inter.noise 2014
MELBOURNE AUSTRALIA
16-19 NOVEMBER

Improvement of body-conducted speech recognition using model estimation

Masashi NAKAYAMA[1]; Shunsuke ISHIMITSU[1]; Satoshi NAKATANI[2]

[1] Graduate School of Information and Sciences, Hiroshima City University, Japan

[2] Department of Electrical and Computer Engineering, Kagawa National College of Technology, Japan

ABSTRACT

One problem with speech recognition is a low performance in noisy environments because it is easily influenced by aerial in air. Although the sound quality of body-conducted speech (BCS) and regular speech are different, with BCS recognition, it is possible to recognize an utterance in noisy environments with a rating of 98 dB sound pressure level (SPL) in our previous study. In this study, we investigate how to improve BCS recognition performance using model re-estimation methods of ML and MAP. An acoustic model uses parameters such as mean vector, covariance matrix, weight, and transition probability. Recognition performance is improved by model re-estimation of speech and BCS using maximum likelihood and maximum a posteriori methods, respectively. We confirmed that improvements in recognition performance are achieved for practical through the re-estimation of the covariance matrix and mean vector.

Keywords: body-conducted speech, speech recognition, model estimation
I-INCE Classification of Subjects Number(s): 69.2.

1. Introduction

Conversation is one of the most important communication methods for human beings; however, noises in the air act as a disturbing factor in this type of communication. Approaches to robust communication methods and instruments are widely proposed and investigated in the research fields of speech signal processing and human interface. In speech signal processing, the robust communication is one of the most significant research topics because speech recognition does not achieve effective performance for practical use.

The approaches for measuring speech in a noisy environment are classified into physical and mathematical approaches. Physical approaches are signal-measuring methods employ physical methods that use special microphones such as microphone arrays and bone-conducted speech microphones, while mathematical approaches include noise reduction methods and sound quality improvement methods [1–4]. Microphone arrays work correctly when there is an approximately 0 to -5 dB signal-to-noise ratio (SNR) in the environment; however, microphone arrays do not work in noise-heavy environments [4]. Conventional mathematical approaches do not achieve a sufficient level of performance in noisy environments because several information and speech samples are required for estimating and clearing speech and/or suppressing noise. Using this research as background information, the authors proposed and discussed body-conducted speech (BCS), which is conducted on skin and bone in a human being. In general, speech is easily influenced by noise in the air, however BCS can measure by accelerometer in a 98-dB SPL-noise (-20 dB SNR) environment. Compared to conventional speech recognition tasks in the noise environments of AURORA and CENSREC [5,6], this environment has one of the heaviest task conditions. An advantage of BCS is its robustness for noise because BCS does not affect by aerial noise ; a disadvantage is its low quality sound at 2 kHz or higher when compared to the regular speech. This disadvantage decreases the performance of speech recognition when using BCS as an input signal, because feature parameters such as cestrum coefficients are different between regular speech and

[1] masashi@hiroshima-cu.ac.jp, ishimitu@hiroshima-cu.ac.jp

BCS. To achieve a sufficiently acceptable level of performance for practical use, users have to improve the sound quality and/or re-estimate the parameters of the acoustic model in speech recognition. Previously, authors investigated both approaches and studied improvements in sound quality and re-estimation of the mean vector in the acoustic model [7]. This paper shows performance improvement using re-estimations of parameters in an acoustic model. The parameters include mean vectors, covariance matrix, transition probability, and weight.

2. Speech and BCS

Speech is an air-conducted sound and is easily influenced by surrounding noise. By contrast, since BCS is a solid, propagated sound, and it is difficult for noise to influence it. Figs. 1 and 2 represent the utterance of a local Japanese place called “Asashi” by a twenty-year-old male. The utterance was chosen from the JEIDA-100-local-place-name database [8]. Table 1 shows the recording environments. Signals were recorded at 16 kHz with 16 bits. Speech was measured by the microphone, which was positioned at a distance of 30 cm from the mouth to lip, is microphone position for practical use and BCS was measured using an accelerometer placed at the upper lip. The distance for speech is assumed to be that of a conventional speech interface such as a car navigation system. The measuring position for BCS has already discussed and proved as suitable location compared with feature parameters between speech and BCS at previous research [7]. However, BCS does not measure 2 kHz or more of higher frequency components, conventional speech recognition does not work for practical use because there are difference quality of sound and feature parameters.

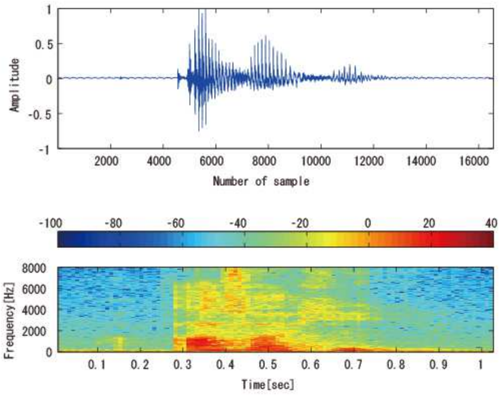

Figure 1. Speech.

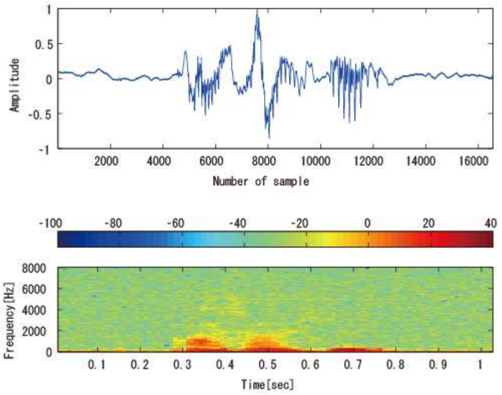

Figure 2. BCS.

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)
Accelerator	Ono Sokki NP-2110
Accelerator amplifier	Ono Sokki PS-602
Accelerator position	Upper lip

3. Experiment

To improve the performance of speech recognition, re-estimations of the acoustic models are experimented with and discussed. 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 cestrum parameters in the models. Model parameters include feature vectors, covariance matrix, weight, and transition probability, thus the authors experimented and discussed that the recognition performances should be evaluated with model re-estimation or not here.

3.1 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 hundreds of local place names in a quiet room. The signals were recorded using a microphone and an accelerometer. Database 20021213 is composed of 900 words that three male speakers uttered during three trials, and database 20030228 is composed of 600 words uttered by two male speakers during three trials.

A speech-recognition decoder, Julius 4.2 [9], is used for large-vocabulary, continuous-speech recognition, including isolated word recognition, and 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 by HTK [10]. 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 is gathered 100 local place name of Japan when conditioned balanced in mora and syllable of appearance ratio. In addition, the acoustic model, uses a tri-phone model as the phoneme and/or syllable, was represented as a hidden Markov model (HMM), which uses the following parameters: mean vectors, diagonal covariance matrices, mixture weight, and the transition probabilities of a particular state. The re-estimations of parameters in HMM are calculated by two algorithms: the maximum likelihood estimation method (ML) and the maximum a posteriori probability estimation method (MAP) [11].

Table 2. Experimental conditions.

Speaker	20021213: 3 males, 20030228: 3 males
Data set	100 words × 3 set/person
Vocabulary	JEIDA 100 local place names
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

3.2 Experimental results

Table 3 shows the recognition results of model re-estimations. Baselines are set using gender-dependent speech models for unspecified speakers without re-estimation. The other data used are results of acoustic models with re-estimations. From the results, the effectiveness of model re-estimations for both sounds at mean vectors, mixture weights, and diagonal covariance matrices were confirmed. On the other hand, this did not improve the conditions of transition probability. 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 time and length because both sounds are synchronized. Re-estimated prompters are almost the same; thus, the efficiency of re-estimation of the transition probability was not obtained.

4. Conclusions and future works

This paper investigated and experimented with improvements to BCS recognition using conventional speech recognition, evaluating 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 was sufficiently improved to allow the practical application of speech recognition.

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

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

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,” J. Acoust. Soc. Amer., 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, Volume 2006, Article ID 95491, pp. 1–11, 2006.
4. M. Nakayama, S. Ishimitsu, and S. Nakagawa, “A study of making clear body-conducted speech using differential acceleration,” IEEJ Transactions on Electrical and Electronic Engineering, vol. 6, issue 2, pp. 144–150, March 2011.
5. H. Hirsch and D. Pearce, “The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions,” in ASR-2000, pp. 181–188, 2000.
6. 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, “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, 2006.
7. S. Ishimitsu, M. Nakayama, T. Yoshimi, and H. Yanagawa, “Noise-robust recognition system making use of body-conducted speech microphone,” AES 122nd Convention, Paper ID: 7105, Austria Center Vienna, Vienna, Austria, 2007.
8. S. Itahashi, “A noise database and Japanese common speech data corpus,” in Journal of ASJ, vol. 47, no. 12, pp. 951–953, 1991.
9. Julius, http://julius.sourceforge.jp/. (in Japanese)
10. HTK, http://htk.eng.cam.ac.uk/.
11. C. H. Lee and J. L. Gauvain, “Speaker adaptation based on map estimation of HMM parameters,” IEEE Internat. Conf. on Acoustics, Speech, and Signal Process (ICASSP-93), pp. 558–561, 1993.