.60	57.66
		body	2.00	62.00

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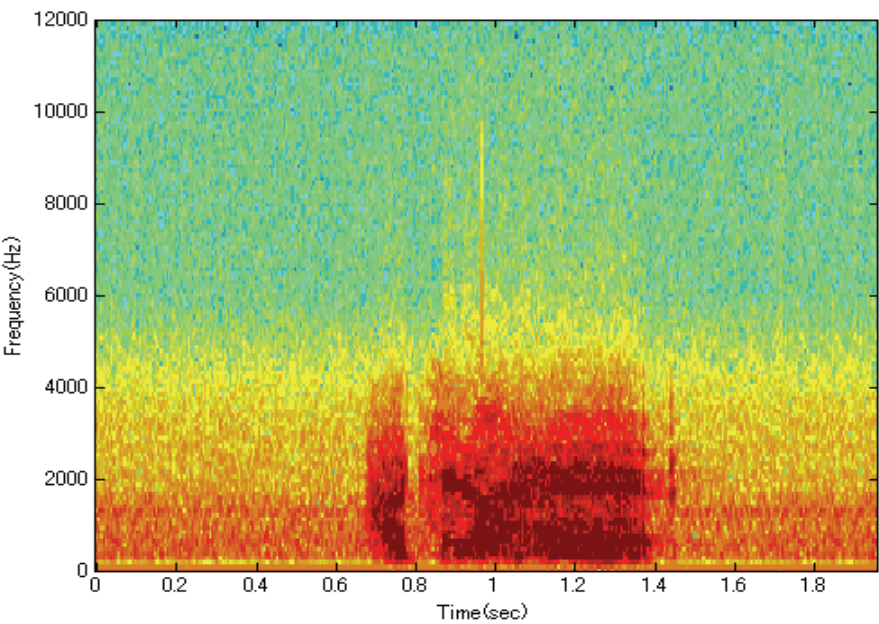

Fig. 12. /hachinohe/ with cable.

Fig. 13. /hachinohe/ with wireless.

7. Conclusion

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 noisy conditions, 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 speech recognition rate. However, in an examination by pretreatment, we could not obtain optimal results when using an adaptation filter and a cross-spectral method. We introduced an adaptive processing method and confirmed the effectiveness of adaptive processing via small repetitions of utterances. In an environment of 98 dB SPL, improvements of 50% or above of recognition rates were successfully achieved within one utterance of the learning data and speech recognition rates of 95% or higher were attained. From these results, it was confirmed that this method will be effective for establishment of the present system.

In a wireless version of the system, the results showed a worsening of recognition rates because of noise in the speech bandwidth. Even when adaptive processing was performed, a sufficient speech recognition rate could not be obtained. Although more testing of this wireless system within the actual environment of the Oshima-maru will be necessary, it will also be necessary to investigate other wireless methods.

8. References

Matsushita, K. and Nagao, K. (2001). Support system using oral communication and simulator for marine engine operation. , *Journal of Japan Institute of Marine Engineering*, Vol.36, No.6, pp.34-42, Tokyo.

Ishimitsu, S., Kitakaze, H., Tsuchibushi, Y., Takata, Y., Ishikawa, T., Saito Y., Yanagawa H. and Fukushima M. (2001). Study for constructing a recognition system using the bone conduction speech, *Proceedings of Autumn Meeting Acoustic Society of Japan* pp.203-204, Oita, October, 2001, Tokyo.

Haramoto, T. and Ishimitsu, S. (2001). Study for bone-conducted spcceh recognition system under noisy environment, *Proceedings of 31st graduated Student Mechanical Society of Japan*, pp.152, Okayama, March, 200, Hiroshima.

Saito, Y., Yanagawa, H., Ishimitsu, S., Kamura K. and Fukushima M.(2001), Improvement of the speech sound quality of the vibration pick up microphone for speech recognition under noisy environment, *Proceedings of Autumn Meeting Acoustic Society of Japan* I, pp.691～692, Oita, October, 2001, Tokyo.

Itabashi S. (1991), *Continuous speech corpus for research*, Japan Information Processing Development Center, Tokyo.

Ishimitsu, S., Nakayama M. and Murakami, Y.(2001), Study of Body-Conducted Speech Recognition for Support of Maritime Engine Operation, *Journal of Japan Institute of Marine Engineering*, Vol.39, No.4, pp.35-40, Tokyo.

Baum, L.E., Petrie, T., Soules, G. and Weiss, N. (1970), A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains, *Annals of Mathematical Statistics*, Vol.41, No.1, pp.164-171, Oxford.

Ishimitsu, S. and Fujita, I.(1998), *Method of modifying feature parameter for speech recognition*, United States Patent 6,381,572, US.

Speech Recognition

Edited by France Mihelic and Janez Zibert

ISBN 978-953-7619-29-9

Hard cover, 550 pages

Publisher InTech

Published online 01, November, 2008

Published in print edition November, 2008

Chapters in the first part of the book cover all the essential speech processing techniques for building robust, automatic speech recognition systems: the representation for speech signals and the methods for speech-features extraction, acoustic and language modeling, efficient algorithms for searching the hypothesis space, and multimodal approaches to speech recognition. The last part of the book is devoted to other speech processing applications that can use the information from automatic speech recognition for speaker identification and tracking, for prosody modeling in emotion-detection systems and in other speech processing applications that are able to operate in real-world environments, like mobile communication services and smart homes.

INTECH

open science | open minds