ICIC Express Letters
Part B: Applications
Volume **10**, Number **5**, May **2019**

ICIC International ©2019 ISSN 2185-2766
pp. **427–431**

BODY-CONDUCTED SPEECH RECOGNITION USING MODEL ADAPTATION

Masashi Nakayama[1], Satoshi Nakatani[2] and Shunsuke Ishimitsu[1]

[1]Graduate School of Information Sciences
Hiroshima City University
3-4-1 Ozuka-higashi, Asaminami, Hiroshima 731-3194, Japan
{ masashi; ishimitu }@hiroshima-cu.ac.jp

[2]Department of Electrical and Computer Engineering
National Institute of Technology, Kagawa College
355 Chokushi, Takamatsu, Kagawa 761-8058, Japan

Received October 2018; accepted January 2019

Abstract. *One of the problems with speech recognition is a low recognition performance in noise environments, as speech is easily influenced by noise. On the other hand, body-conducted speech (BCS) can be measured under any environment, as it is not affected by airborne noise. With this feature, a noise robust speech recognition system can be realized with BCS. However, the BCS recognition system should re-estimate an acoustic model for high performance, as the parameter characteristics have large differences between speech and BCS. In this study, we aim to improve the acoustic model for BCS recognition with a focus on parameters such as mean vector, covariance matrix, weight, and transition probability in acoustic model, along with model adaptation for BCS, by using the maximum likelihood (ML) and maximum a posteriori method (MAP), respectively. As the result, BCS recognition with model adaptation was achieved to about 95% and more in word recognition rate.*

Keywords: Body-conducted speech, Speech recognition, Model adaptation

1. **Introduction.** Speech conversation is one of the most important communication method for us. However, noisy environments make it difficult to communicate using speech easily, as the noise becomes a disturbing factor for understanding spoken words and sentences. To avoid the effects of noise sound, several researchers have investigated noise suppression and signal extraction in a noise environment [1-4]. Microphone arrays can measure from about −5 dB to 0 dB signal-to-noise ratio (SNR) [4]. Body-conducted speech (BCS) is one of the solutions for measuring speech sound in a noise environment.

BCS can be used to conduct the noise robust speech recognition, as the BCS sound is not affected by the noise sound. However, it is difficult to measure clear sound using BCS, compared to speech of 2 kHz and more. In order to achieve high performance in recognition, the acoustic model has to re-estimate BCS [5]. Furthermore, for the improvement of recognition performance in an acoustic model, we need to discuss parameters such as covariance matrix, transition probability, and weight, which can be expected to contribute to recognition performance with model adaptation.

The remainder of the paper is organized as follows. In Section 2, we show the differences between speech and BCS, and then discuss advantages and weaknesses of the sounds. Next, Section 3 provides an overview of an acoustic model, and Section 4 discusses experiments on speech recognition with model adaptation of MAP and ML methods. Finally, Section 5 offers conclusions and recommendations for future research in this field.

DOI: 10.24507/icicelb.10.05.427

2. **Speech and BCS.** Speech is an air-conducted sound, and is easily influenced by surrounding noise. In contrast, BCS is a solid-propagated sound that is difficult to be influenced by noise. Figures 1 and 2 represent the utterance of the name of a local Japanese place called "Asahi" by a 20-year-old male. The utterance was chosen from the JEIDA-100-local-place-name database [6]. Table 1 shows the recording environments. The signals were recorded at 16 kHz with 16 bits. Speech was measured using a microphone positioned at a distance of 30 cm from the mouth, which is the ideal microphone position for practical use, and BCS was measured using an accelerometer placed on the upper lip. The distance for speech is assumed as that of a conventional speech interface, such

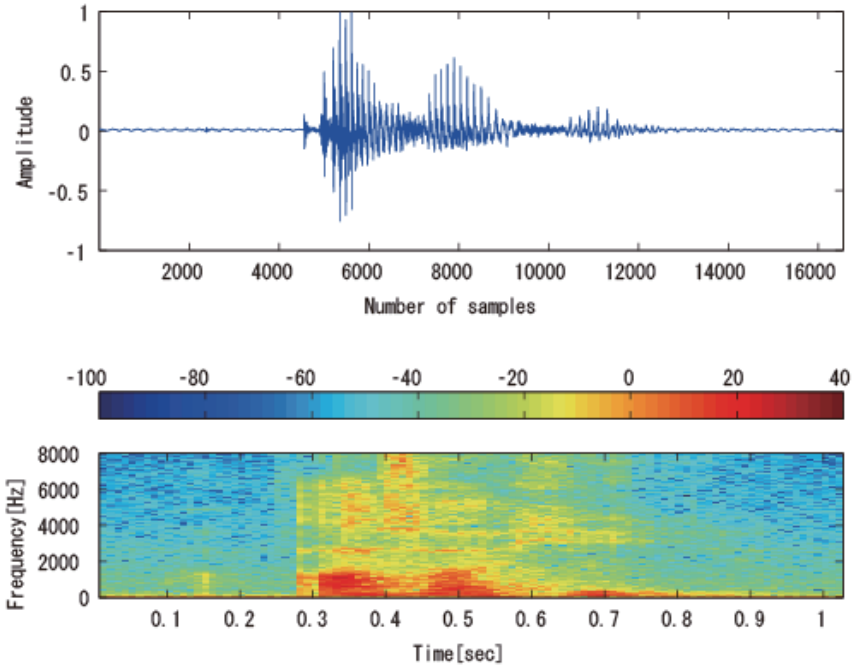

FIGURE 1. Speech

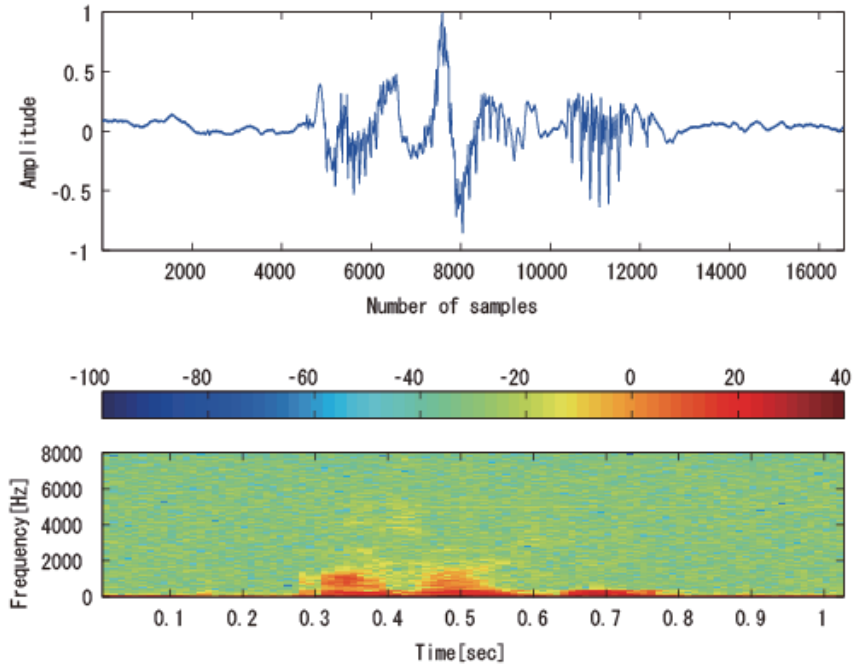

FIGURE 2. BCS

Table 1. Recording environments

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

as a car navigation system. The measuring position for BCS has already been discussed and proven suitable, compared with the feature parameters between speech and BCS, in previous research [5]. However, BCS does not measure 2 kHz or more of higher frequency components, and conventional speech recognition does not work for practical use due to differences in the quality of sound and feature parameters.

3. **Acoustic Model.** The fundamentals of speech recognition can be described and implemented using information theory for estimating a probabilistic model. The decoding of speech recognition means calculation of the likelihood of acoustic and linguistic matching with speech and models. In this study, we decode to the word utterance with only acoustical matching in order to focus on how to achieve high likelihood of acoustic and linguistic matching with acoustic model and feature parameters of sounds. Conventional speech recognition is optimized using an unspecified speaker's speech; hence, it is difficult to achieve high likelihood when we used BCS directly for the system. To achieve high recognition performance, we should improve sound quality or the acoustic model. Therefore, in this study, we focused on the re-estimation of an acoustic model that is composed from a statistical model, including multi-dimensional Gaussians and transition probability of HMM (Hidden Markov Model).

4. **Experiment.** Experiments were performed to improve the performance of speech recognition with re-estimations of the acoustic models. The parameters in the model should be re-estimated for speech into for BCS, because speech recognition estimates result candidates of words chosen by matched with feature parameters of sound and cepstrum parameters in the models. The model parameters include feature vectors, covariance matrix, weight, and transition probability; thus, the authors experimented and discussed whether the recognition performances should be evaluated using model re-estimation or not.

4.1. **Experimental setup.** Table 2 shows the experimental conditions for the isolated word recognition. The experiment used two databases: 20021213 and 20030228. In both databases, the speaker uttered JEIDA 100 hundred local place name in a quiet room. The signals were recorded using a microphone and an accelerometer. Database 20021213 comprises 900 words uttered by three male speakers during three trials, and database 20030228 comprises 600 words uttered by two male speakers during three trials.

A speech-recognition decoder, Julius 4.2 [7], employed for isolated word recognition as well, was used in this experiment. The experiments were performed under two conditions: open and close test. The re-estimations of acoustic models were only used for database 20021213, and were then re-estimated using HTK [8]. However, our recognition experiments used both databases. Database 20021213 was used for the closed test, and database 20030228 was used for the open test. The dictionary for recognition is a 100-local-place-name dictionary from JEIDA, which includes 100 names of local places in Japan, in which

Table 2. Experimental conditions

Speaker	20021213: 3 males; 20030228: 2 males
Data set	100 words × 3 set/person
Vocabulary	JEIDA 100 names of local places
Decoder	Julius 4.2
Acoustic model	Gender-dependent tri-phone
Model condition	16 mix, clustered 3,000 states
Parameter	MFCC(12) + ΔMFCC(12) + ΔPOW(1)
Training for baseline model	20,000 samples of speech with HTK 2.0
Model re-estimation condition	600 samples of speech or BCS, 20021213 with HTK 3.4.1

Table 3. Recognition results of model re-estimations

		20021213				20030228			
		Speech		BCS		Speech		BCS	
		Correct	Diff.	Correct	Diff.	Correct	Diff.	Correct	Diff.
Baseline		94.83	+0.00	54.33	+0.00	96.11	+0.00	46.56	+0.00
ML	Mean	99.82	+4.99	99.17	+44.83	99.81	+3.70	99.15	+52.59
	Variance	100.00	+5.17	99.72	+45.39	99.89	+3.78	99.52	+52.96
	Transition	94.67	−0.17	55.61	+1.28	96.56	+0.44	46.81	+0.26
	Weight	96.39	+1.56	73.39	+19.06	97.67	+1.56	67.30	+20.74
	All	100.00	+5.17	100.00	+45.67	99.96	+3.85	100.00	+53.44
MAP	Mean	99.11	+4.28	94.22	+39.89	99.59	+3.48	94.48	+47.93
	Variance	99.00	+4.17	91.28	+36.94	99.74	+3.63	90.41	+43.85
	Transition	94.83	+0.00	54.33	+0.00	96.11	+0.00	46.56	+0.00
	Weight	95.83	+1.00	60.00	+5.67	97.37	+1.26	54.52	+7.96
	All	94.83	+0.00	54.33	+0.00	96.11	+0.00	46.56	+0.00

phonemes in the database were balanced at mora and syllable of appearance ratio. In addition, the acoustic model, which uses a tri-phone model as the phoneme and/or syllable, was expressed as an HMM, which was composed from parameters such as mean vectors, diagonal covariance matrices, mixture weight, and the transition probabilities of a particular state because there are main parameters at HMM. The re-estimations of parameters in HMM are calculated using two algorithms: the maximum likelihood estimation method (ML) and the maximum a posteriori probability estimation method (MAP) [9]. ML is a stochastic approach to calculating the acoustic model; however, there is a possibility of falling into a local solution in case of non-suitable initial parameters. MAP is a method based on the likelihood maximization criterion, which has the advantage that even if the initial value is unstable, it is difficult to fall into a local solution.

4.2. **Experimental result and discussion.** Table 3 shows the recognition results of model re-estimations. The baselines use gender-dependent models for unspecified speakers without re-estimation. The other data used are the results of acoustic models with re-estimations. From the results, we confirmed the effectiveness of model re-estimations at mean vectors, mixture weights, and diagonal covariance matrices. On the other hand, there is no effect to use transition probability. At first, we focused on the result of mean vectors, which were confirmed on about 40% to 50% and more because it is one of most important factors. Next, covariance is also achieved to about 40% to 50%, as the covariance can cover distribution of each parameter when the mean vector is the same. It can be seen that even with the same mean vector, distributions between speech and BCS are covered with the re-estimation of covariance only. Transition probability refers

to the staying probability of each state of HMM. However, the time duration and its boundary at each state of HMM are always the same, as both sounds are synchronized. The re-estimated boundary of each syllable and phonemes are almost the same; thus, the efficiency of re-estimation of the transition probability was not obtained.

5. **Conclusion and Future Work.** This study investigated and experimented with improvements to BCS recognition using conventional speech recognition, and evaluated recognition performance using model re-estimations. It was confirmed that the recognition performances significantly improved after the re-estimation of the mean vector, mixture weights, and covariance matrices, using two re-estimation algorithms: ML and MAP. The level of performance improved sufficiently to allow the practical application of speech recognition.

In future, the authors plan to conduct these performance improvements using model re-estimations with sound quality improvement method, combined with a differential acceleration and noise reduction method [4].

REFERENCES

[1] Y. Gong, Speech recognition in noisy environments: A survey, *Speech Communication*, vol.16, pp.261-291, 1995.

[2] H. Hermansky, Perceptual linear prediction (PLP) analysis of speech, *Journal of the Acoustical Society of America*, vol.87, no.4, pp.1738-1752, 1990.

[3] L. Wang, N. Kitaoka and S. Nakagawa, Robust distant speech recognition by combining multiple microphone-array processing with position-dependent CMN, *EURASIP Journal on Applied Signal Processing*, vol.2006, pp.1-11, 2006.

[4] M. Nakayama, S. Ishimitsu and S. Nakagawa, A study of making clear body-conducted speech using differential acceleration, *IEEJ Trans. Electrical and Electronic Engineering*, vol.6, no.2, pp.144-150, 2011.

[5] S. Ishimitsu, M. Nakayama, T. Yoshimi and H. Yanagawa, Noise-robust recognition system making use of body-conducted speech microphone, *AES 122nd Convention*, Vienna, Austria, 2007.

[6] S. Itahashi, A noise database and Japanese common speech data corpus, *Journal of ASJ*, vol.47, no.12, pp.951-953, 1991.

[7] *Julius*, http://julius.sourceforge.jp/.

[8] *HTK*, http://htk.eng.cam.ac.uk/.

[9] C. H. Lee and J. L. Gauvain, Speaker adaptation based on map estimation of HMM parameters, *IEEE International Conference on Acoustics, Speech, and Signal Process (ICASSP-93)*, pp.558-561, 1993.