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

154
Countries delivered to

Our authors are among the

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

BOOK CITATION INDEX
CLARIVATE ANALYTICS
INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index
in Web of Science™ Core Collection (BKCI)

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

Improvement of Sound Quality on the Body Conducted Speech Using Differential Acceleration

Masashi Nakayama[1,3], Shunsuke Ishimitsu[2] and Seiji Nakagawa[3]
[1]Kagawa National College of Technology,
[2]Hiroshima City University
[3]National Institute of Advanced Industrial Science and Technology,
Japan

1. Introduction

During recent years, applications using speech recognition have been developed to aid dictation during lectures and to advance voice-prompted car navigation systems. Research in speech recognition has been conducted to improve recognition performance and spoken document processing (Nakagawa, 2007). However, even with developments in speech recognition technology, high recognition performance can be compromised due to noisy environments. Standard rate scales, such as CENSREC (Kitaoka et al., 2006) and AURORA (Hirsch and Pearce, 2000), are typically used for evaluating speech recognition performance in noisy environments and have shown that speech recognition rates are approximately 50–80% when under the influence of noise, demonstrating the difficulty of achieving high recognition percentages. To achieve a high recognition performance, background noise should be minimal, and normal speech should be clear, because the system estimates recognition using a feature vector from its signal. This signal can be affected by sound quality or by an utterance style due to noise in the surrounding environment.

When the noise level is low, sound quality becomes clear. However, when the noise level is high, the speech is buried in the noise, causing a change in the speaker's utterance style, termed the Lombard effect. This change causes the basic frequency to rise because a speaker does not hear the feedback sound from the ear. The method of extracting normal speech under these complex conditions because environment always changes. Several methods have been investigated to extract clear speech under these conditions, such as a noise reduction method, the use of a microphone array or a body-conducted signal. Of these, noise reduction is most commonly used for retrieving a noisy signal and can extract clear speech effectively as long as the background noise is not too high. The microphone array is typically combined with noise reduction. Body-conducted speech is a robust signal extraction method that differs from the other techniques, because it provides a solid signal that propagates through skin and bone.

Previously, we built a body-conducted speech recognition system to recognize speech in a noisy environment (98 dB SPL), specifically in the engine room of Oshima-maru, a training ship in Oshima National College of Maritime Technology (Ishimitsu et al., 2004). We found

that this system exhibited an average recognition rate of greater than 95 %. However, the recognition for body-conducted speech needs the suitable acoustic model to achieve a high recognition performance. Here, we aimed to extract a clear signal using body-conducted speech and to evaluate its efficacy by signal frequency characteristics and recognition performance. Though body-conducted speech gives a robust signal, it does not have a clear quality.

Conventional retrieval methods for body-conducted speech include Modulation Transfer Function (MTF), Linear Predictive Coefficients (LPC), direct filtering and the use of a throat microphone. However, these methods need the direct input of speech (Tamiya and Shimamura, 2006; Vu et al., 2006; Liu et al., 2004; Dupont et al., 2004). Additionally, a conventional microphone does not extract speech in a noisy environment. Thus, we proposed retrieval methods with only body-conducted speech.

2. Body-conducted speech

2.1 Characteristics of body-conducted speech

Speech is an air-conducted signal and is easy to influence with surrounding noise. In contrast, body-conducted speech is a solid propagated signal and is less influenced by noise. Figures 1 and 2 demonstrate the word, "Asahi", obtained from the database of JEIDA which contains 100 local place words (Itahashi, 1991). They were uttered by a 20-year-old male. Speech was measured 30 cm from the mouth using a microphone, and body-conducted speech was extracted from the upper lip with an accelerator. This microphone position is commonly used for a speech input of the car navigation system. The upper lip, as a signal extraction position, can provide the best cepstral coefficient characteristic as a feature vector for recognition (Ishimitsu et al., 2004). The signals were recorded at 16 kHz and 16 bits. Table 1 shows the recording environments used in this research. Speech generally provides a clear signal; however, body-conducted speech is not always clear because it lacks a high frequency component (2 kHz or more). Thus, the recognition performance can be low when its signal is used for the recognition directory.

Recorder	TEAC RD-200T
Microphone	Ono Sokki MI-1431
Microphone amplifier	Ono Sokki SR-2200
Microphone position	30cm (Between mouth and microphone)
Accelerator	Ono Sokki NP-2110
Accelerator amplifier	Ono Sokki PS-602
Accelerator position	Upper lip

Table 1. Recoding environments
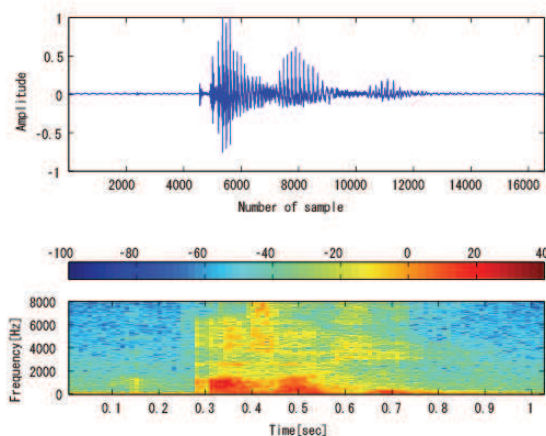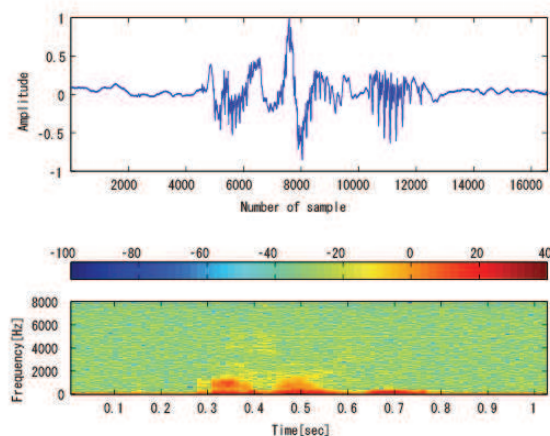
Fig. 1. Speech   Fig. 2. Body-conducted speech

2.2 Differential acceleration

Because body conducted speech does not exhibit a high frequency component, conventional retrieval techniques may also be needed to extract clear speech. However, speech is not well measured with a microphone in noisy environments. Therefore, we aimed to investigate the signal retrieval from body-conducted speech itself. To improve the sound quality, we focused on a high frequency component of 2 kHz or more, however this signal has little gain and includes an effective frequency. Subsequently, we developed for a novel signal retrieval method using differential acceleration calculated from the difference of body-conducted speech in each sample. Even with the changing of sampling frequency and speaker, it was not necessary to design a new filter because the proposed technique allowed signal retrieval with only differential procedure.

Figure 3 shows the differential acceleration signal estimated from Figure 2. Figure 3 becomes a emphasized signal however the stationary noise was included. For this reason, differential acceleration involving a retrieval signal introduces conventional noise reduction (Gong, 1995). To obtain a signal that approximates natural speech and includes effective frequencies components, conventional noise reduction techniques were employed for the differential acceleration.

3. Noise reduction for differential acceleration

3.1 Spectral subtraction method

The spectral subtraction method subtracts the spectrum of the noise sections from the average of spectrum of the noisy signal (Gong, 1995). Equations (1) and (2) describe this method.

$$x(i) = s(i) + n(i) \tag{1}$$

$$S(\omega) = \left(|X(\omega)| - |N(\omega)| \right) \exp^{j \arg X(\omega)} \tag{2}$$

It is assumed that differential acceleration $x(i)$ consists of the speech signal $s(i)$ and the noise signal $n(i)$. An estimated spectrum $S(\omega)$ can be obtained using the spectral subtraction method from Equation (2). The phase information on the input signal spectrum,

$X(\omega)$, is represented by $\arg X(\omega)$. Figure 4 shows the results from the spectral subtraction method when the filtering was repeated seven times with a setting frame width of 128 samples. The stationary noise is not removed completely. The characteristics of the high frequency component cannot be fully recovered because musical noise is produced in the signal (Nomura et al., 2006; Yamashita et al., 2005). So we concluded that it was difficult to recover frequency characteristics with this method.

3.2 Wiener filtering method

The Wiener filtering method is a technique which estimates a speech spectrum envelope from noisy speech (Li and O'Shaughnessy, 2003). The speech spectral envelope is estimated using linear prediction coefficients to obtain a clear signal that nears the frequency component of speech. The following equation describes the Wiener filtering method.

$$H_{Estimate}(\omega) = \frac{H_{Speech}(\omega)}{H_{Speech}(\omega) + H_{Noise}(\omega)} \tag{3}$$

The estimated signal spectrum, $H_{Estimate}(\omega)$, is calculated from the noisy speech, $H_{Speech}(\omega)$, and noise spectrum, $H_{Noise}(\omega)$. $H_{Estimate}(\omega)$ is expressed as a transfer function that converts a noisy signal to a clear signal. To estimate $H_{Estimate}(\omega)$, autocorrelation functions and linear prediction coefficients by the Levinson Durbin algorithm are required (Durbin, 1960). Noise spectrum, $H_{Noise}(\omega)$, is then estimated by autocorrelation functions and is used as a noise signal in differential acceleration. Figure 5 shows the result of each signal when the coefficient of both linear prediction coefficients and autocorrelation functions were equal to 1, frame width was 764 sample, and repetition was three times.

To estimate $H_{Speech}(\omega)$ and $H_{Noise}(\omega)$, the number of the linear prediction coefficients and autocorrelation function used the same number because this setting allows us to solve the problem simply and easily. The number of coefficients was changed from 1 to 32, the frame width was changed from 128 to 4,096, and the number of repetitions was also changed from 1 to 5. The best results were obtained using the conditions shown in Figure 5. We confirmed the recovery of frequency components at 2 kHz or more when using the Wiener filtering method, and musical noise was not found. Next, we compared the signal difference using spectrograms. The experimental results showed that the Wiener filtering method was suitable for differential acceleration.

4. Improvement using a combination between the proposed method and conventional method

We attempted to extract a clear signal using a combination of differential acceleration and the Wiener filtering method. We next evaluated the combination of the proposed method with various conventional methods including the cross spectrum method and the adaptive filter method. These results are shown in the following sub-sections.

4.1 Cross spectrum method

The cross spectrum method (Morise et al., 2007) was introduced using the auto spectrum of body-conducted speech, $H_{BCS}(\omega)$, and the cross spectrum between normal speech and

body-conducted speech, $H_{Speech-BCS}(\omega)$. To estimate the retrieval signal, $H_{BCS_{Estimate}}(\omega)$, the following equation was used.

$$H_{BCS_{Estimate}}(\omega) = \frac{H_{Speech-BCS}(\omega)}{H_{BCS}(\omega)} \tag{4}$$

Figure 6 shows the retrieval signal using a transfer function in a word estimated by the cross spectrum method. Sufficient recovery of the frequency characteristic was not observed when processed by the cross spectrum method. Since a transfer function is estimated by the signal in a word that contains two or more phonemes and syllables, it is difficult to retrieve sound quality. However, it is possible to improve sound quality using a transfer function for sub-word, because sub-word is a minimum unit of uttered speech (Ishimitsu et al., 2007).
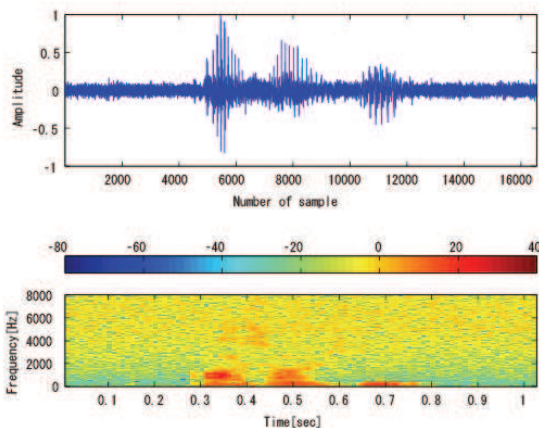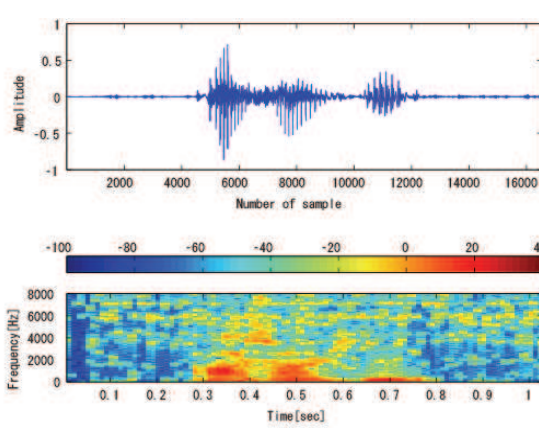

Fig. 3. Differential acceleration

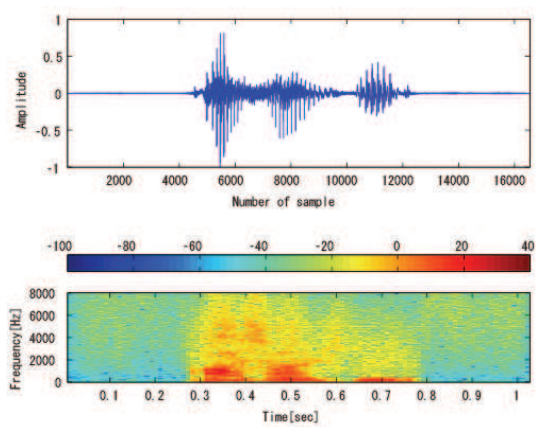Fig. 4. Spectral subtraction method
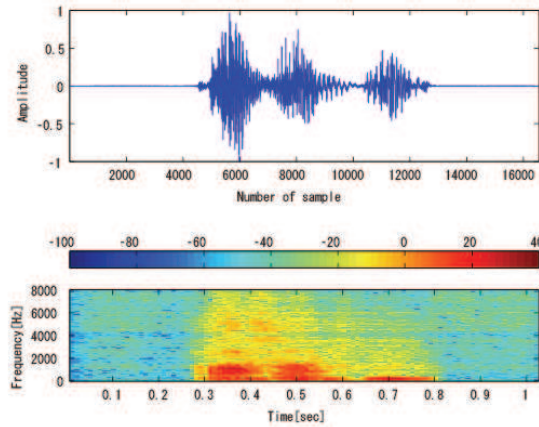

Fig. 5. Wiener filtering method
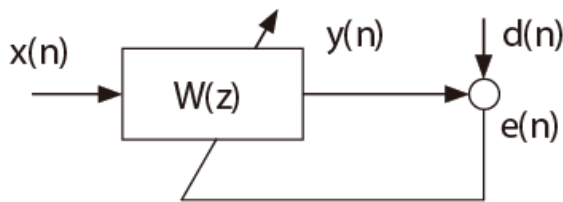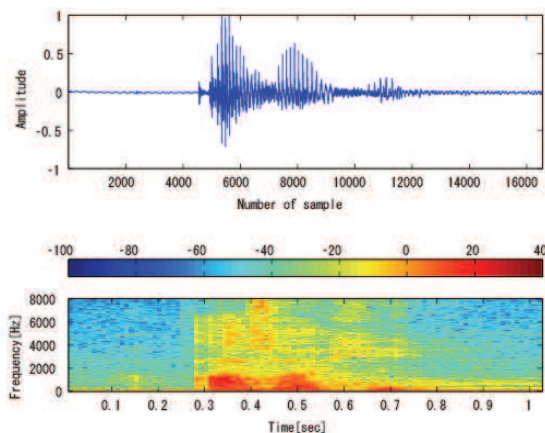

Fig. 6. Cross spectrum method

Fig. 7. LMS algorithm
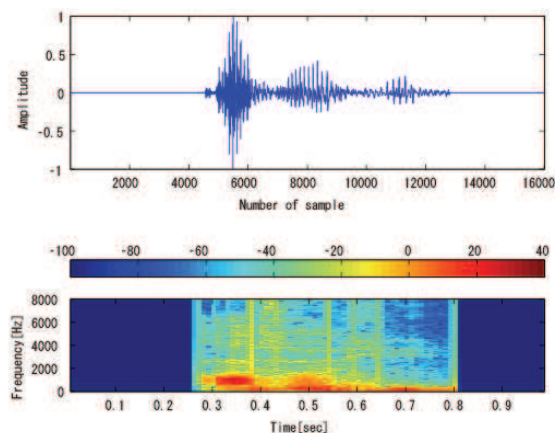
Fig. 8. Adaptive filter for word unit   Fig. 9. Adaptive filter for sub-word unit

4.2 Adaptive filter method for word unit

In previous research (Ishimitsu et al., 2004), the adaptive filter method (Haykin, 1996) was not enough to recover a sufficient frequency characteristic when body-conducted speech was used as a reference signal. The result in previous research is not improved the frequency characteristic because bodyconducted speech is low quality (Ishimitsu et al., 2004). However, since a retrieval signal using differential acceleration is clear, the estimation of clearest signal can be expected by using the retrieval signal and adaptive filter (Haykin, 1996). The updated filter method is shown in the following equation and in Figure 7.

$$w(n+1) = w(n) + \mu \cdot e(n) \cdot x(n) \tag{5}$$

where, $w(n+1)$ is new filter, μ is the convergence coefficient, and n is time index, respectively. Output signal $y(n)$ is provided by input signal $x(n)$ and current filter $w(n)$. And then, error signal $e(n)$ is calculated from the difference between the input signal and the output signal. By using an error signal, the system can be used to estimate a new filter part using Equation 5. Figure 8 shows the retrieval signal using an adaptive filter when the convergence coefficient was set to 0.01, and the filter length of taps was 16,384. Subsequently, the coefficients of the final stage of updated adaptive filter were used. Thus, a clear signal was produced using an adaptive filter. Moreover, the formant was not observed in the retrieval signal with the cross spectrum method. And it was also confirmed in the spectrogram.

4.3 Adaptive filter method for sub-word unit

The result of previous sub-section confirmed the efficacy of the proposed method. Next, we proposed the use of an adaptive filter using a sub-word unit. Previously, we constructed the speech support system that used a transfer function for sub-word to create a clear speech from body-conducted speech (Ishimitsu et al., 2007). Thus, the sound quality was improved by a transfer function for a sub-word.

Consequently, we experimented whether supervised speech recognition could be achieved by this system. The proposed system would be able to estimate the boundary of a sub-word or a word using the recognition decoder. The continuous sub-word unit recognition decoder

is made by Julian that is the Large vocabulary continuous speech recognition system (Kawahara et al., 1999; Lee et al., 2001). Our subjects are Japanese, so we choose the Mora unit as a sub-word. The Mora unit is constructed of a vowel or consonant and a vowel.

There is a problem in length of sample because the length of an adaptive filter for a sub-word is very short. So adaptive filter for sub-word unit could not be estimated by conventional methods. As a result, the adaptive filter for sub-word unit is required the long sample. The input signal can be a long sample using following equation.

$$x = iN \qquad (6)$$

where x is the length of long sample; i is the number of each sub-word sample, and N is the number of connection. Figure 9 shows the results of using a adaptive filter for a sub-word. A convergence coefficient of 0.3 and number of connection of 6 times were used. It was possible to recover the high frequency components and the formant frequencies. Additionally, since an impulsive noise was mixed at the boundary of each sub-word, the calculated results always include the errors. The adaptive filter for word unit compared with the adaptive filter for sub-word unit, so we decided that the adaptive filter for word unit is suitable for retrieval signal with differential acceleration.

5. Investigation of characteristics difference

To evaluate the efficacy of the retrieval signals, the following signals were compared with speech:

- Body-conducted speech (BCS)
- Differential acceleration with the spectral subtraction method (SS)
- Differential acceleration with the Wiener filtering (Wiener)
- Retrieval signal with Wiener filtering and the cross spectrum method (Cross)
- Retrieval signal with Wiener filtering and the adaptive filter for word (Adaptive 1)
- Retrieval signal with Wiener filtering and the adaptive filter for sub-word (Adaptive 2)

Figure 10 shows the difference between speech and body-conducted speech using a time-frequency representation. Figure 11 shows the difference between speech and the signal retrieves by the spectral subtraction method, and Figure 12 represents the difference between speech and the signal retrieves using Wiener filtering. As shown in Figure 10, there was a large difference between speech and body-conducted speech, including the formant frequencies greater than 2 kHz. However, body-conducted speech had no formant frequencies which are characteristic of the Japanese vowels. Comparing Figures 11 and 12, there was little difference in the frequency component. Particularly, as shown in Figure 12, the difference of formant frequencies was minimal. Wiener filtering was used to calculate the linear predictive coefficients, and the stationary noise was then reduced. This technique worked effectively because the predictive coefficients provided suitable parameters in each retrieval phase. Therefore, we concluded that the most suitable noise reduction method was Wiener filtering combined with differential acceleration.

Figures 13, 14 and 15 demonstrate the differences between speech and each retrieval signal that was used in combination with a conventional method. These include the retrieval signals obtained using differential acceleration calculated from the cross spectrum method or an adaptive filter for a word or for a sub-word. Each figure was compared with its original signal obtained with Wiener filtering. In the lower range of 1.5 kHz, there is a large

difference in spectrograms between speech and retrieval signal with cross spectrum method. The difference was similarly extended in Figure 15. The retrieval signal calculated from the adaptive filter for a sub-word could not extract a clear signal from the body-conducted speech because the number of samples was reduced in each sub-word duration. However, there was a marginal difference of spectrograms in all frequencies of Figure 14. Therefore, the adaptive filter for a word was confirmed as the most suitable retrieval method when it was combined with retrieval signal of differential acceleration.
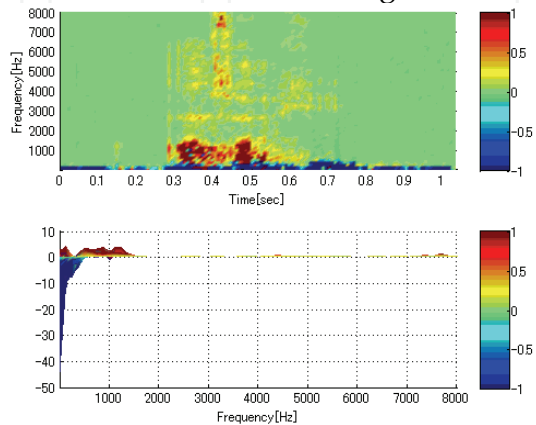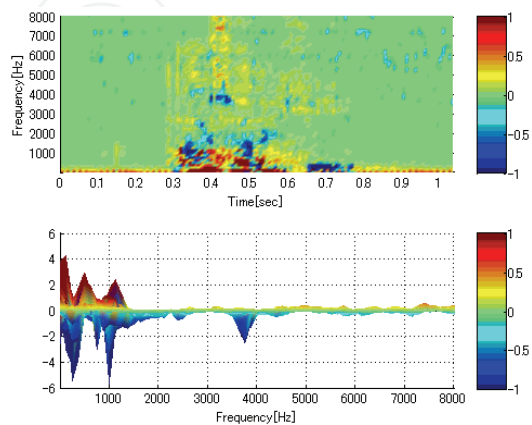

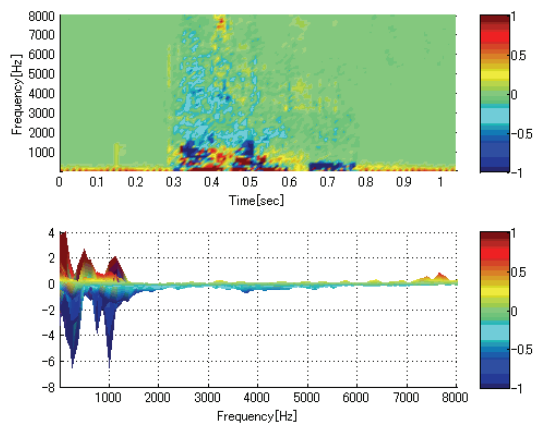Fig. 10. Difference of BCS

Fig. 11. Difference of SS

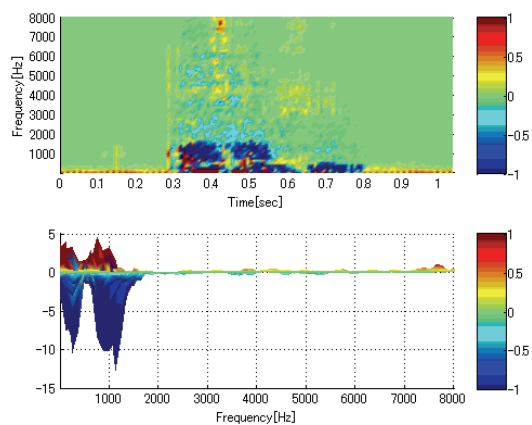Fig. 12. Difference of Wiener filtering

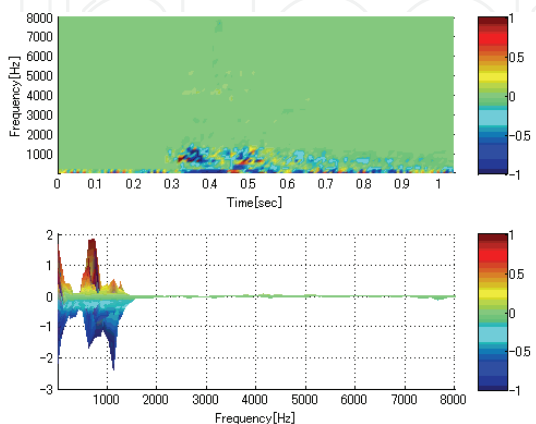Fig. 13. Difference of Cross spectrum

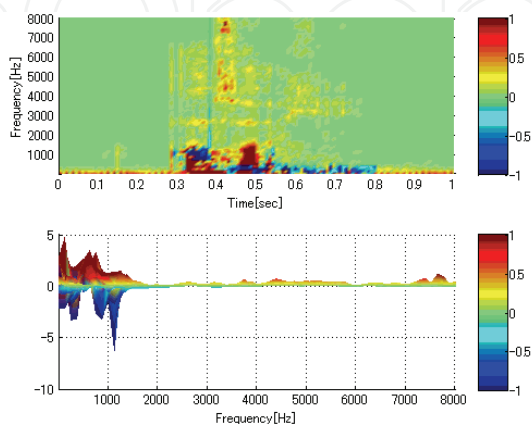Fig. 14. Difference of Adaptive 1

Fig. 15. Difference of Adaptive 2

6. Recognition experiment

Next, we evaluated the performance of the proposed method with an isolated word recognition experiment. In previous research, we constructed the body-conducted speech recognition system, which could perform in noisy environments, specifically in the engine room of a training ship (Ishimitsu et al., 2004). Here, to obtain a high recognition performance, the acoustic model needed to re-estimate the parameter using body-conducted speech. Because it is possible to estimate a speech from body-conducted speech, it expects that the retrieval signal can be directory used to speech recognition. The recognition performance can be improved when its signal becomes a clear speech. Therefore, to perform an objective experiment with statistical analysis, the recognition rate of isolated word recognition was evaluated using an acoustic model for unspecified speakers built with a speech. Generally, a subjective listening experiment is often used to evaluate the sound quality however it needs many people and much data as a admitting result.

Speech recognition was matched with the feature vector of each signal. The model parameters were able to approximate a natural speech when the recognition rate compared with body-conducted speech to each retrieval signal. Additionally, because HMM is an acoustic model that evaluates each feature vector using the output of a multi-dimensional normal distribution, it can be evaluated statistically. Thus, HMM can confirm whether the speech nearness characteristic is acquired at the feature parameter level by comparing word recognition rates.

Table 2 shows the experimental environments for isolated word recognition. The recognition decoder, Julius (Kawahara et al., 1999; Lee et al., 2001), was used in this experiment. Because Julius is a decoder for large vocabulary continuous speech recognition, it can be changed into isolated word recognition. Thus, it became possible to recognize words without a language model. JEIDA 100 local place words were used as candidates for word recognition (Itahashi, 1991). The words consisted of a database of Japanese place names after consideration of phoneme balance. The acoustic model was the context-dependent type tri-phone model in a word for unspecified speakers. The differential acceleration was processed with the parameter used in the previous chapter. The candidates for recognition in this experiment included each of the following signals:

- Speech : Speech
- BCS : Body-conducted speech
- ret. BCS : Retrieval signal with Wiener filtering

The body-conducted speech was compared with the retrieval signal from the recognition experiment. It was expected that the recognition performance would improve if the signal approximated natural speech from the body-conducted speech. Tables 3 - 5 show the recognition rate in each speaker. And Table 6 shows the average of all speakers. There was an improvement of 3 - 9 % in Speakers B and C but little improvement in Speaker A. About 5 % of the improvement was obtained through average of the recognition rate. From Table 6, the retrieval signals with differential acceleration become clear signals compare to original body-conducted speeches. Using an adaptation technique to estimate the new parameter in an acoustic model, the recognition performance increased to greater than 95 %. Here, we focused on the investigation of the signal retrieval, so the feature vectors in the acoustic models did not need new parameters in this experiment. Though the improvement was

marginal, this result demonstrated the effectiveness for signal retrieval without speech. We expect that these recognition rates can be greatly improved by fine-tuning parameters for each speaker using an adaptation technique.

Speaker	two 20 and one 37 years old male
Number of data sets	100 words × 3 set/person
Vocabulary	JEIDA 100 local place names
Recognition system	Julius-3.4
Acoustic model	gender dependent triphone model
Model conditions	16 mixture gaussian, clustered 3000 states
Feature vectors	MFCC(12)+ΔMFCC(12)+ΔPow(1)=25 dim.
Training condition	20,000 sample and more (Itou et al., 1999), HTK 2.0 (Young et al.,2000)

Table 2. Experimental environments

Signal type	Set 1	Set 2	Set 3	Average
Speech	90%	90%	91%	90.3%
BCS	63%	56%	61%	60.0%
ret. BCS	62%	57%	63%	60.7%

Table 3. Speaker A (20 years old male)

Signal type	Set 1	Set 2	Set 3	Average
Speech	93%	94%	92%	93.0%
BCS	53%	50%	48%	50.3%
ret. BCS	63%	57%	58%	59.3%

Table 4. Speaker B (20 years old male)

Signal type	Set 1	Set 2	Set 3	Average
Speech	92%	94%	92%	92.7%
BCS	60%	68%	61%	63.0%
ret. BCS	65%	68%	65%	66.0%

Table 5. Speaker C (37 years old male)

Signal type	Average
Speech	92.0%
BCS	57.8%
ret. BCS	62.0%

Table 6. Recognition results of all speakers

7. Investigation of body-conducted speech in a noisy environment

7.1 Signal recording in a noisy environment

In order to acquire, a noisy environment, we used the engine room of the 'Oshima-maru' training ship from the Oshima National College of Technology, Japan. Noise within the engine room, under the two conditions of anchorage and cruising, were 93 and 98 dB SPL, respectively, and the SNR measurements from microphone. There was –20 and –25 dB SNR, respectively. The signals of 100 words, from the database of JEIDA were read three times in each environment by three males aged 20, 20 and 37 years old. For body-conducted speech extraction, measurements were taken from the upper lip. In this study, we experimented under anchorage condition to estimate retrieval signals.

Figures 16 and 17 demonstrate the word "Ageo" that was obtained from the database of JEIDA. Figures 16 and 17 exhibit speech and body-conducted speech signals from the engine room of the Oshima-maru. Because body-conducted speech is a structure bone sound, it is less influenced by noise than normal speech. Unlike body-conducted speech, normal speech signals do not detect the utterance of speakers. Comparing Figures 2 and 17, there was little difference of frequency characteristic between quiet and noisy environment, so it expects that the proposed method for signal retrieval can apply to the signal in noisy environment.
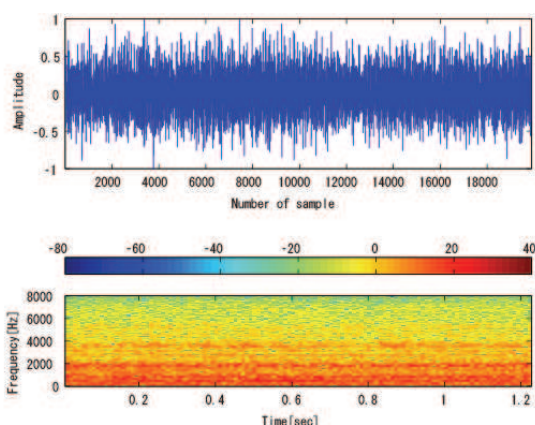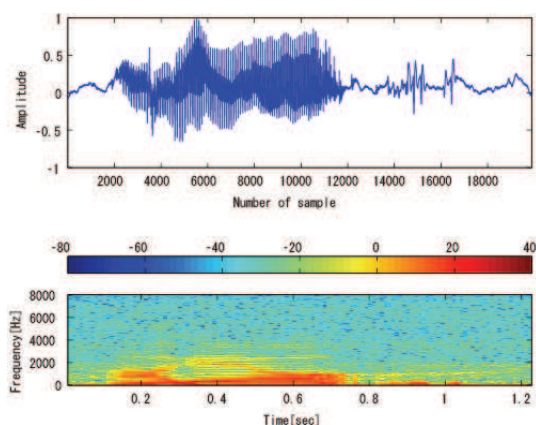

Fig. 16. Speech in noise   Fig. 17. Body-conducted speech in noise

7.2 Differential acceleration in a noisy environment

The signals can be a clear with same setting of signal retrieval in quiet room with proposed method. Difference in signals characteristics between speech and body-conducted speech are little because that reason is shown a difference of Figures 2 and 17. Thus, we also estimate the differential acceleration directory which the difference between the signals is calculated. Figure 18 shows the differential acceleration estimated from Figure 17. Body-conducted speech was extracted with an accelerator from the noisy environment. The signal level in each frequency is low compares to signal in quiet environment. Therefore, differential acceleration in the noisy environment exhibited a clearer signal compared with that extracted in the quiet environment.

7.3 Signal retrieval for body-conducted speech in a noisy environment

The signal characteristic of body-conducted speech in a noisy environment is not affected by noise; however, the basic frequency of the signal rises by the Lombard effect. Generally,

since the decoder only uses spectral envelopes as recognition parameters, the problem is a matter of no importance.

Figures 19, 20 and 21 show the retrieval signals using the proposed methods that differ in the number of Wiener filtering repetitions. The signals exhibited frequency characteristics of 2 kHz and more. The estimated signal rejected stationary noise from the differential acceleration completely. The effect of repetitions was also clearly observable. We found that three repetitions allowed stationary noise to be completely reduced. In this study, parameter settings in Wiener filtering were the same as for a quiet room because difference in signal characteristic between quiet and noise is little. From Figure 21, we expect that performance of speech recognition can also be improved.

8. Decoding algorithm for differential acceleration

8.1 Problem of recognition for differential acceleration

The differential acceleration exhibited a distorted signal when a consonant was present. Consonants were removed together with the stationary noise because its signal level is low. However, the signal levels of vowels are high because of the formant frequencies. Thus, vowels were kept intact. The stationary noise level changed according to the environment of
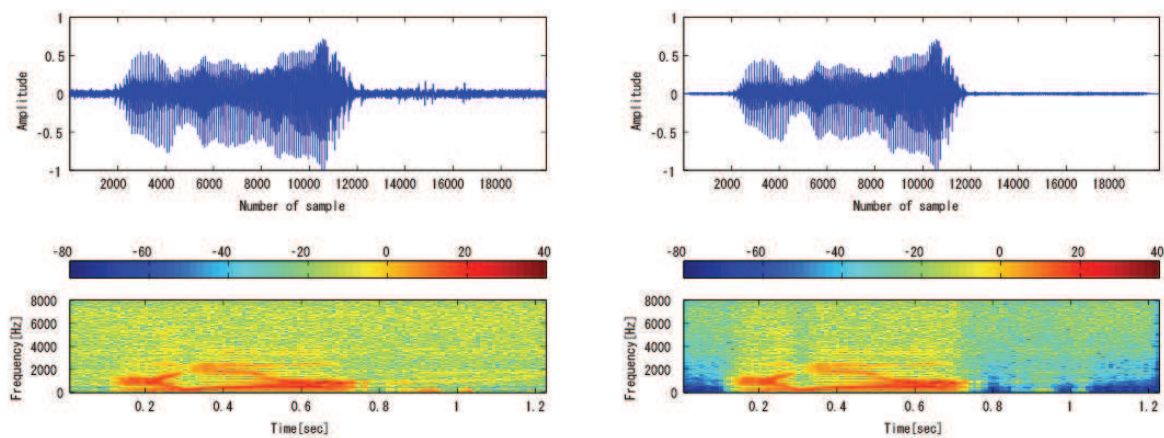

Fig. 18. Differential acceleration in noise   Fig. 19. Retrieval signal with Wiener 1
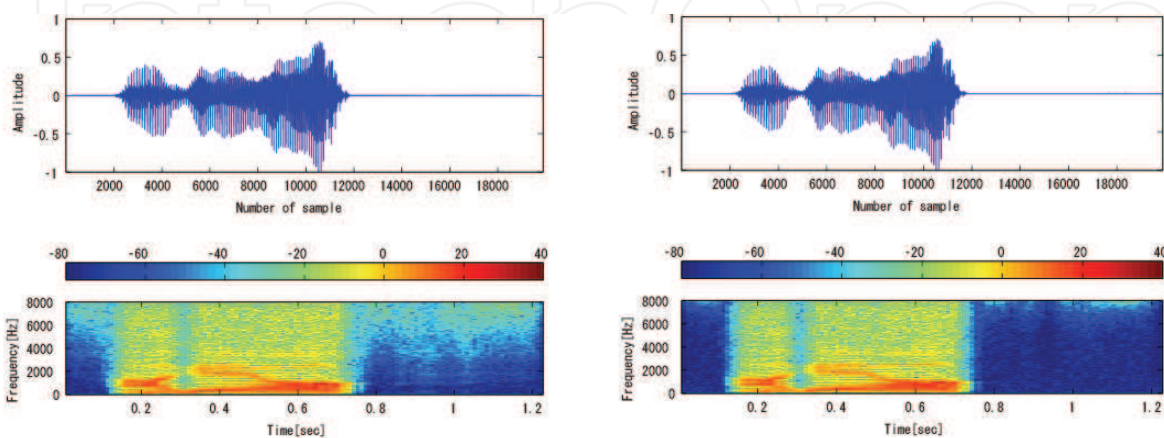

Fig. 20. Retrieval signal with Wiener 2   Fig. 21. Retrieval signal with Wiener 3

the recording signal, so the number of the suitable repetitions should be selected appropriately for recognition with the proposed retrieval method and noise environment. The recognition performance decreases if the repetitions are not sufficient when the stationary noise level is changed.

To solve the problem, we have to invent decision method for number of repetition in Wiener filtering. Then, we focused on the recognition system that calculates the likelihood between signal and model in recognition parameter. The algorithm chooses the clearest signal because it is appeared a highest likelihood from the recognition decoder. So we apply this idea to the decoding algorithm for differential acceleration.

8.2 Fundamental speech recognition

The fundamental speech recognition is described as follows briefly (Rabiner, 1993). Speech were recorded with a microphone and sampled at 16 kHz and 16 bits. As part of the recognition parameters, speech data were converted to the feature vectors as cepstrum coefficients. Speech recognition systems often use cepstrum techniques that consist of LPC and melfrequency cepstral coefficients (MFCCs) (Davis and Mermelstein, 1980). MFCCs were used in 12 dimensions computed every 10 ms. The differences in coefficients (ΔMFCC) and power (ΔLogPow) were also incorporated into the feature vectors. Thus, the feature vectors in each frame consisted of 25 (= 12 + 12 + 1) variables. The recognition system then uses these parameters to discriminate sub-word or word unit candidates. Each sub-word or word unit model consists of hidden Markov models (HMM) (Rabiner, 1993), which have transition probability and output probability distributions. A neural network (NN) is also used in speech recognition previously. However ANN is applied to other application (Rivara et al., 2009), speech recognition systems uses only HMM recently. Usually, the HMM's state transitions are in a left-to-right model, and the initial and final states are limited. This system calculates the likelihood (nearness of each acoustic model to an input parameter) using feature vectors and HMM. From the highest likelihood, the system can determine the candidate word/sub-word HMM. Generally, likelihood of acoustic models is computed with Viterbi algorithms. This recognition system includes dictionary and acoustic models, however no language models. The dictionary indicates the sub-word sequence for a word, because a word consists of the sub-word unit acoustic model. Equation (6) shows the fundamental formulation of this isolated word/sub-word recognition.

$$\hat{w} = \arg\max_{w \in W} P(X \mid w) \tag{6}$$

In equation (6), $\hat{w}$ is an estimated candidate, W is set of all candidates, w represents each candidate in W, and $X = x_1, x_2, \cdots, x_n$ are feature vectors. $P(X \mid w)$ is likelihood from the acoustic model. Generally, the acoustic model represents the sub-word or word unit. The system determines the estimated candidate by the likelihood from each HMM.

8.3 Proposed algorithm

To avoid decreasing of the recognition performance, the decoder algorithm was improved for signal retrieval using differential acceleration. The proposed algorithm using differential acceleration is shown in Figure 22 and represents three input signals. First, the speaker uttered a word, and its signal was then extracted with the accelerator. Thus, the signals $X_1, X_2, \cdots, X_N$ were performed for speech recognitions as input signals. The signals of

Figure 22 represent original body-conducted speech, differential acceleration and retrieval signals using Wiener filtering. This experiment focused on the body-conducted speech recognition in a noisy environment, so the speech signal is removed in this experiment. Although the recognition performance is improve that its decoder combines speech as input signal when the noise level is low. Equation (7) was applied for speech recognition with signals of differential acceleration.

$$w_i = \arg\max_{w \in W} P(X_i \mid w) \tag{7}$$

The recognition decoders give the recognition results that are the candidates of words $w_1, w_2, \cdots, w_N$. The parameters of the frame lengths and acoustic scores differ in scale for each signal because Wiener filtering causes a problem in sample length. To compare the acoustic scores, it was necessary to regulate the number of samples. The regulated acoustic score was calculated by the acoustic score and flame length. With equation (8), each score was compared and then the decoder determined the final candidate using the regulated acoustic scores.

$$w_{final} = \arg\max_{\substack{w \in W \\ i=1,\cdots,N}} \left\{ \frac{P(X_i \mid w)}{l_i} \right\} \tag{8}$$

8.4 Experimental setup

Finally, we evaluated the signal retrieval from body-conducted speech in a noisy environment with conventional decoding and the decoding algorithm for differential acceleration. The following signals and decoding algorithms were performed in this experiment:

- BCS: body-conducted speech
- Diff. Acc.: differential acceleration of BCS
- Wiener 1: differential acceleration using Wiener filtering that is repeated one time
- Wiener 2: differential acceleration using Wiener filtering that is repeated two times
- Wiener 3: differential acceleration using Wiener filtering that is repeated three times
- Max: the decoding algorithm for differential acceleration using regulated acoustic scores from the retrieval signals and body-conducted speech.

To evaluate the performance of signal retrieval using differential acceleration, the Julian (Kawahara et al., 1999; Lee et al., 2001), the Japanese speech recognition system that consists of grammar and an acoustic model, was used as the recognition decoder. The experimental environments are shown Table 2.

The improvement efficacy was evaluated by the following measurements:

$$Correct = \frac{Correct\ words}{Rercognition\ words} \times 100\,[\%] \tag{9}$$

$$Relative\ Improvement = \frac{Correct\ rate\ B - Correct\ rate\ A}{100 - Correct\ rate\ A} \times 100\,[\%] \tag{10}$$

Equation (9) is a word correct rate, which calculates a result whether the word is recognized or not. Equation (10) is relative improvement, depicting major improvements above base line levels.

8.5 Recognition results

Retrieval signals with Wiener filtering exhibited marginal improvement with a decrease in recognition because the algorithm rejects a stationary noise and consonant frequency. This problem is described in Section 8.1. We speculated that the noise reduction made improved the recognition performance when the stationary noise level was louder. Clear body-conducted speech retrieval using differential acceleration is not always associated with noisy environment, because the environment often varies according to factors such as the speaker, equipment. Tables 7–9 are shown the recognition results of each speaker. And Table 10 is the result of average of all speakers. The meaning of each data is described in Section 5. In all speakers, the recognition performances are improved on 'Diff. Acc.' and 'Max' however Wiener's results are little improvement. 'Diff. Acc.' becomes a clear signal in this experiment because it does not produce a stationary noise from differential acceleration. And 'Max' is worked on the recognition experiment correctly, so it is proven by Table 10 that the effectiveness of the combination method using retrieval method of signal and its decoding algorithm. So, the recognition performances are improved about 3–4% in word correct rate and about 10–19% in relative improvement.
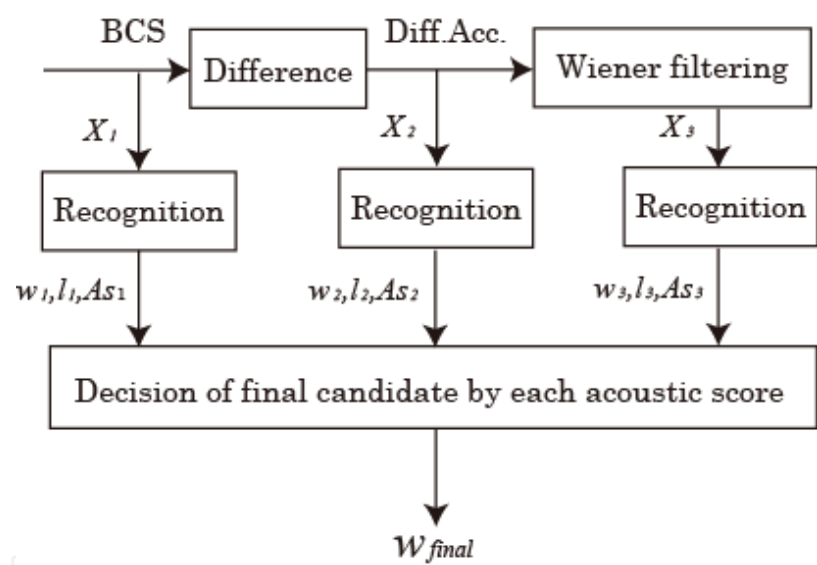

Fig. 22. Decoding algorithm for differential acceleration

	Set 1	Set 2	Set 3	Average	RI
BCS	72%	67%	75%	71.3%	-
Diff. Acc.	76%	74%	80%	76.7%	18.6%
Wiener 1	65%	69%	80%	71.3%	0.0%
Wiener 2	70%	71%	79%	73.3%	7.0%
Wiener 3	60%	74%	81%	71.7%	1.2%
Max	76%	71%	80%	75.7%	15.1%

Table 7. Recognition results of Speaker A in noisy environment

	Set 1	Set 2	Set 3	Average	RI
BCS	70%	71%	71%	70.7%	-
Diff. Acc.	71%	72%	77%	73.3%	9.1%
Wiener 1	67%	65%	72%	68.0%	-9.1%
Wiener 2	68%	65%	70%	67.7%	-10.2%
Wiener 3	72%	67%	72%	70.3%	-1.1%
Max	72%	72%	77%	73.7%	10.2%

Table 8. Recognition results of Speaker B in noisy environment

	Set 1	Set 2	Set 3	Average	RI
BCS	84%	76%	75%	78.3%	-
Diff. Acc.	86%	79%	79%	81.3%	13.8%
Wiener 1	87%	74%	76%	79.0%	3.1%
Wiener 2	88%	69%	77%	78.0%	-1.5%
Wiener 3	86%	74%	77%	79.0%	3.1%
Max	86%	74%	77%	79.0%	3.1%

Table 9. Recognition results of Speaker C in noisy environment

	Average	RI
BCS	73.4%	-
Diff. Acc.	77.1%	13.8%
Wiener 1	72.8%	-2.5%
Wiener 2	73.0%	-1.7%
Wiener 3	73.7%	0.8%
Max	76.1%	10.0%

Table 10. Recognition results of all speakers in noisy environment

9. Conclusions and future work

Here, we investigated a signal retrieval from body-conducted speech using differential acceleration combined with conventional noise reduction methods. Specifically, differential acceleration was used to emphasize the high frequency component of body-conducted speech. Additionally, this method is little cost in calculation. Although the differential acceleration of body-conducted speech became a retrieval signal when stationary noise was present. Thus, it was possible to remove noise effectively using conventional noise reduction methods i.e. spectral subtraction or Wiener filtering. From the experimental results, the Wiener filtering method proved to be a suitable noise reduction method for differential acceleration, as evidenced by the difference between normal speech and each retrieval signal. So, this method can be used to estimate a clear speech using only body-conducted speech. Thereafter, we combined the proposed method and conventional signal retrieval

methods to make a clear signal. Combined methods using an adaptive filter for a word, its effectiveness is shown the results in the difference of spectrograms between each signal and speech. Thus, it appears that the proposed method was effective when pre-processing for the conventional signal retrieval techniques existed. The recognition experiment using the differential acceleration followed by the Wiener filtering method, demonstrated the efficacy of differential acceleration that the recognition performance improved 3 - 5 % in isolated word recognition. These results suggest that the retrieval signal approximated natural speech. So we concluded that the proposed method was able to estimate clear speech from body-conducted speech in quiet room.

As a next step of the proposed method, we applied the noise reduction methods to body-conducted speech in a noisy environment. From the experimental results, the Wiener filtering method proved to be a suitable noise reduction method for differential acceleration as evidenced by comparing spectrograms with normal speech and each retrieval signal in a noisy environment. To decide the suitable number of repetitions on Wiener filtering, we proposed a decoding algorithm that was used to regulate the acoustic scores. From the recognition experiments, the differential acceleration signals exhibited the most improvement, and the proposed algorithms improved when compared with base-line body-conducted speech signals. Furthermore, the system does not depend on the environments and speakers because the system chose the highest likelihood from the signals.

As a future works, we will be examined as a pre-processing for the speech support system using body-conducted speech for disorders that converts a clear speech from body-conducted speech of disorders (Ishimitsu and Nakayama, 2009). And, the signal retrieval using differential acceleration is applied to body-conducted speech in noisy environments. Furthermore, we will also construct the microphone using body-conducted speech and differential acceleration for noise environment.

10. References

S. Nakagawa (2007). To spoken document processing from spontaneous speech transcription, in proc. 2007 Autumn Meeting ASJ CD-ROM, pp.1-4

N. Kitaoka, T. Yamada, S. Tsuge, C. Miyajima, T. Nishiura, M. Nakayama, Y. Denda, M. Fujimoto, K. Yamamoto, T. Takiguchi S. Kuroiwa, K. Takeda, and S. Nakamura (2006). CENSREC-1-C: Development of evaluation framework for voice activity detection under noisy environment, in IPSJ SIG Technical Report, 2006-SLP-63, pp.1-6

H. Hirsch and D. Pearce (2000). The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions, in proc. ISCA ITRW ASR2000, pp. 181-188

S. Ishimitsu, M. Nakayama, and Y. Murakami (2004). Study of Body-Conducted Speech Recognition for Support of Maritime Engine Operation, in Journal of the JIME, Vol.39 No.4, pp.35-40

T. Tamiya and T. Shimamura (2006). Improvement of Body-Conducted Speech Quality by Adaptive Filters, in IEICE Technical Report, SP2006-191, pp.41-46

T. T. Vu, M. Unoki, and M. Akagi (2006). A STUDY ON RESTORATION OF BONE-CONDUCTED SPEECH WITH LPC-BASED MODEL, in IEICE Technical Report, SP2005-174, pp.67-78

Z. Liu, Z. Zhang, A. Acero, J. Droppo, and X. Huang (2004). Direct Filtering for Air- and Bone-Conductive Microphones, in proc. IEEE International Workshop on Multimedia Signal Processing (MMSP'04), pp.363-366

S. Dupont, C. Ris, and D. Bachelart (2004). Combined use of closetalk and throat microphones for improved speech recognition under non-stationary background noise, in proc. COST278 and ISCA Tutorial and Research Workshop (ITRW) on Robustness Issues in Conversational Interaction, paper31

S. Itahashi (1991). A noise database and Japanese common speech data corpus, in Journal of ASJ, Vol.47 No.12, pp.951-953

Y. Gong (1995). Speech recognition in noisy environments: A survey, Speech Communication 16, pp.261-291

Y. Nomura, H. Tozawa, J. Lu, H. Sekiya, and T. Yahagi (2006). Musical Noise Reduction by Spectral Subtraction Using Morphological Process, in Trans. of IEICE on information and systems, Vol.89 No.5, pp.991-1000

K. Yamashita, S. Ogata, and T. Shimamura (2005). Improved Spectral Subtraction Utilizing Iterative Processing, in Trans. of IEICE on Inst. of Electronics, Information and Communication Engineers, Vol.J88-A No.11, pp.1246-1257

D. Li and D. O'Shaughnessy (2003). Speech Processing: A Dynamic and Optimization - Oriented Approach, Marcel Dekker Inc

J. Durbin (1960). The Fitting of Time-Series Models, Review of the International Statistical Institute, Vol.28 No.3, pp.233-244

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

S. Ishimitsu, M. Nakayama, and K. Oda (2007). Study of Speech Support System for Disorders Using Body-Conducted Speech Recognition, in proc. 2007 Autumn Meeting ASJ CD-ROM, pp.715-716

S. Haykin (1996). Adaptive filter theory (3rd ed.), Prentice-Hall, Inc.

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

A. Lee, T. Kawahara, and K. Shikano (2001). Julius - an open source real-time large vocabulary recognition engine, in Proc. European Conference on Speech Communication and Technology (EUROSPEECH), pp. 1691-1694

K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi (1999). JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research, in Journal of ASJ, Vol.20, No.3, pp.199-206

S. Young, J. Jansen, J. Odell, and P. Woodland (2000). The HTK Book for V2.0, Cambridge University

L. Rabiner (1993). Fundamentals of Speech Recognition, Prentice Hall

S.B. Davis and P. Mermelstein (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, IEEE Trans. Acoustics, Speech Signal Proc., ASSP-28, Vol. 4, pp.357–366

N. Rivara, P.B. Dickinson, and A.T. Shenton (2009). Constrained variance control of peak-pressure position by spark-ionisation feedback for multi-cylinder control', *International Journal of Advanced Mechatronic Systems (IJAMechS)*, Vol. 1, No. 4, pp.242–250

S. Ishimitsu and M. Nakayama (2009). Construction of Speech Support System Using Body-Conducted Speech Recognition for Disorders, in proc. of The Third International Conference on Innovative Computing, Information and Control (ICICIC2008), CD-ROM
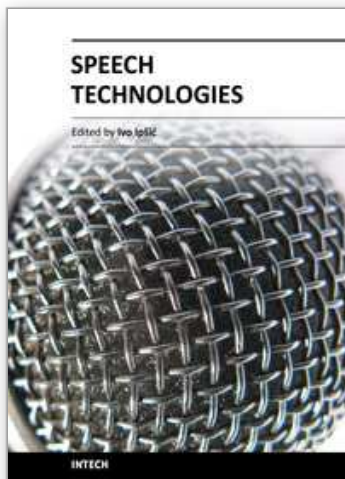
Speech Technologies

Edited by Prof. Ivo Ipsic

This book addresses different aspects of the research field and a wide range of topics in speech signal processing, speech recognition and language processing. The chapters are divided in three different sections: Speech Signal Modeling, Speech Recognition and Applications. The chapters in the first section cover some essential topics in speech signal processing used for building speech recognition as well as for speech synthesis systems: speech feature enhancement, speech feature vector dimensionality reduction, segmentation of speech frames into phonetic segments. The chapters of the second part cover speech recognition methods and techniques used to read speech from various speech databases and broadcast news recognition for English and non-English languages. The third section of the book presents various speech technology applications used for body conducted speech recognition, hearing impairment, multimodal interfaces and facial expression recognition.

INTECH
open science | open minds