4

Body-Conducted Speech Recognition and its Application to Speech Support 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 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. 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.

On the other hand, people with speech disabilities face communication problems in daily conversation. They can communicate with substitute speech, but this does not have the required frequency to be readily understood in daily conversation. Therefore, we have proposed the speech support system with body-conducted speech recognition. The system retrieves speech from the body-conducted speech via a transfer function using recognition to decide on a subword sequence and the duration. Before constructing the system, we examined the effectiveness of body-conducted speech recognition for communication disorders. The first step in constructing the system involved investigating continuous word unit speech recognition, using an acoustic model not suited to body-conducted speech for communication disorders. In this study, we analyzed each parameter of these speeches and experimented with body-conducted speech recognition. We concluded that an adaptation using body-conducted speech recognition to achieve high recognition performance for disorders is valid.

Source: Advances in Speech Recognition, Book edited by: Noam R. Shabtai,
ISBN 978-953-307-097-1, pp. 164, September 2010, Sciyo, Croatia, downloaded from SCIYO.COM

2. Noise-robust body-conducted speech recognition system

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

Since the number of 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).

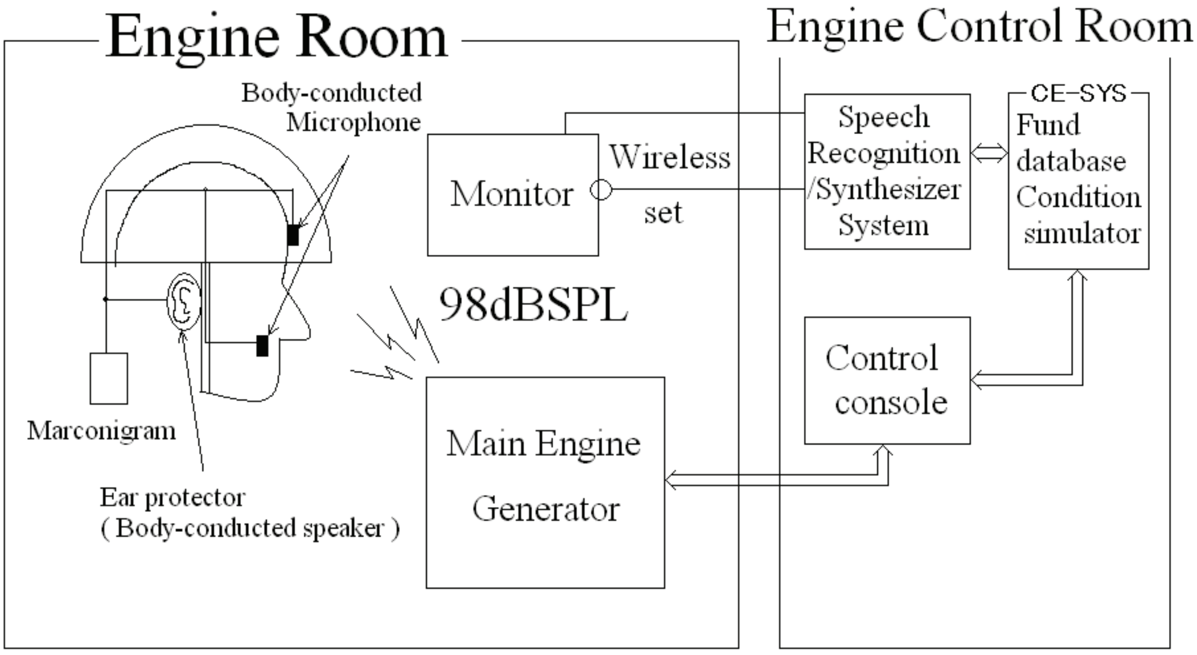

Engine Room
Body-conducted Microphone
Monitor
Wireless set
98dBSPL
Marconigram
Ear protector (Body-conducted speaker)
Main Engine Generator
Engine Control Room
Speech Recognition /Synthesizer System
CE–SYS
Fund database Condition simulator
Control console

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

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)

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

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.

2.2 Investigation into identifying sampling locations for body-conducted speech

2.2.1 Investigation through frequency characteristics

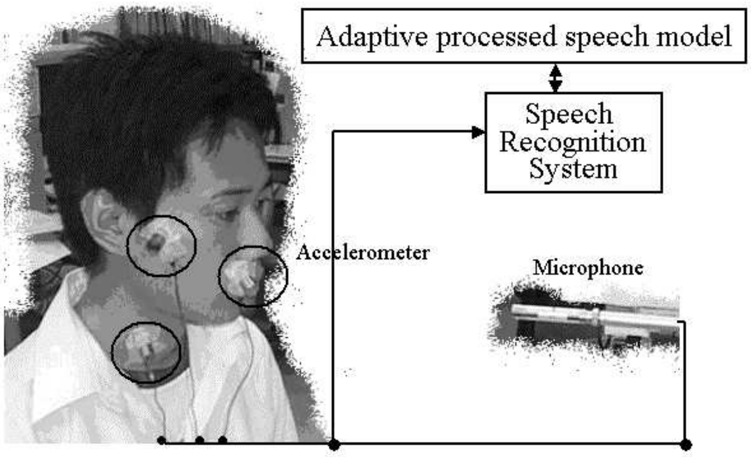


Fig. 2. Sampling location for body-conducted speech.

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. The clarity of vibration signals from body-conducted speech was poorer using signals from all sites except the upper left part of the upper lift 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.

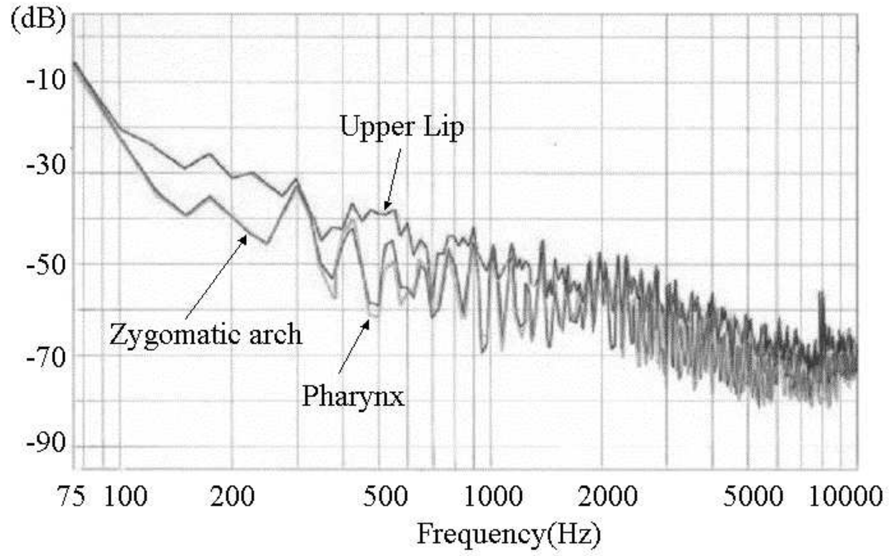


Fig. 3. Frequency characteristics of body-conducted speech.

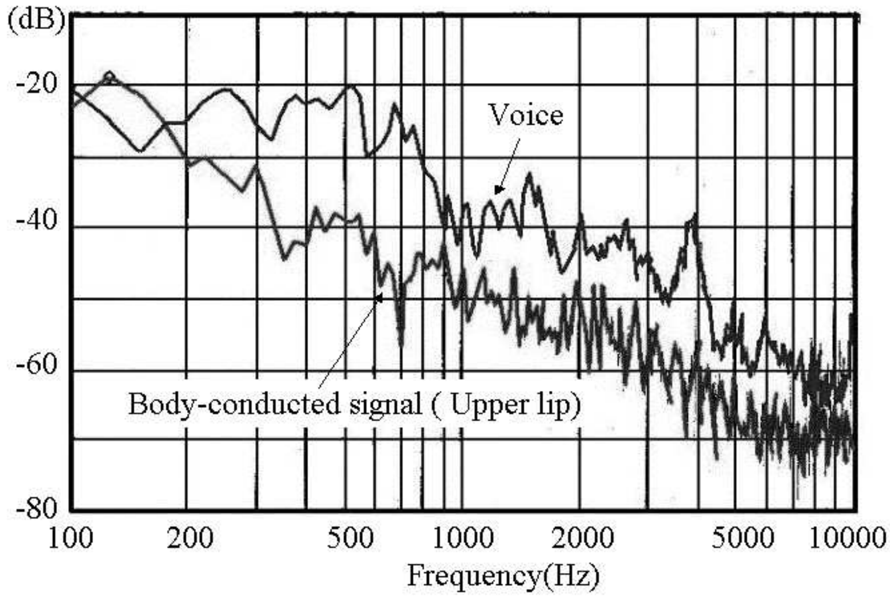


Fig. 4. Frequency characteristics of body-conducted speech and speech.

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.

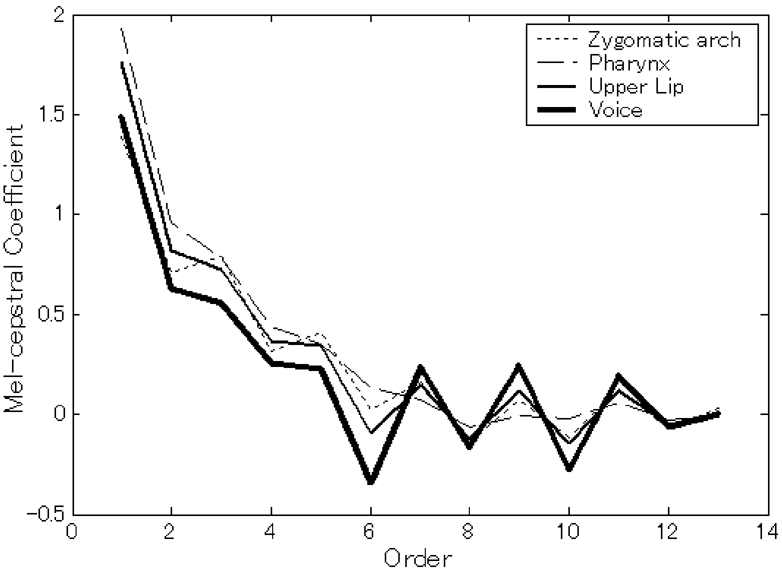


Fig. 5. Mel-cepstrum difference between speech and body-conducted speech.

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.

2.3 Recognition experiments

2.3.1 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.

2.3.2 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 3, and in Figures 6 and 7. 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 6, 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 7. 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.

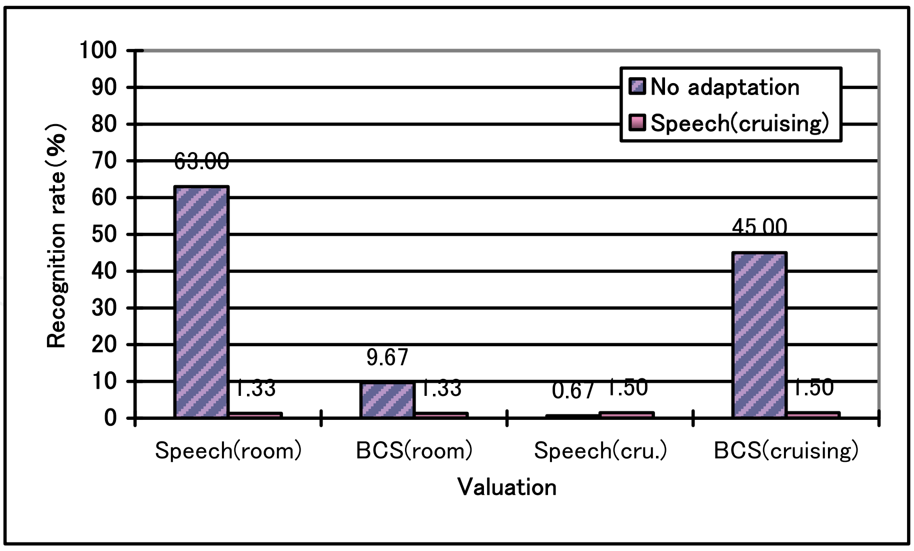


Fig. 6. Signal adaptation with speech (crusing).

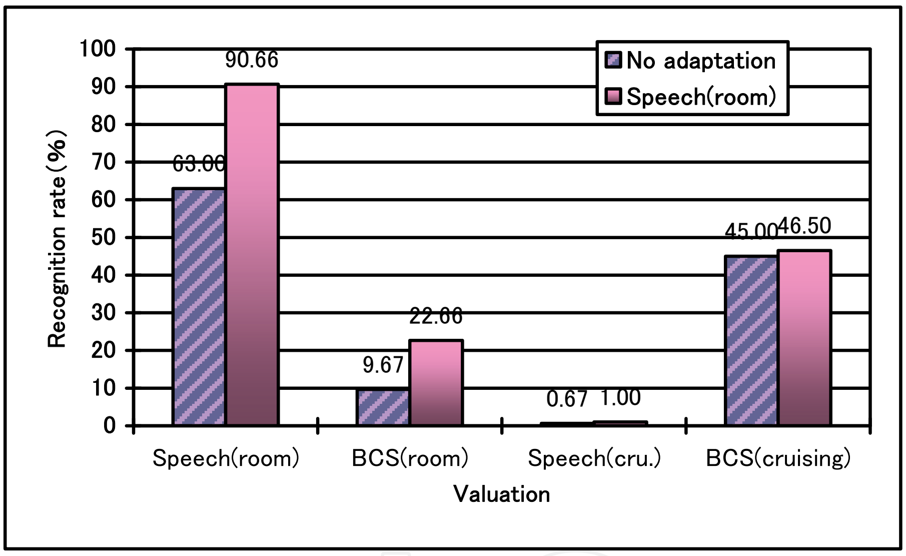


Fig. 7. Signal adaptation with speech (room).

Valuation	Candidate for adaptation		
	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 3. Result of adaptation processing with speech (%)

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 4, and in Figures 8 and 9. Similar to the case where adaptive processing was performed using speech, when the environment where adaptive processing and the environment where

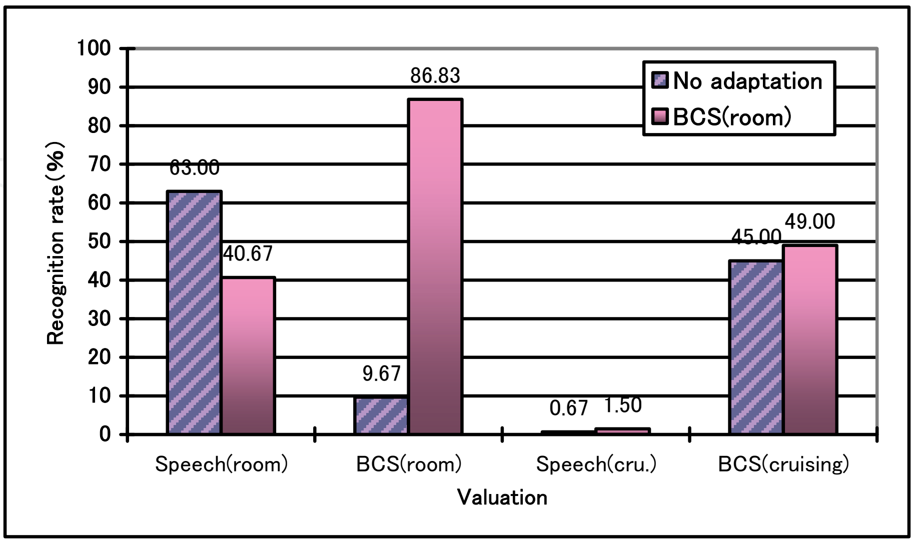


Fig. 8. Signal adaptation with body-conducted speech (room).

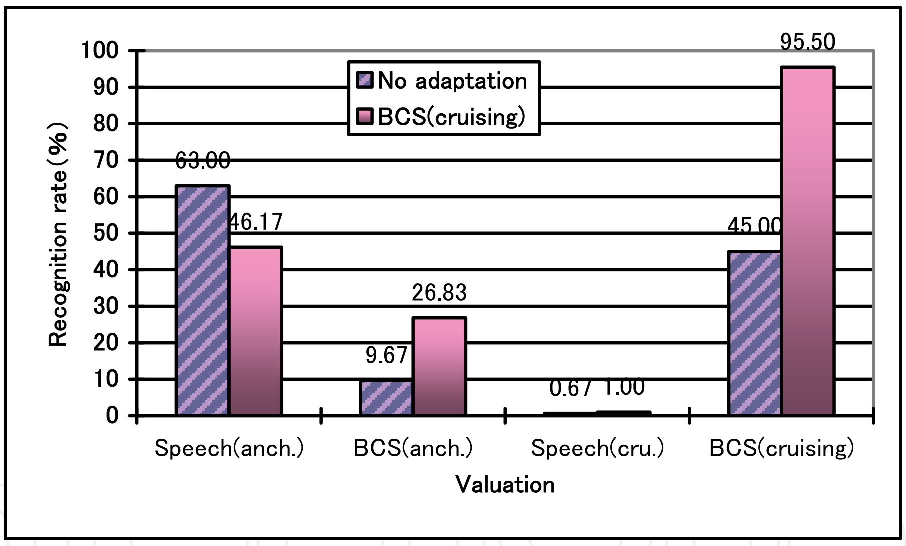


Fig. 9. Signal adaptation with body-conducted speech (crusing).

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

recognition was performed were equivalent, high recognition rates of around 90% were obtained, as shown in Figure 8. In Figure 9. 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.

3. Speech support system using body-conducted speech recognition for disorders

In late year, the number of people with disabilities that impede normal speech communication has recently increased. Pharyngeal cancer is one of the many disorders affecting such people confirmed by the increasing number of pharynx-related surgery. Although most affected patients recover well after surgery, they develop speech disorders. As a result, they have to deal with speech communication problems in their daily conversations.

The most common solution used for speech disorders is esophagus vocalization, which is inexpensive and does not require surgery. Such vocalization involves inhaling air into the stomach, and then breathing it out into the surrounding air. The new glottis in the lower pharyngeal mucous membrane then vibrates, changing air into esophageal speech through the articulation organ between the pharynx and mouth. In this way, a functionally disordered individual can generate esophageal speech. However, esophageal speech does not provide optimal fundamental frequency, high-frequency component, and power for daily conversations. Therefore, people with esophagus vocalization still have problems of communication in noisy situations encountered in daily life. Many researchers have attempted to improve the quality of esophageal speech and have looked at methods to achieve clear vocalization from body-conducted speech and the construction of speech synthesis systems. Here, we describe relevant prior research for retrieving good quality esophageal speech.

Akimoto, et al. are improved its quality retrieval on fundamental frequency (Akimoto et al., 2002). Nakamura, et al. are constructed voice conversion system using transmitted artificial speech (Nakamura et al., 2007). Ando, et al. proposed speech synthesis system for Chinese language training system (Ando & Takagi, 2007). We propose speech support system using body-conducted speech recognition for disorders. This system is able to extract a signal in a noisy environment using an accelerator.

However, conventional techniques cannot create clear speech, including the speaker's particular speech characteristics. To resolve this problem, we use continuous sub-word body-conducted speech recognition and a sub-word unit transfer function database. We propose a new solution for disorders based on a speech support system that uses bodyconducted speech recognition. Typically, the system uses body-conducted speech as the vocal chord signals, so it differs from that using the vocal chords with an impulse response to the input signal (Fukushima & Kido, 2007, Morise et al., 2007).

3.1 Proposed system

Here, we describe the speech support system using body-conducted speech recognition and sub-word transfer functions. Figure 10 shows an outline of the speech support system for disorders.

First, a disabled person makes an utterance through esophageal speech, and the system extracts body-conducted speech with an accelerator pickup. Second, the system estimates

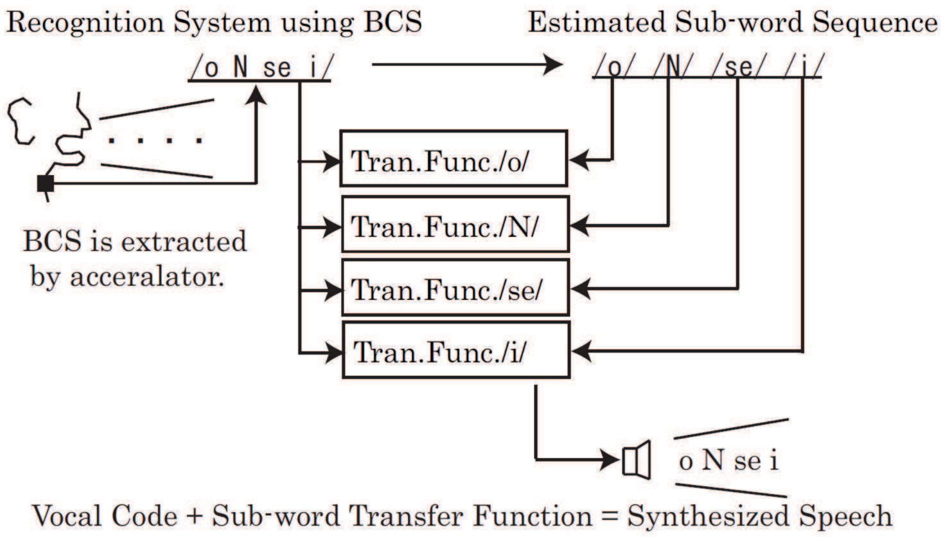

Recognition System using BCS
/o N se i/
Estimated Sub-word Sequence
/o/ /N/ /se/ /i/
BCS is extracted by acceralator.
Tran.Func./o/
Tran.Func./N/
Tran.Func./se/
Tran.Func./i/
o N se i
Vocal Code + Sub-word Transfer Function = Synthesized Speech

Fig. 10. Speech support system for disorders.

the sub-word unit sequence and its duration. Esophageal speech is then changed into recovery speech using the transfer function of the presumed sub-word unit through recognition of the output information. Finally, the system connects each recovery signal of the sub-word unit, and recreates the utterance with them.

This system has several advantages. Esophageal speech does not have sufficient volume compared with normal speech, and therefore, a speech disabled person faces a variety of problems in conversations with typical everyday noise. This becomes a problem when the conversation partner cannot hear the esophageal speech. However, with our system, even in a noisy environment, esophageal speech can be heard using body-conducted speech. Because the transfer function used by our system expresses each speaker's characteristics, the proposed system becomes a refection of each speaker. As well, because body-conducted speech is used as vocal cord signals, the signals hold linguistic informations such as fundamental frequency. When body-conducted speech is used, it is expected that the recovered speech will contain recognition errors and the system can then choose different transfer functions.

3.2.1 Advantages of the system

The system has following several advantages.

- The system works on high noisy environment
- Transfer functions has possess a robust individuality of each disorders characteristics
- The system uses vocal code user's body-conducted speech
- It is expected that the retrieved speech can approximate clear speech when recognition errors are considered.

Esophageal speech does not have sufficient volume compared with normal speech, so disabled people have a problem when conversing in noisy environments. However, this problem can be solved using body-conducted speech, since the signal can function correctly in noisy environments. Transfer functions in the system each express the individual characteristics of a user. The reason for this is explained in the next section. Moreover, using body-conducted speech as vocal chord information means that it contains linguistic information, such as the fundamental frequency and so on. Also, the recognition system can be amended when the system retrieves speech using a different transfer function.

3.2.2 Controversial issues in constructing the system

To construct the system, it has to examine following kinds.

- Effectiveness of continuous sub-word unit recognition system.
- Construction of continuous sub-word unit cross spectrum transfer function database.
- Effectiveness of the retrieved speech with respect to the frequency component and the ability to hear it.

Here, we discuss the effectiveness of the system for healthy people only. As a next step, we will construct a system for the speech disabled, which, as such, is beyond the scope of this paper.

3.3 Continuous sub-word recognition

3.3.1 Decoding algorithm of continuous sub-word recognition

Continuous sub-word unit recognition is important for body-conducted speech recognition in the system, since it is necessary to estimate each sub-word sequence and the duration times. This decoding system, constructed using the Julian/Julius tools, is known as Japanese Large Vocabulary Continuous Speech Recognition (LVCSR) (Kawahara et al., 1999). Although the Julius speech recognition engine needs a language model, our decoding system does not. Instead of a language model, our system contains a descriptive grammar. The continuous sub-word unit recognition includes the grammar, and is executed iteratively by a sound model and silent model of the mora or syllable unit. The decoding system is involved in sub-word continuous recognition. We have already demonstrated the effectiveness of body-conducted speech recognition using an acoustic model with the parameters estimated by body-conducted speech. By using this technique, the recognition system using body-conducted speech can correctly estimate a sub-word sequence and its duration.

3.3.2 Determination of signal sampling location for body-conducted speech

In a previous section, we examined signal sampling locations for body-conducted speech by comparing recognition parameters for each location. For this experiment, the upper lip was chosen as the signal sampling location for body-conducted speech. In the system, we use the pharynx as the body-conducted signal sampling location. This position is very close to the pharynx, so we expect this to be a suitable location for body-conducted speech as vocal code. If this sampling location is not suitable for executing this system, we will use the upper lip. The upper lip and pharynx have already been used effectively in isolated word recognition systems using body-conducted speech.

3.4 Construction of sub-word unit transfer function database

In this section, first, we explain fundamental transfer function between speech and body-conducted speech. Then we consider a transfer function between speech and body-conducted speech. We examine the word unit transfer function using a cross spectrum method as in previous research, however, this result is not effective since a word contains several consonants, and is complex compared with a sub-word. So we need to examine the effectiveness of several sub-word units of the retrieved speech, such as the syllable, semi-syllable and mora.

3.4.1 Relationship of transfer functions

Speech is synthesized by the vocal chords and the transfer function expressed by the oral and nasal cavities, while body-conducted speech is expressed by the body and skin. There is a relationship between the transfer functions of speech and body-conducted signals as shown in Figure 11, where disabled people are those with disorders from cancer of the pharynx, and healthy people are those that are able to utter spoken speech. The Esophagus and BCS are the utterance styles for each group, respectively. BCS means body-conducted speech while Esophageal denotes esophageal speech. In this study, we propose sub-word transfer functions that allow those using body-conducted speech to speak as healthy individuals. These transfer functions are estimated using a cross spectrum method where each signal is a sub-word.

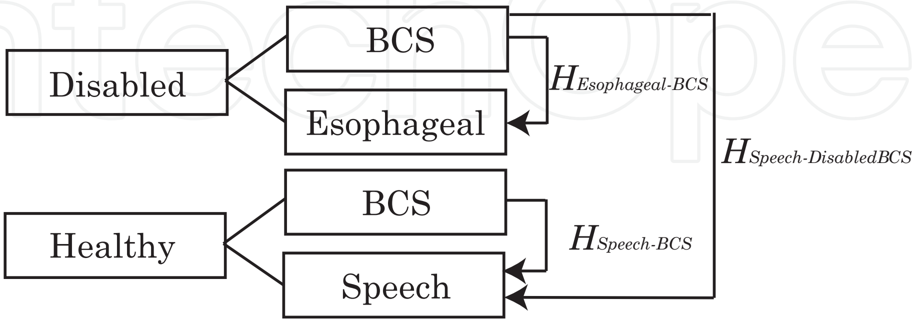


Fig. 11. Relationships of transfer functions between speech and body-conducted speech

3.4.2 Cross spectrum method

In this section, first, we will explain the basic principles of the transfer function between normal speech and body-conducted speech. Second, we describe the technique of making a sub-word unit transfer function using a cross-spectral method that makes use of speech and body-conducted speech healthy. In a previous study, we developed a word unit transfer function that used a cross-spectral method. Therefore, we investigated the validity of speech recovery with several sub-word units such as the syllable, semi-syllable, and Mora. Speech consists of a transfer function expressed as vocal cord signals, in the mouth and the nasal cavity. Moreover, as for body-conducted speech, the signals involve the body or skin. Figure 11 shows the relationship between the transfer function in speech and body-conducted speech. For every speaker, the utterance styles can be body-conducted speech body-conducted speech and esophageal. Here, we propose the use of a sub-word transfer function that converts disordered body-conducted speech into that of a healthy person. This transfer function was estimated using the cross-spectral method that makes use of each sub-word signal. Although speech from a disabled person was not available, speech sounds had previously been recorded, and our proposed system allows the recovery of these speech sounds. In the absence of any historical speech records, a transfer function is used to estimate the speech sounds from speakers such as a relative.

In applying the system, we investigated the following issues.

- Effectiveness of sub-word unit transfer functions made by cross spectrum method
- Examination for deciding sub-word unit

The system constructed for Japanese, so we examined several sub-word units.

- Phoneme
- Syllable and Semi-syllable
- Mora

Phonemes and semi-syllables are the smallest sub-word units. In pilot experiments, it was found that these do not estimate enough of each sub-word parameter of the cross spectrum transfer functions. Thus in further experiments, we examined the syllable and mora, which are longer than the other candidates. These candidates were found to estimate stable parameters for each sub-word transfer function. Because the Japanese language is constructed of several moras, we chose the mora as the unit in our system. Next, we discuss what should be used in the system as the transfer function unit. In this paper, we discuss the recognition sub-word unit and making transfer functions for context independent models only. However, the system performance is expected to improve if transfer functions can be created for context dependent models, and recognition performance should improve accordingly.

3.4.3 Transfer function database

To construct a transfer function database, we need to consider the following issues.

- An estimate of how many transfer functions need each type of signal samples
- The problem of difference phonetic contexts for each sub-word environment

The cross spectrum method expects transfer function parameters to have only one set of signals for each pair of samples. However, these transfer functions have to use all contexts of the sub-word sequence when using an acoustic model for recognition and speech retrieval. To estimate a transfer function, we use all context samples to create a transfer function database. However, as samples often contain silence at the start and end of the sample, the transfer function is not able to capture the characteristics of the frequency magnitude. This problem is discussed in the next section. As the first step in the system, we focus on context-dependent sub-word transfer functions and creating transfer functions from one pair of set signals of speech and body-conducted speech for each sub-word. We have already explained that if a context dependent transfer function is used, the techniques used in the system are significantly improved.

3.5 Investigation of the effectiveness of transfer function with speech

In this section, we examine the effectiveness of a cross spectrum method in speech retrieval. If a recognition system contains recognition errors, it does not function correctly. To investigate this problem, we divided the experiment into two cases with different experimental conditions. One system carries out recognition correctly, while the other contains errors.

3.5.1 Experimental setup for speech retrieval experiments

Speech is recorded with a microphone placed 30 cm from the speaker. Body-conducted speech is extracted with an accelerator and its amplitude is then boosted by a suitable amplifier with the accelerator position set as the upper lip. These experiments focus only on the effectiveness of speech retrieval using the proposed method. This position is best for picking up body-conducted speech clearly with an accelerator. Each signal is recorded with 16 bit, 48 kHz sampling, and then both signals are synchronized after each signal is converted from 48 kHz to 16 kHz on a computer. In the experiment, words read by a 20-year-old male are recorded by the microphone.

One of the words is ″Asahi (/a/, /sa/, /hi/)″ and it is also contained in the JEIDA database with 100 locality names. This word has several different phonetics. The system uses Julian as the recognition decoder. The purpose of this experiment is to estimate only the boundary of each sub-word, because we use Julius for supervised recognition.

The recognition system consists of a 2-stage decoder with a decoding algorithm. The first stage uses a bi-phone and 2-gram model to calculate approximately the N-best results, while the second stage calculates details of each of the N-best results using a tri-phone and 3-gram model.

Recognition errors are generated from correct results, by changing correct to fail in each subword. The following labels are examined in this experiment.

- Correct: /a/, /sa/, /hi/
- Incorrect: /hi/, /hi/, /a/

These labels are used when esophageal speech is converted to retrieved speech.

3.5.2 Investigation of speech retrieval from body-conducted speech

Here we discuss details of the results of the retrieval experiment. Figure 12 shows the speech that is extracted using the microphone, while Figure 13 shows the body-conducted speech that is picked up with the accelerator. The upper parts of the figures show wave form data while the lower parts show the corresponding spectrograms. The speech is very clear, and thus speech characteristics such as formant frequency and high resolution frequency can be found. On the contrary, the body-conducted speech does not have these characteristics and this signal is not as clear as that of the normal speech. Comparing speech and body-conducted speech, the body-conducted signal cannot capture high frequency components of 2 kHz or more, which indicates that body-conducted signals do not have any formant frequency. Therefore, the body-conducted signal is not a naturally produced signal and is a lower quality signal compared with speech signals. Figure 14 shows the retrieved speech using correct recognition results, whereas Figure 15 shows the retrieved speech using incorrect recognition results. In Figure 14, we observe frequency retrieval at 2 kHz or more and formant frequencies. Focusing on each sub-word signal, each signal represents several formant frequencies using the subword unit transfer function. For this reason, it is clear that the system is effective. In Figure 15, we see that frequency retrieval at 2 kHz is not adequate to obtain the same retrieval results compared with Figure 14. However, each recognition result is not correct, and therefore, its signal contains other signal formant frequencies.

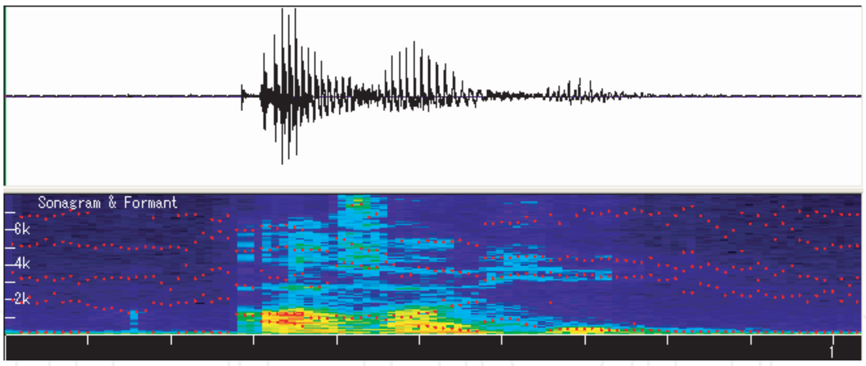


Fig. 12. Speech

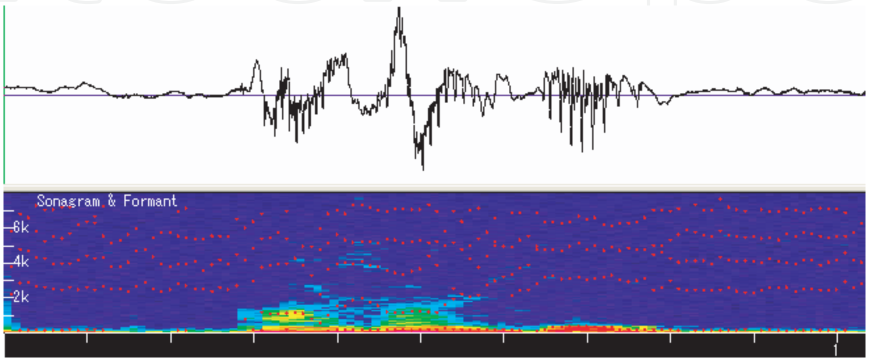


Fig. 13. Body-conducted speech

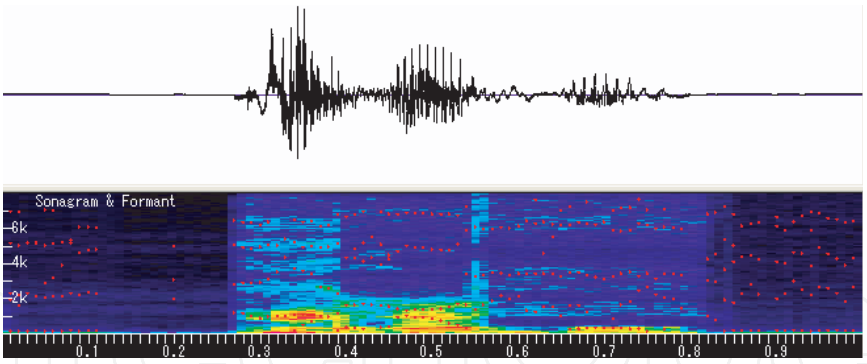


Fig. 14. Retrieved speech using correct recognition results

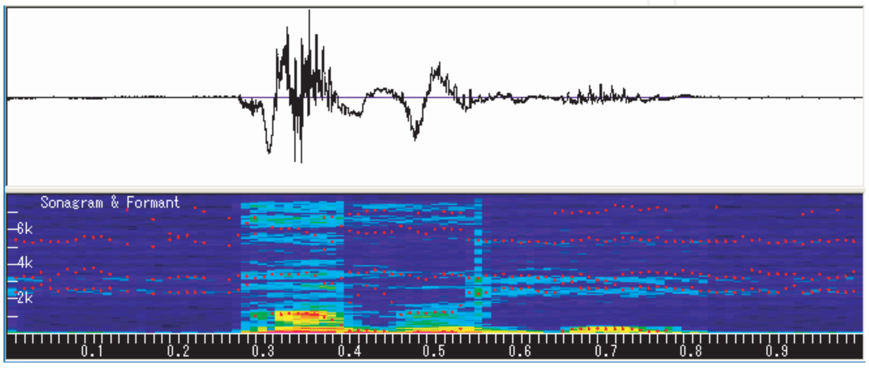


Fig. 15. Retrieved speech using incorrect recognition results with errors

4. Conclusion

First, 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. 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.

Second, we have proposed a speech support system using body-conducted speech recognition. Such a recognition system can provide people with disorders related to cancer of the pharynx with a new speech communication tool for conversation. The system consists of a body-conducted speech recognition method and a transfer function database. The recognition system provides each sub-word and its duration per sentence in speech conversation. Based on this information, the system is able to retrieve the speech using the sub-word unit transfer function. In recognizing correct and erroneous results, we confirm each signal improvement based on its waveform and spectrogram. In particular, the experiments confirmed that retrieved speech of healthy people approximates the retrieval of speech signals with high frequency and formant information. In future work, we will apply the system to those with speech disorders, and the new system will examine the possibility of a recognition system to assist disabled people with conversation and to estimate natural speech retrieval.

5. 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.

Akimoto, H., Fujii, K., Mori H., and Kasuya H.(2002), Improvement of prosody and voice quality of esophageal speech, *in IEICE Technical Report*, SP2002-94, pp.59-64.

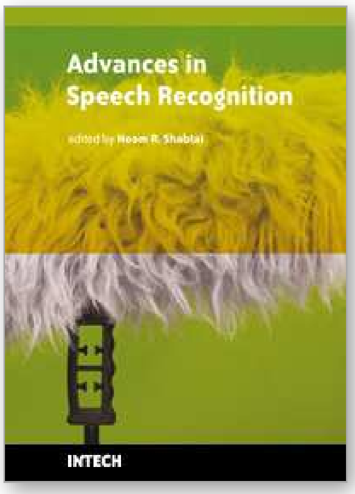
Nakamura, K., Toda, T., Saruwatari, H., and Shikano, K.(2007), A Speech Communication Aid System for Total Laryngectomees Using Voice Conversion of Body Transmitted Artificial Speech, *Journal of IEICE*, Vol.J90-D no.3, pp.780-787.

Ando, A., and Takagi, T.(2007), High-quality Speech Synthesis and Speech Processing Technology, *Journal of ICICE*, Vol.90, No.2, pp.91-94.

Fukushima, M., and Kido, K.(2007), Investigation of estimation error in impulse response by using cross spectral technique, *Journal of the ASJ*, Vol.55 N0.4, pp.265-274.

Morise, M., Irino, T., and Kawahara, H.(2007), Error Evaluation of Impulse Response Estimation by Cross Spectral Method Using Speech Signal, *Journal of IEICE*, Vol.J90-A N0.7, pp.559-566.

Kawahara, T., Lee, A., Kobayashi, T., Takeda, K., Minematsu, N., Itou, K., Ito, A., Yamamoto, M., Yamada, A., Utsuro, T., and Shikano K.(1999), Japanese Dictation Toolkit -1997 version-, *Journal of ASJ*, Vol.20 No.3, pp.233-239.



Advances in Speech Recognition
Edited by Noam Shabtai

ISBN 978-953-307-097-1
Hard cover, 164 pages
Publisher Sciyo
Published online 16, August, 2010
Published in print edition August, 2010

In the last decade, further applications of speech processing were developed, such as speaker recognition, human-machine interaction, non-English speech recognition, and non-native English speech recognition. This book addresses a few of these applications. Furthermore, major challenges that were typically ignored in previous speech recognition research, such as noise and reverberation, appear repeatedly in recent papers. I would like to sincerely thank the contributing authors, for their effort to bring their insights and perspectives on current open questions in speech recognition research.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821