ICSV13 - Vienna
The Thirteenth International Congress on Sound and Vibration
Vienna, Austria, July 2-6, 2006

INVESTIGATION OF CONSTRUCTING A NOISE-ROBUST RECOGNITION SYSTEM MAKING USE OF BODY-CONDUCTED SPEECH

Shunsuke Ishimitsu*[1]

[1]Department of Mechanical Engineering, University of Hyogo
2167, Shosha, Himeji, Hyogo, 671-2201, Japan
ishimitu@eng.u-hyogo.ac.jp

Abstract

In recent years, speech recognition systems have been introduced in a wide variety of environments such as vehicle instrumentation. Speech recognition plays an important role in ships' chief engineer systems. In such a system, speech recognition supports 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 such as engine rooms which are low SNR environments, construction of a system with a high recognition rate can be expected. However, 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 in 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.

INTRODUCTION

In recent years, speech recognition systems have been used in a wide variety of environments, including, for example, 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, the electric generator, and other equipment, and control support within the engine room is also performed. Here, 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 speech in such low SNR environments, and speech recognition rates have been remarkably low. This has prevented the

introduction of recognition systems, and up 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, and not 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, construction of a system with a high recognition rate can be expected. However, within the construction of such systems, in order to create a dictionary specialized for body-conducted speech, learning data consisting of sentences that must be read in 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.

DIALOGUE-TYPE MARINE ENGINE OPERATION SUPPORT SYSTEM USING BODY-CONDUCTED SPEECH

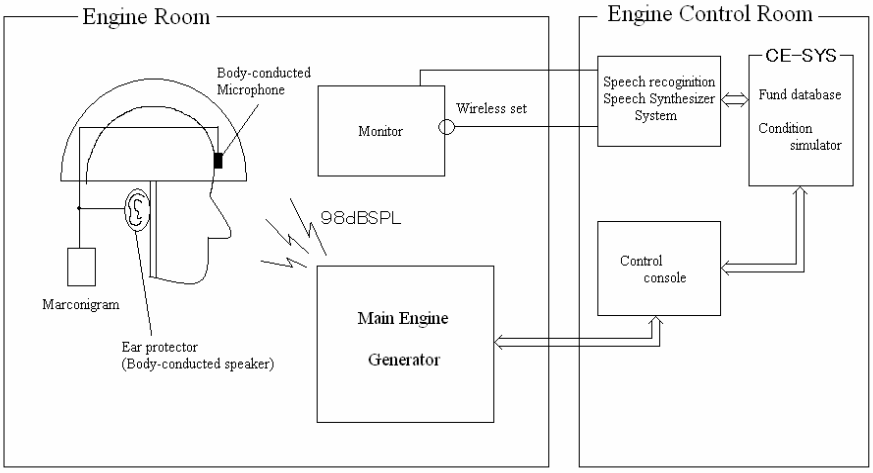

Figure 1 – Dialogue-type marine engine operation support system using body-conducted speech.

Figure 1 shows a conceptual diagram of a dialogue-type marine engine operation support system using body-conducted speech. Signals taken up with a body-conducted microphone are 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 concerning these, the 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. 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 an 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].

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

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 a few utterances, we focused only on the relearning of mean vector μ within the output probability, and thus created a user-friendly system for performing adaptive processing [7].

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 also 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.

INVESTIGATION ON IDENTIFYING SAMPLING LOCATIONS FOR BODY-CONDUCTED SPEECH

Investigation through Frequency Characteristics

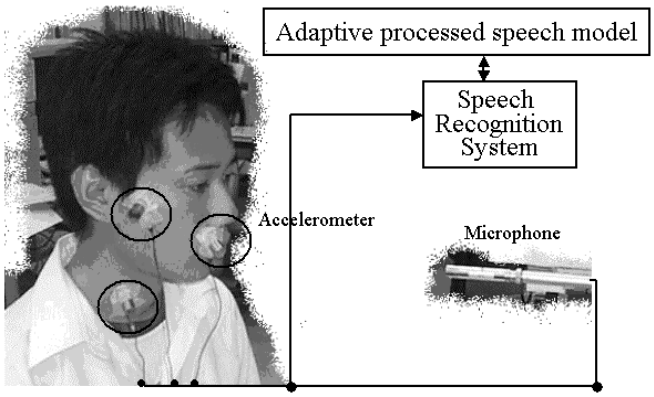

Figure 2 – Investigation on location for body-conducted speech.

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 might have 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 [4].

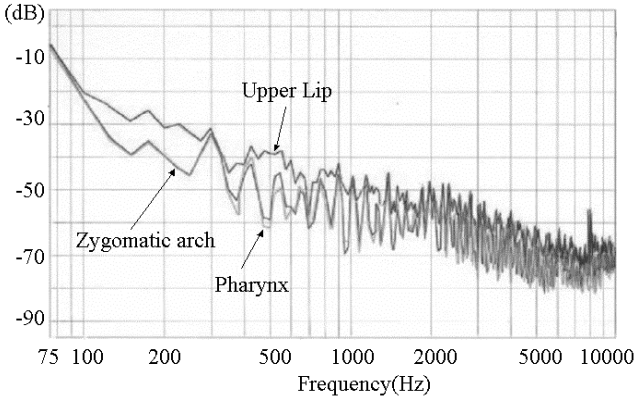

Figure 3 – Frequency characteristics of body-conducted speech.

Figure 3 indicates amplitude characteristics of body-conducted speech signals at each location, and Figure 3 shows the difference between the body-conducted signal on the upper lip and 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 amplitudes of the body-conducted speech at the zygomatic arch and the pharynx are 10 to 20 dB lower than the body-conducted speech at the upper left part of the upper lip. Also, in the listening experiment using vibration signals of the body-conducted speech, its clarity is inferior except at the upper left part of the upper lip. 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, 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 the frequency characteristics, we believe that recognition of the body-conducted signal will 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 think it will be possible to recognize the body-conducted speech with the acoustic model built from acoustic speech by using adaptive signal processing or speaker adaptation.

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 mel-cepstrum as the feature vectors to build HMM. This system is widely used for parameters of speech

recognition [6]. From the 1st to the 13th coefficients are used as the feature vector. The analysis conditions were: 12 kHz sampling, analysis frame length 22 msec, frame period 7 msec, analysis window hamming window.

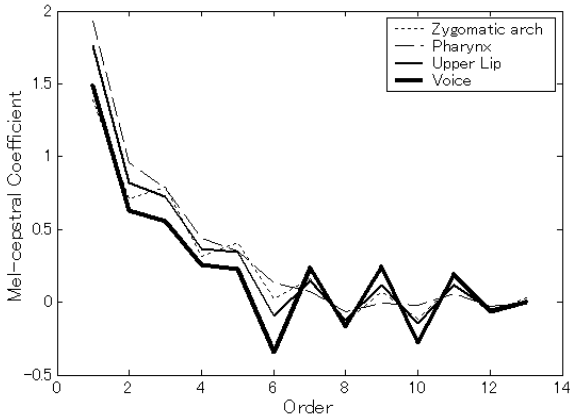

Figure 4 –Mel-cepstrum difference between speech and body-conducted speech.

In this study, we examined a word recognition system. First, 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 and parameters at each location and parameter differences amongst words. Figure 4 shows the difference on mel-cepstrum between speech and body-conducted speech at all frame averages. Body-conducted speech concentrate energy at low frequencies so that they converge 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 mel-cepstrum of speech. They have robust value at the seventh, ninth and eleventh orders and exhibit the outward form of the frequency property unevenly. Although the upper left part of the upper lip has the closest proximity to voice characteristics, it is not enough to 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 the model at the upper left of the upper lip and optimizing speech-conducted speech signals based on a voice model.

RECOGNITION EXPERIMENTS

Selection of the Optimal Model

The speech recording equipment we 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
- Body-conducted speech within the engine room of the Oshima-maru as the ship was running

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 month to about 20cm
Accelerator position	The upper left part of the upper lip

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 lip, upper left portions [3],[4], 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 3. In the speech recognition experiments, the recognition rate in anchorage dropped dramatically to 0 or 1 % as a result of the recognition environment. On the other hand, body-conducted speech in the same conditions achieved around 45 % recognition performance.

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%

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.

Underlined portions show the results of 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. This is thought to have been due to the fact that, as the engine room noise was of greater extent than the speech sounds, extraction of 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. 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, from these results, we find 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. For these results, 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 obtained. It could be observed, especially, that signal adaptation using engine-room interior body-conducted speech and recognition results were 95 % and above, with 50 % and above improvements, and that we had thus attained a level for practical usage.

Table 4 – Result of adaptation processing with speech (%)

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 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)	**86.83**	26.83	9.67
Speech(Engine)	1.50	1.00	0.67
Body(Engine)	49.00	**95.50**	45.00

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. Examinations were made for a body-conducted signal recognition system to be used in a low-SNR environment. Locations for body-conducted signals were discussed, and sample signals were derived using a recognition system. The word recognition experiment using this system resulted in far higher recognition performance at high noise levels than voice signal systems, achieving a 45% recognition rate in a -25 dB SNR environment. 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 recognition rates were successfully achieved, and recognition rates of 95 % or higher were attained. From these results, in the case of establishing the present system, it was confirmed that this method will be effective.

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 Mr. Toshikazu Yoshimi of Pioneer Corporation. We also thank Mr. Masashi Nakayama, Mr. Yasuki Murakami and students of Oshima National College of Maritime Technology who cooperated in the experiment. The work described was supported by Electric Technology Research Foundation of Chugoku, Japan.

REFERENCES

[1] Matsushita, K. and Nagao, K.,` `Support system using oral communication and simulator for marine engine operation ", Journal of Jap. Inst. Mar. Eng. Vol.36, No.6, pp.34-42. (2001)
[2] Ishimitsu, S., Kitakaze, H., Tsuchibushi, Y., Takata, Y., Ishikawa, T., Saito Y., Yanagawa H. and Fukushima M., ``Study for constructing a recognition system using the bone conduction speech", Proc. Autumn Meet. Acoust. Soc. Jpn. pp.203-204 (2001).
[3] Haramoto, T. and Ishimitsu, S. , ``Study for bone-conducted spcceh recognition system under noisy environment", Proc. 31st gra. Stu. Mech. Soc. Jpn., pp.152 (2001).
[4] Saito, Y., Yanagawa, H., Ishimitsu, S., Kamura K. and Fukushima M.,``Improvement of the speech sound quality of the vibration pick up microphone for speech recognition under noisy environment", Proc. Autumn Meet. Acoust. Soc. Jpn. I, pp.691-692 (2001).
[5] Itabashi S.,``Continuous speech corpus for research", Japan Information Processing Development Center (1991).
[6] Nakagawa S.,``Speech recognition using the probability model", IEICE (1988).
[7] Ishimitsu, S. and Fujita, I. ``Method of modifying feature parameter for speech recognition", United States Patent 6,381,572. (1998).