Acoustics'08 Paris
June 29-July 4, 2008
www.acoustics08-paris.org

Speech recognition with body-conducted speech using differential acceleration

Masashi Nakayama[a], Shunsuke Ishimitsu[a] and Seiji Nakagawa[b]

[a]Hiroshima City University / National Institute of Advanced Industrial Science and Technology, 3-4-1 Ozuka-Higashi, Asa-Minami-Ku, 731-3194 Hiroshima, Japan
[b]National Institute of Advanced Industrial Science and Technology (AIST), 1-8-31 Midorigaoka, 563-8577 Ikeda, Osaka, Japan
m.nakayama@aist.go.jp

Speech-recognition rates decrease in noisy environments. The body-conducted speech, conducted in solids such as body and skins, has a noise-robust characteristics and can be served for recognition systems even in 98 dBSPL (-20 dBSNR) noise environments. However, the body-conduction could not capture high frequency sounds. Conventional methods for the improvement in sound quality of body-conducted speeches needs both speeches and body-conducted speeches. In this paper, a new body-conducted speech retrieval technique in sound quality without a speech signal itself is proposed. First, high-frequency components in the body-conducted speech were emphasized using differential acceleration. Second, conventional noise reduction method was adopted to make a clear body-conducted speech from a retrieval speech which contains constant noise. The recognition experiments using the proposed method showed that it improved recognition rate in all speakers.

1 Introduction

During recent years, applications using speech recognition have been developed and the technology is being used in various settings, including note taking for meetings, dictation during lectures, and car navigation systems. Research into speech recognition is being conducted to improve recognition accuracy and improve spoken document processing [1]. However, even with developments in speech recognition technology, a good enough performance cannot be obtained in a noisy environment. Activity which offers a standard rate scale for evaluating speech recognition performance in a noisy environment, such as CENSREC [2] and AURORA [3] has been discussed. We built a body-conducted speech recognition system which can be recognized in a noise environment of 98 dB SPL in the engine room of Oshima-maru, a training ship in Oshima National College of Maritime Technology [4]. Body-conducted speech is a solid signal propagated through skin, bone and so on. Therefore, it is a more noise robust signal than an air conducted voice. This was confirmed in experiments using the system where an average recognition rate of 95% or more was obtained. However, in order to obtain a sufficient recognition rate, it is necessary to estimate a parameter to correlate body-conducted speech to the acoustic model built from voices of unspecified speakers. This is because body-conducted speech and voices have different frequency characteristics. In voice and body-conducted speech, since the feature vectors also differ, a good enough recognition rate cannot be obtained. Moreover, since a body-conducted speech system does not include the high frequency component of 2 kHz and more, it is not clear as a voice. As a result many researchers have studied ways of creating a clear sound from body-conducted speech. The conventional method used until now was also examined. Even if it found the integral signal directly from body-conducted speech it was still not possible to acquire the displacement signal. However, with the proposed method it is possible to estimate a clear voice and signals of other dimensions from body-conducted speech [5]. We confirmed the effectiveness of the retrieval signal by LAR distance and an isolated word recognition experiments that performed on input comparing a body-conducted speech and a difference-acceleration signal. An improvement in signal recognition rate was obtained using the proposed method.

2 Body-conducted speech

A voice is an air conducted signal and is influenced by surrounding noise. On the other hand, since body-conducted speech is a solid propagated signal, it is hard to be influenced by noise. Figs. 1 and 2 are the word 'Asashi' that was obtained from the database of JEIDA [6]. This database contains 100 local place words which are uttered by a 20-year-old male. A voice was measured 30cm from the mouth with a microphone, and body-conducted speech was extracted from the upper lip with acceleration pickup. The microphone was set 30 cm from the mouth because the microphone position used when carrying out voice input for a car navigation system was assumed. Body-conducted speech was extracted from the upper lip which was able to provide the best cepstrum coefficient characteristics as a feature vector for recognition from previous research [4]. From Figs. 1 and 2, body-conducted speech has a lower quality signal of articulation score than a voice since the high frequency component of 2 kHz or more is smaller. The signal was recorded at 16 kHz and 16 bits in this research. Table 1 shows the recording environments used in this research. As shown in Fig. 3, it is difficult to estimate dimension signals, such as the signal of velocity and displacement, by simply integrating with a body-conducted speech directory. However, if you use our proposed method, it can estimate other dimension signals from body-conducted speech. Moreover, if a sampling frequency and a speaker change, it is necessary to design a filter because the proposed technique is designed to perform signal retrieval automatically with differential acceleration and a noise reduction filtering method. In order to fully emphasize the high frequency component contained in body-conducted speech at little calculation cost, waveform signal difference was used.

Table 1: Recording environments

Recorder	TEAC RD-200T
Microphone	Ono Sokki MI-1431
Microphone amplifier	Ono Sokki SR-2200
Microphone position	30cm (Mouth to mic.)

3 Differential acceleration

In the following section, we describe the use of differential acceleration. First, the difference-acceleration signal $x_{differential}(i)$ is estimated from the acceleration signal $x(i)$ with the formula (1), where $x(i)$ is waveform data of time frame i.

$$x_{differential}(i) = x(i+1) - x(i) \qquad (1)$$

Since $x(i)$ is a difference signal between each time frame, and it becomes a smaller signal, it requires gain adjustment. Fig. 4 shows the difference-acceleration signal estimated from Fig. 2 with formula (1). From Fig. 4, it can be considered that the difference-acceleration signal is a voice mixed stable noise. So we have invented a differential acceleration and conventional noise reduction method.

4 Noise Reduction

In this chapter, in order to obtain an effective frequency component from a difference-acceleration signal, the following noise reduction techniques were performed and compared to differential acceleration.

- Spectral subtraction method
- Wiener filtering method

4.1 Spectral Subtraction Method

The Spectral Subtraction Method works by subtracting the spectrum of the noise section from the overall spectrum [7]. An algorithm is an easy and adaptive filter [4] [8]. MTF, and LPC [9], etc. need information such as a voice. However, since it cannot measure a voice in a noisy environment, the conventional method is not realistic. In order to solve this problem, we propose the technique of estimating a clear signal only by body-conducted speech, without using a target signal voice. If it is possible to estimate a clear signal only by body-conducted speech, it is not necessary to estimate parameters in an acoustic model and a speech recognition system can be used as it is. Moreover, it is also possible to use a microphone which can perform voice extraction under noise. Using this method, a clear signal was estimated by body-conducted speech alone using acceleration difference. It also confirmed that it was possible to acquire a clearer signal by using the proposed method and the conventional technique together. In addition, the technique of estimating the same displacement signal as a voice and the signal of other dimensions from body-conducted speech is widely used in the field of research into speech recognition as a noise reduction technique [5] [10]. A spectral subtraction method is shown in the following formulas, (2) and (3).

$$\begin{aligned} x(i) &= s(i) + n(i) && (2) \\ S(\omega) &= (|X(\omega)| - |N(\omega)|) \exp^{j \arg X(\omega)} && (3) \end{aligned}$$

It assumes that the difference-acceleration signal $x(i)$ is constituted from the voice signal $s(i)$ and the noise signal $n(i)$. An estimated spectrum $S(\omega)$ can be obtained using the spectral subtraction method. $\arg X(\omega)$ means the phase information on input signals spectrum $X(\omega)$. Fig. 5 shows the results from the spectral subtraction method when repeated 7 times with a setting frame width of 128. The number of times of repetition was

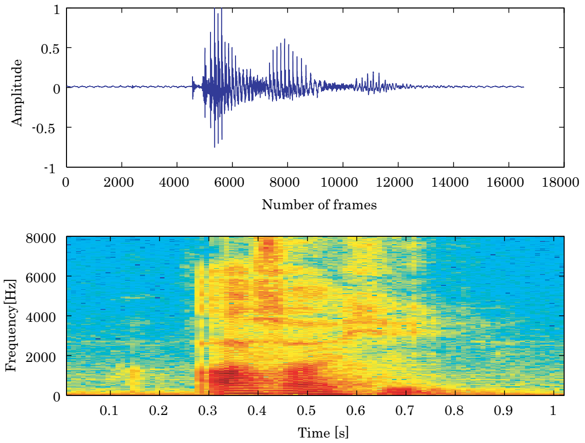

Figure 1: Speech

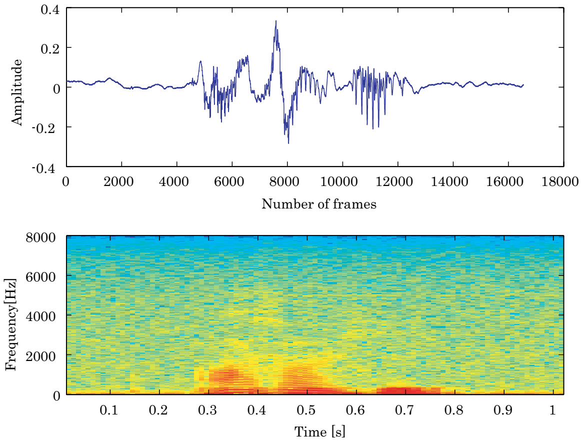

Figure 2: Body-conducted speech

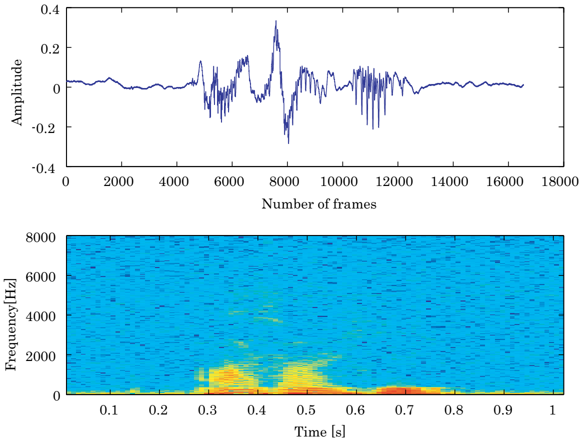

Figure 3: BCS in velocity (directly integration)

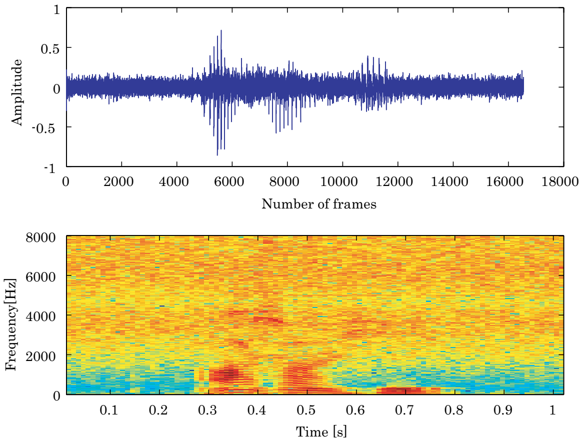

Figure 4: BCS in differential acceleration

processed using the spectral subtraction method. However, noise is not removed completely with the spectral subtraction method, the characteristics of the high frequency component cannot be fully recovered, and musical noise produces mixed results [11] [12]. From these results, we conclude that it is difficult to recover frequency characteristics with the spectral subtraction method.

4.2 Winer Filtering Method

We tried to use the spectral subtraction method as a noise reduction technique for differential acceleration. However, in the spectral subtraction method, the improvement in frequency component could not be obtained when compared with the voice. Therefore, we tried to extract a clear signal using the Wiener filtering method [7]. The Wiener filtering method is a technique which estimates a voice spectrum envelope from a noisy voice. Although the noise spectrum is simply subtracted in the spectral subtraction method, the voice spectral envelope is estimated using linear prediction coefficients and the effective frequency component can be obtained. The following formula (4) shows the Wiener filtering method.

$$H_{Estimate}(\omega) = \frac{H_{Speech}(\omega)}{H_{Speech}(\omega) + H_{Noise}(\omega)} \tag{4}$$

The estimated signal spectrum $H_{Estimate}(\omega)$ can be calculated from the estimated voice spectrum $H_{Speech}(\omega)$ and noise spectrum $H_{Noise}(\omega)$. $H_{Estimate}(\omega)$ can be expressed as a transfer function that converts a clear signal from a noisy signal. At this time, it becomes possible to estimate $H_{Speech}(\omega)$ by voice spectrum, calculating autocorrelation functions and linear prediction coefficients by the Levinson Durbin algorithm [13]. Noise spectrum $H_{Noise}(\omega)$ is then estimated by autocorrelation functions. Fig. 6 shows the resulting signal when each coefficient of both linear prediction coefficients and autocorrelation functions is 1, the frame width is 764, and there are 3 repetitions. In order to estimate $H_{Speech}(\omega)$ and $H_{Noise}(\omega)$, the same number was used for the linear prediction coefficients and autocorrelation functions because we hoped to solve the problem simply and cheaply. Even when linear prediction coefficients and autocorrelation functions were changed from 1 to 32, the frame width was changed from 128 to 4,096, and the number of repetitions was changed from 1 to 5, the best results were obtained using the conditions shown in Fig. 6. With the resulting signal, the high frequency component was close to Fig. 1, and the musical noise was not found. From this result, it is thought that the Wiener filtering method is a suitable technique for reducing stable noise in differential acceleration.

5 Evaluations

In this chapter, we discuss the effectiveness of the proposed method in each of the following experiments. Evaluation of signal distance between each retrieval signal and body-conducted speech in the parameter of Log-Area Ratio. Experiments into isolated word speech recognition.

5.1 LAR distance

We compared the effectiveness of signal retrieval with LAR distance between a voice and each signal [8]. LAR distance is the distance measurement which used reflective coefficients and linear prediction coefficients for a voice and for each retrieval signal. The distance can be calculated with linear prediction coefficients and reflective coefficients using the following formulas (5) and (6).

$$\begin{aligned} g(j) &= \frac{1}{2}\log\frac{1+\phi_j}{1-\phi_j} = \tan h^{-1}\phi_j, j = 1, 2, \ldots, P(5) \\ d_{LAR} &= \sqrt{\frac{1}{P}\sum_{j-1}^{P}[g_x(j) - g'_x(j)]^2} \qquad (6) \end{aligned}$$

P means number of linear predictive coefficients and ϕ_j is the reflective coefficient. LAR distance was calculated by having set up 12 linear prediction coefficients and having set up a frame width of 256. LAR distance is calculated by the voice and each of the following signals:

- BCS(acc): Body-conducted speech
- Pre(acc diff): Retrieval body-conducted speech
- Pre-adapted(syllable): Retrieval body-conducted speech with mora unit transfer function
- Pre-adapted(Word unit): Retrieval body-conducted speech with word unit transfer function

In this experiment, in order to compare with body-conducted speech, LAR distance for BCS between a voice and body-conducted speech was made as the baseline. From Fig.7, it is thought that the distance of the other retrieval signal is smaller than the BCS distance. In this research, since it is not necessary to compare results from the silent section and the noise section, only 17-48 frames, equivalent to the utterance section, are discussed. The results of this experiment show when compared with a BCS, the distance moves each retrieval signal closer to that of a voice. Therefore, differential acceleration is an effective technique, as shown by experimentation with LAR distance. Furthermore, clearer signal estimation could be completed by using the adaptive filter and a difference-acceleration signal. It confirmed that estimation had produced the very near signal especially in the adaptive filter of both units [5].

5.2 Recognition Experiments

Finally we evaluated the validity of the proposed method with an isolated word recognition experiment. In previous research, we constructed the body-conducted speech recognition system which performed in noisy environments, such as the engine room in the training ship [4]. This time, in order to obtain a high recognition performance, the acoustic model needed to estimate the parameter by body-conducted speech. However, if it becomes possible to estimate a voice from body-conducted speech, it will be possible to use a speech recognition decoder with body-conducted speech. It was very difficult to evaluate if the proposed method was close to a voice because a listening experiment with a large amount of data was required and there were differences between each person. Therefore, in order to perform an objective and statistical evaluation, the recognition rate of isolated word recognition was evaluated using the acoustic model for unspecified speakers built with a voice. Speech recognition matting with the feature vector from each signal and models parameters could evaluate the nearness of a voice if the recognition rate comparing body-conducted speech to each signal improved with the proposed technique. In addition, since HMM is an acoustic model which evaluates each feature vector by the value of the output of a multidimensional normal distribution, it can be judged statistically. Thus, it can confirm that the characteristic of nearness of a voice 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 [14], was used in this experiment. Since Julius was a decoder for large vocabulary continuous speech recognition, it had to be changed into isolated word recognition. Therefore, it became possible to recognize words without using a language model. JEIDA 100 local place words were used as candidates for word recognition [6]. The words were the word database of the Japanese name of a place after consideration of phoneme balance. The acoustic model used the context-dependent type tri-phone model in a word for unspecified speakers. At this time, the acceleration difference was processed with the parameter used in the previous chapter. The candidates for recognition in this experiment were each of the following signals:

- BCS : Body-conducted speech
- ret.BCS : Retrieval body-conducted speech

This experiment compared the signals processed by the acceleration difference to body-conducted speech. Tables 3 - 5 show the recognition rate in each speaker, and Table 6 shows the average word recognition rate for all people. 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 the recognition rate average. From this result, the validity of processing using acceleration difference changes a lot depending on the setup of a speaker or a parameter. It is thought that these recognition rates can be greatly improved by tuning up parameters for each speaker.

6 Conclusions

In this research, methods for estimating a clear sound from body-conducted speech which can acquire a robust signal from high noise were evaluated. We used acceleration difference for a waveform signal to examine body-conducted speech, in order to emphasize the high frequency component. Although the acceleration difference signal of body-conducted speech was mixed when stable noise was present, it was possible to remove noise effectively by noise reduction using the Wiener filtering method. The proposed method does not dependent on sampling frequency and users. It costs little to perform calculations because it uses differentiation to emphasize the signal. And when speakers differ, turning up a parameter can be used to gain higher precision signal estimation. We confirmed that the recognition experiment using the proposed method, body-conducted speech and the proposed method showed the effectiveness of an acceleration difference signal. As future work, we will examine the possibility of effectiveness to make a clear body-conducted speech that measured in noisy environments.

Table 2: Experimental Environments

Speaker	3 males
Data sets	100 words × 3 set/person
Vocabulary	JEIDA 100 local place names
Decoder	Julius-3.4
Acoustic models	gender dependent triphone
Model conditions	16 mix, clustered 3000 states

Parameters	MFCC(12)+ΔMFCC(12)
	+ΔPow(1)

Training condition	20,000 samples, HTK 2.0

Table 3: Result 1 (Speaker A: 20 years old male)

	set 1	set 2	set 3	Average
BCS	63 %	56 %	61 %	60.00 %
ret.BCS	62 %	57 %	63 %	60.67 %

Table 4: Result 2 (Speaker B: 20 years old male)

	set 1	set 2	set 3	Average
BCS	53 %	50 %	48 %	50.33 %
ret.BCS	63 %	57 %	58 %	59.33 %

Table 5: Result 3 (Speaker C: 37 years old male)

	set 1	set 2	set 3	Average
BCS	60 %	68 %	61 %	63.00 %
ret.BCS	65 %	68 %	65 %	66.00 %

Table 6: Result 4 (All speakers averages)

	Average
BCS	57.78 %
ret.BCS	62.00 %

References

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

[2] 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.

[3] H. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions", *in ASR-2000*, 181-188, 2000.

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

[5] M. Nakayama, S. Ishimitsu, and S. Nakagawa, ”A study of making clear body-conducted speech using differential acceleration”, *in 22nd IEICE SIP Symposium*, pp.622-627, 2007

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

[7] D. Li and D. O’Shaughnessy, ”Speech Processing: A Dynamic and Optimization -Oriented Approach”, *Marcel Dekker Inc*, 2003.

[8] T. Tamiya and T. Shimamura, ”Improvement of Body-Conducted Speech Quality by Adaptive Filters”, *in IEICE Technical Report*, SP2006-191, pp.41-46, 2006.

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

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

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

[12] K. Yamashita, S. Ogata, and T. Shimamura, ”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, 2005.

[13] J. Durbin, ”The Fitting of Time-Series Models”, *Review of the International Statistical Institute*, Vol.28 No.3, pp.233-244, 1960.

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

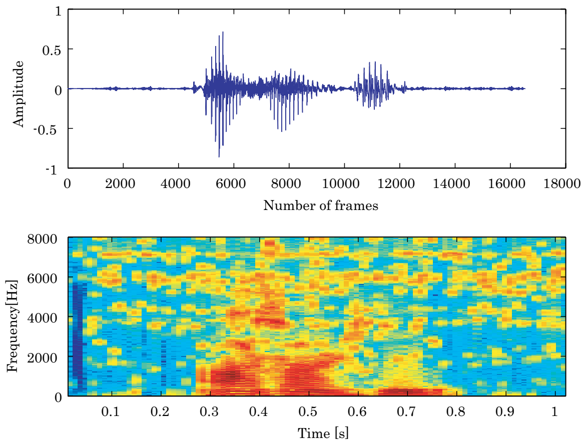

Figure 5: Retrieval speech with spectral subtraction

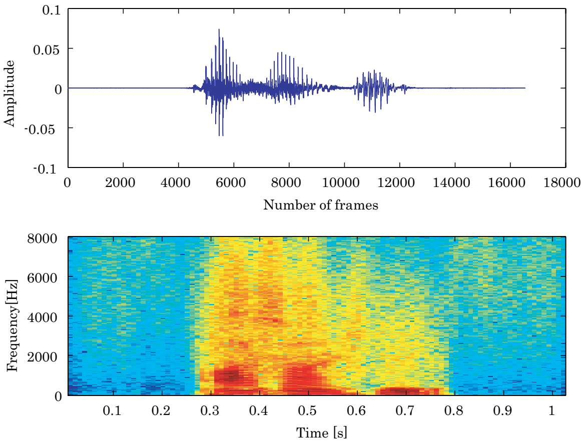

Figure 6: Retrieval speech with wiener filtering

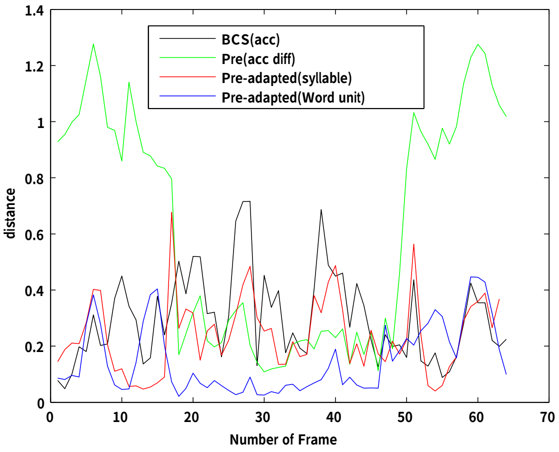

Figure 7: LAR distance