Acoust. Sci. & Tech. **33**, 1 (2012)

PAPER

Voice activity detection in noise using modulation spectrum of speech: Investigation of speech frequency and modulation frequency ranges*

Kimhuoch Pek[1,†], Takayuki Arai[1,‡] and Noboru Kanedera[2,§]

[1]*Graduate School of Science and Technology, Sophia University, 7–1 Kioi-cho, Chiyoda-ku, Tokyo, 102–8554 Japan*
[2]*Ishikawa National College of Technology, Kitachujo, Tsubata, Kahoku-gun, Ishikawa, 929–0392 Japan*

(*Received 2 March 2011, Accepted for publication 2 August 2011*)

Abstract: Voice activity detection (VAD) in noisy environments is a very important preprocessing scheme in speech communication technology, a field which includes speech recognition, speech coding, speech enhancement and captioning video contents. We have developed a VAD method for noisy environments based on the modulation spectrum. In Experiment 1, we investigate the optimal ranges of speech and modulation frequencies for the proposed algorithm by using the simulated data in the CENSREC-1-C corpus. Results show that when we combine an upper limit frequency between 1,000 and 2,000 Hz with a lower limit frequency of less than 300 Hz as speech frequency bands, error rates are lower than with other bands. Furthermore, when we use the frequency components of the modulation spectrum between 3–9, 3–11, 3–14, 3–18, 4–9, 4–11, 4–14, 4–18, 5–7, 5–9, 5–11, or 5–14 Hz, the proposed method performs VAD well. In Experiment 2, we use one of the best parameter settings from Experiment 1 and evaluate the real environment data in the CENSREC-1-C corpus by comparing our method with other conventional methods. Improvements were observed from the VAD results for each SNR condition and noise type.

Keywords: Voice activity detection, Modulation spectrum, Noise, Modulation frequency range

PACS number: 43.72.Ar, 43.72.Ne, 43.72.Dv [doi:10.1250/ast.33.33]

1. INTRODUCTION

Voice activity detection (VAD) is a very important preprocessing scheme in many fields, such as speech recognition, speech coding, and digital hearing aids. Recently, the growth of high-speed internet communication has encouraged people to electronically exchange thousands of videos around the world. Sharing foreign movies creates a large demand for translation work. When translating videos or movies, the end points of speech are often hand labeled by translators. This extra work for translators adds significant time and expense. Thus, VAD systems have played an important role in video captioning.

In previous studies, researchers have proposed various approaches to VAD and peripheral techniques. Rabiner and Sambure [1] propose a technique using the power and zero-crossing rate (ZCR) of the signal to detect word boundaries in speech. The International Telecommunication Union adopted an international speech standard known as G.729 for VAD [2]. This algorithm uses ZCR, a linear prediction spectrum, and full-/low-band energy as its features. Another method for automatic speech detection using linear prediction was also introduced [3]. This technique analyzes the characteristics of speech, determines the end points of each speech portion, and automatically creates a time code, which supports the translation work for video captioning. However, these approaches are not robust in low signal-to-noise ratio (SNR) environments. Thus, many researchers have studied different strategies for detecting speech in noise [4]. To improve detection rates, many researchers focus on different acoustic properties of the

*This paper is a summary of our previous work: Pek, Arai, Kanedera, and Yoshii, "Voice activity detection by using modulation filtering and its multi-language comparison," *Proc. Spring Meet. Acoust. Soc. Jpn.*, pp. 133–136 (2009), and Pek, Arai, Kanedera, and Yoshii, "Voice activity detection by using modulation spectrum in noise: Investigation on sound band frequency and modulation band frequency," *Proc. Autumn Meet. Acoust. Soc. Jpn.*, pp. 155–158 (2009). Additional experiments and analyses are also introduced.

†e-mail: k-pek@sophia.jp
‡e-mail: arai@sophia.jp
§e-mail: kane@ishikawa-nct.ac.jp

speech signal and noise, such as the frequency characteristics of speech. Other studies deal with noise in speech by estimating the noise spectrum and using the SNR of a signal. Sohn *et al.* [5] proposed a robust VAD technique based on a statistical model which requires prior knowledge of noise. European telecommunications standard has been recommended ES 202 050 (advanced front-end) [6] for VAD based on energy values across the whole spectra to design Wiener filters, the spectral sub-region, and spectral variance of each frame. The spectral divergence proposed by Ramirez *et al.* [7], periodic to aperiodic component ratio (PARADE) of speech covers a wide range of noise [8]. These methods work well in stationary noise but have problems with non-stationary noise.

Other acoustic characteristics have also been studied such as acoustic features based on temporal processing. Markaki and Stylianou [9], design an algorithm to exploit long term information inherent in the modulation spectrum. This study used joint acoustic and modulation frequency subspaces with higher energy. Modulation frequency components up to 250 Hz are used. This modulation frequency range is much higher than that used for automatic speech recognition [10–12] and faces a dimensionality reduction problem. Another VAD scheme based on the modulation spectrum was presented in Shadevsky and Petrovsky [13], in which a speech signal is split into multiple frequency bands and temporal envelopes are calculated for the frequency bands. In each temporal envelope, the whole modulation frequency components between 1 to 16 Hz are compared with other modulation components outside the range to enhance speech components and suppress noise components. Finally, this algorithm uses summed speech envelopes obtained from each frequency band. Unfortunately, these techniques do not investigate either an optimal speech frequency range (SFR) or an optimal modulation frequency range (MFR) in the lower modulation frequencies between 1 and 16 Hz. Moreover, it is less effective to use frequencies higher than 4,000 Hz as speech information. In contrast, noise often contains a certain amount of energy in such high frequencies. In addition, the modulation spectrum in a broad SFR is generally more flat than the one in a narrow SFR, because the temporal modulation among several frequency bands are averaged and have fewer sharp peaks from the modulation spectrum when they are combined [14]. Thus, the broad SFR may not be able to detect speech portions as accurately as the narrow SFR. In fact, different SFRs yield different temporal envelopes, and as a result, temporal contours of the feature value are different (see Section 2, Fig. 3). Therefore, limiting SFR is also very important as well as limiting MFR in noisy environments. Consequently, in this study we will find the common optimal SFR among various noise types, while previous studies, such as, Shadevsky and Petrovsky [13], use all of the frequency bands including higher frequencies.

In this paper, we propose a VAD technique based on the modulation spectrum where an optimal SFR is investigated for speech data. In Section 2 we describe this technique which uses the modulation index as a VAD feature value in background noise with an SNR of less than 30 dB. We further investigate an optimal MFR in lower modulation frequencies for each SFR by using fine bandpass filters in the modulation frequency domain to determine which components among them are important for VAD.

This paper is organized as follows. In Section 2, we describe a method to detect voice activity based on the modulation spectrum. In Section 3, we introduce Experiment 1, the aim of which is to investigate the optimal speech and modulation frequency ranges for this algorithm. In Section 4, we describe Experiment 2, which uses the best parameter settings from Experiment 1 to evaluate various data using the receiver operating characteristic curve and compare our proposed method with other conventional methods. Finally, a conclusion and future work are presented in Section 5.

2. THE MODULATION SPECTRUM

The modulation spectrum of speech is a frequency representation of temporal envelopes in which the horizontal axis is the modulation frequency and the vertical axis is the modulation index. Perceptual experiments [15] have shown that some components of the modulation spectrum are more important than others. In an environment with no noise, or a clean environment, the modulation frequency components between 1 and 16 Hz are important in preserving the intelligibility of speech [16–18]. Kanedera *et al.* [10–12] suggested that most of the information in the modulation frequency necessary for automatic speech recognition in a clean environment occurs in the same modulation frequency range, 1 to 16 Hz. In particular, the syllabic rate, at a modulation frequency of 4 Hz, is the most important component [17,19]. Additionally, researchers have used modulation components around 4 Hz to identify speech and music [20]. In keeping with these findings, using information above 16 Hz and below 2 Hz is not as useful for noisy speech, and it may actually degrade recognition performance.

In this study, we initially conducted an experiment using fixed modulation frequencies of 2–16 Hz containing speech information to investigate the optimal SFR. We then conducted a second experiment using the optimal SFR obtained from the first experiment to investigate the best MFR using our proposed method. The algorithm below is necessarily an offline method because it requires lengthy temporal information where delay is unavoidable.

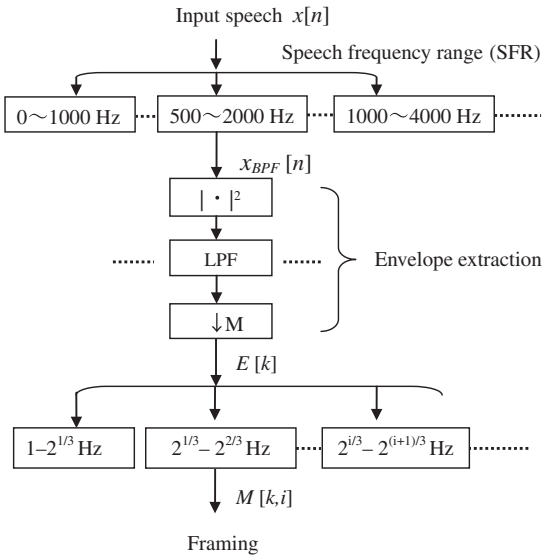

Fig. 1 Overview of feature extraction based on the modulation spectrum.

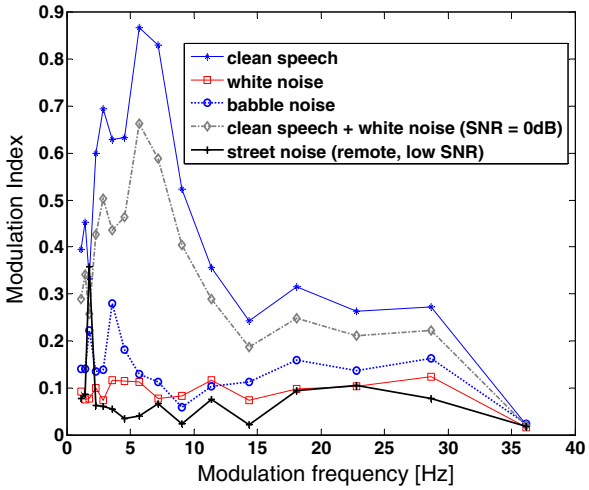

Fig. 2 Modulation spectra of speech, white noise, babble noise and street noise.

The calculation of modulation-based VAD features proposed in this study is shown in Fig. 1. First of all, we limited a full-band input speech $x[n]$ to several subbands. This is different from [21], where the input speech data were first divided into temporal frames. To determine the optimal SFR for VAD using the proposed method, we tested each one of the subbands. For each subband, we calculated the modulation spectrum based on the following algorithm.

First, we extracted the temporal envelope $E[k]$ of each subband by starting from the square amplitude of the band limited signal $x_{\mathrm{BPF}}[n]$. We put the square amplitude into a low-pass filter (LPF) with a cutoff frequency of 30 Hz and then downsampled to 80 Hz (↓M) to obtain $E[k]$, where k is a new sample index, or the frame index, for the down-sampled time domain. Next, the slowly varying modulation movements, $M[k,i]$ ($i = 0$–15), were calculated by putting $E[k]$ into a series of modulation bandpass filters with lower and upper cutoff frequencies of $f_{\mathrm{L}} = 2^{i/3}$ [Hz] and $f_{\mathrm{U}} = 2^{(i+1)/3}$ [Hz], respectively. We computed the root mean square (RMS) of $M[k,i]$ as the unnormalized modulation index. Based on the original definition by Houtgast and Steeneken [22], we determined the modulation spectrum by dividing the unnormalized modulation index by the average value of the whole temporal envelope $E[k]$, that is,

$$MI[k,i] = \frac{\sqrt{\dfrac{1}{L}\displaystyle\sum_{k'=k}^{k+L-1} |M[k',i]|^2}}{\dfrac{1}{K}\displaystyle\sum_{k'=0}^{K-1} E[k']}, \tag{1}$$

where the numerator is the RMS of $M[k,i]$, L is the length of the time frame and K is the number of whole samples of the temporal envelope. In this study, we define a modulation spectrum as the modulation index $MI[k,i]$ as a function of the modulation frequency of $mf(i) = (f_{\mathrm{L}} + f_{\mathrm{U}})/2$.

The normalized modulation indices $MI[k,i]$ were then averaged over an MFR and shifted frame by frame to obtain a single feature value per frame. The average of normalized modulation indices, $\overline{MI[k]}$, is defined as follows:

$$\overline{MI[k]} = \frac{1}{i_{\mathrm{U}} - i_{\mathrm{L}} + 1} \sum_{i=i_{\mathrm{L}}}^{i_{\mathrm{U}}} MI[k,i], \tag{2}$$

where i_{U} and i_{L} correspond the upper and lower limits of the modulation frequency bands, so that the MFR is defined as $[mf(i_{\mathrm{L}}), mf(i_{\mathrm{U}})]$.

Figure 2 shows examples of the modulation spectra of speech and noise signals in a frame. The modulation spectra of clean speech and street noise were individually calculated from the signals in the CENSREC-1-C corpus [23]; the noises were white noise and babble noise obtained from the NOISEX-92 corpus [24]. In this figure, the peak of the modulation spectrum of speech is located between 2 and 8 Hz (round off values), whereas the modulation spectrum of white noise is distributed over a wide range of modulation frequencies, with several small peaks. Because babble noise is what occurs when multiple speakers talk at the same time, the modulation spectrum of babble noise has a shape similar to that of speech in many cases. Still, the modulation index value of babble noise is often lower than that of speech. Furthermore, the three types of data: speech, white noise, and babble noise used in Fig. 2 are put into the algorithm separately, so the modulation index values of speech put in separately are higher than those of data in which speech and noise are combined, as in the case of speech and white noise.

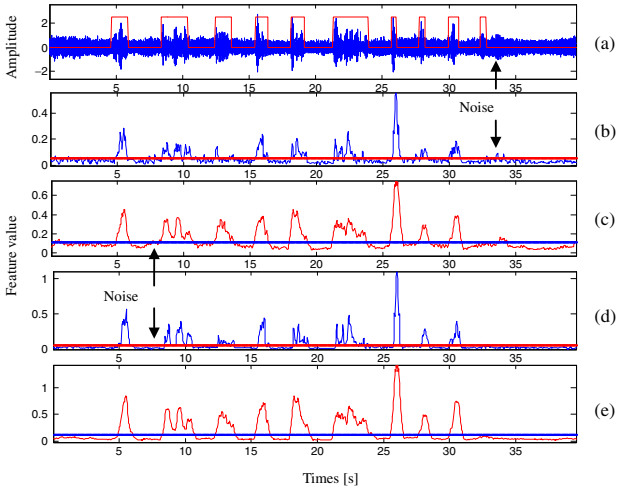

Fig. 3 A real noisy speech of street noise with low SNR. (a) Original waveform with boxes which indicate speech segments labeled by hand. (b)–(e) Contours of the feature values; (b) SFR = 0–4,000 Hz, MFR = 2–16 Hz, (c) SFR = 0–4,000 Hz, MFR = 4–16 Hz, (d) SFR = 200–4,000 Hz, MFR = 2–16 Hz, and (e) SFR = 200–4,000 Hz, MFR = 4–16 Hz. The horizontal straight lines are threshold in each panel.

Figure 3 shows the contours of the feature values based on the proposed method with various SFRs and MFRs of real noisy speech from CENSREC-1-C. Figure 3(b) and (d) show the feature contours when SFR = 0–4,000 Hz and 200–4,000 Hz, respectively. Figure 3(b) has a noisier feature contour than the one in Fig. 3(d). When using MFR = 4–16 Hz as shown in Fig. 3(c) and (e), the modulation index contours of both SFRs are less noisy and the speech portion can be detected more accurately than when MFR = 2–16 Hz. In addition, the contour is enhanced during some speech portions. Many kinds of noise contain high frequencies; and in addition, some noises have energy in the lower frequencies. Limiting the frequency bands of noisy speech may well reduce the influence of noise, but some speech information could also be lost with too much filtering. Thus, we designed the next two experiments to investigate the optimal SFR and MFR.

3. EXPERIMENT 1: INVESTIGATION OF SPEECH AND MODULATION FREQUENCY RANGES

3.1. Experimental Data

In Experiment 1, we used Japanese digits in the speech corpus CENSREC-1-C [23]. This corpus has a sampling frequency of 8 kHz, a 16 bit quantization, and an 11 word vocabulary [the digits one through nine, zero, and "maru (circle)"]. The data consisted of two types: simulated data and data recorded in real environments. The speech data in the simulated environments were subway, babble, car, exhibition, restaurant, street, airport and station, where SNR was between −5 dB and 20 dB (in 5 dB increments) and a clean environment (SNR of infinity). Speech data in real noisy environments were recorded in school cafeterias (restaurant) and at a place close to a highway (street) with two SNR levels (low and high). The low SNR was defined as a crowded restaurant (avg. 69.7 dBA) and near a main highway (avg. 69.2 dBA). The high SNR was defined as an uncrowded restaurant (avg. 53.4 dBA) and near a subsidiary highway (avg. 58.4 dBA).

3.2. Procedure and Evaluation

In a previous study [25], we used only one type of noise (subway noise) to investigate SFR; however, we recognize that different types of noise might yield different results. Therefore, in Experiment 1, we used the proposed method to examine the detection performance of speech in various environments: subway, car, babble, restaurant and street noises, which were provided in the corpus only in the simulated environment. First, we fixed the MFR of the input speech data in order to examine SFR. The fixed MFR was between 2 and 16 Hz, because most of the speech information crucial for intelligibility occurs in this range [16–18]. Next, the optimal SFR was fixed and used to examine MFR using different ranges, for example: 3–5 Hz, 4–7 Hz, and 5–10 Hz. The simulation data in the CENSREC-1-C corpus are filtered by an ITU-T G.712 filter and the frequency band is limited to between 300 and 3,400 Hz. The frequencies below 300 Hz and over 3,400 Hz might be attenuated but they are not exactly cut off at 300 Hz or 3,400 Hz. In fact, the frequencies of the simulation data tend to have speech information between 200–3,600 Hz. Consequently, we can use this speech frequency range in Experiment 1.

The frame length was set at 112.5 ms ($L = 9$), the frame shift was set to 1/3 of the frame length and K is approximately 3,200 samples depending on the input sentence. The initial threshold of the modulation index used to determine whether the target frame is speech or not was determined by using the threshold selection method in [23,26]. The optimal threshold, THR, is determined by the following equation [23]:

$$THR = THR_{\text{int}} + (r \times \alpha) \tag{3}$$

$$\alpha = (POW_{\text{h}} - POW_{\text{l}})/R, \tag{4}$$

where POW_{h} is the average of the logarithmic frame feature value (modulation index) which is equal to or greater than the initial threshold, and where POW_{l} is the average of the logarithmic frame feature value which is less than the initial threshold. The initial threshold (THR_{int}) is determined by using the threshold selection method in Otsu [26]. In this experiment, R is set to 45 and r is set to 1.

The modulation indices within MFR were averaged to obtain a single feature value $\overline{MI[k]}$; if the value was greater

than the threshold, we determined that frame to be speech, if not, we considered it non-speech.

Previous studies have shown that when viewers watch movies with captions, they prefer it when the captions appear at the same time, and disappear just as or slightly after the speaker finishes speaking [27,28]. Additionally, Kitaoka *et al.* [23] showed that extending the speech boundary by 200–300 ms can retain the accuracy of speech recognition performance. Thus, we considered the 250 ms after the end point of a speech portion to be a speech period. Our goal is to apply this technique for a subtitle-making system, which is an offline process that marks the end points of speech segments and allows users to input subtitles. Although movies contain natural speech, we use a standard corpus to compare our proposed method with other VAD techniques. This is because CENSREC-1-C is a corpus that is often used to evaluate VAD performance. Therefore, our purpose is to investigate whether our method outperforms over other well-known methods before testing its performance with natural speech.

To evaluate the results, we used false rejection rates (FRRs) and false acceptance rates (FARs), defined as follows:

$$FRR = \frac{number\ of\ incorrectly\ detected\ speech\ frames}{number\ of\ hand\ labeled\ speech\ frames} \quad (5)$$

$$FAR = \frac{number\ of\ incorrectly\ detected\ non\text{-}speech\ frames}{number\ of\ hand\ labeled\ non\text{-}speech\ frames} \quad (6)$$

When calculating multiple target data, we computed FRR and FAR of every data point first and then evaluated their mean value. Because FRR and FAR represent the error rates of speech and non-speech, lower values are desirable.

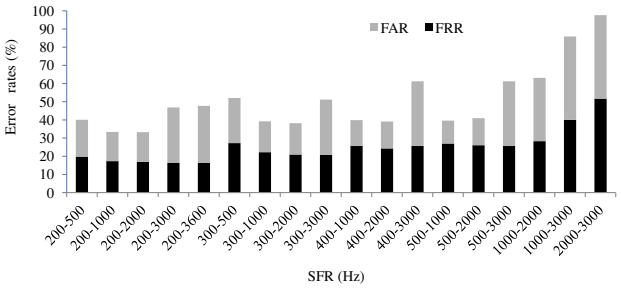

Fig. 4 Error rates of SFR (subway, SNR = 0 dB).

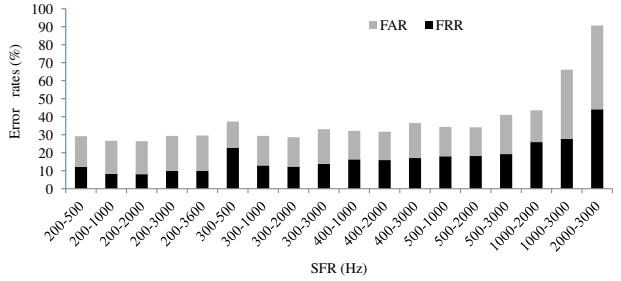

Fig. 5 Error rates of SFR (subway, SNR = 5 dB).

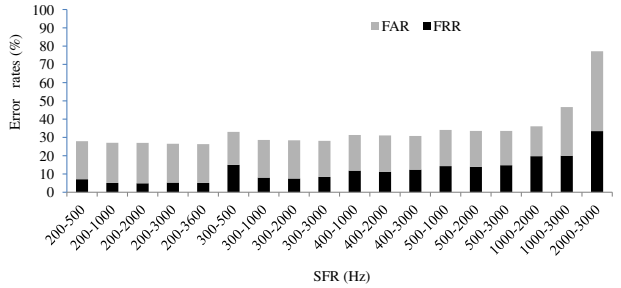

Fig. 6 Error rates of SFR (subway, SNR = 10 dB).

3.3. Experiment 1.1: Examining the Speech Frequency Ranges (SFR)

The experimental results with subway, car, babble, restaurant and street noises are shown in Figs. 4–12. Figures 4, 5 and 6 show FRR and FAR with subway noise when SNR = 0 dB, 5 dB and 10 dB, respectively. Figures 7, 8 and 9 show FRR and FAR with car noise when SNR = 0 dB, 5 dB and 10 dB, respectively. Figures 10, 11 and 12 show the FAR and FRR with babble, restaurant and street noises when SNR = 5 dB.

For subway noise, FRRs for SNR = 0 dB are around 17 to 20% when combining the lower limit of SFR = 200 Hz and the upper limit of SFR between 500 and 3,600 Hz. When the lower limit is higher than the above frequency ranges, FRRs increase. For example, when SFR = 300–500 or 500–2,000 Hz, FRR is approximately 27%. When SFR = 2,000–3,000 Hz, FRR is around 52% and FAR is about 46%. When SFR = 400–3,000, 500–3,000, or 1,000–2,000 Hz, FAR is about 36%. FAR is around 31% when SFR = 200–3,000, 200–3,600, or 300–3,000 Hz. However, FAR decreases to around 17% when SFR's lower limit between 200 and 300 Hz is combined with an upper limit of SFR between 1,000 and 2,000 Hz. The same trend was observed when SNR = 10 dB and 5 dB, although the error rates are lower than when SNR = 0 dB.

For car noise, FARs for SNR = 10 dB, 5 dB are between 17% and 20% when combining SFR's lower limit of 200–300 Hz with an upper limit of 500–3,000 Hz. The exception is SFR = 300–500 Hz and 300–1,000 Hz, which have lower error rates (about 15% for SNR = 5 dB). When the lower limit increases from 200 Hz to 300 Hz, the FRRs of these SFRs increase from 6% to 12% for SNR = 10 dB and from 8% to 16% for SNR = 5 dB. When combining the lower limit of SFR = 400 and 500 Hz with the upper limit of 1,000–3,000 Hz, the results of the FARs are constant but the FRRs tend to increase more than when

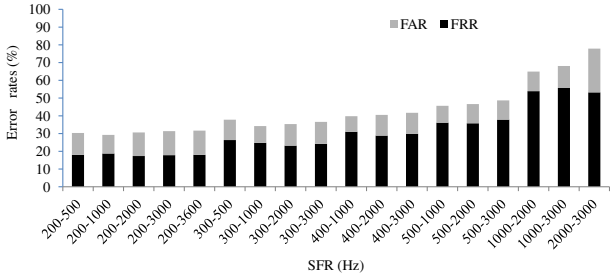

Fig. 7 Error rates of SFR (car, SNR = 0 dB).

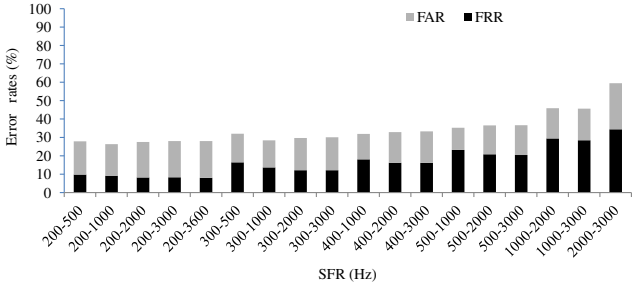

Fig. 8 Error rates of SFR (car, SNR = 5 dB).

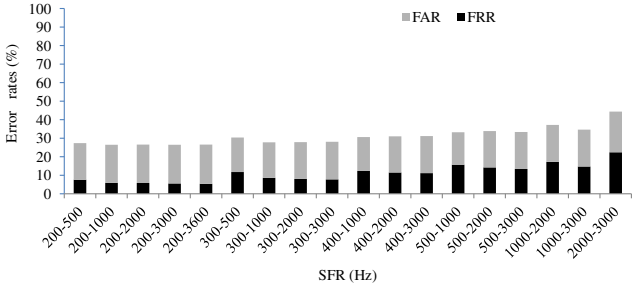

Fig. 9 Error rates of SFR (car, SNR = 10 dB).

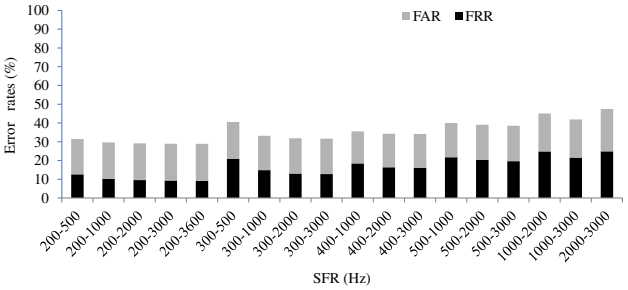

Fig. 10 Error rates of SFR (babble, SNR = 5 dB).

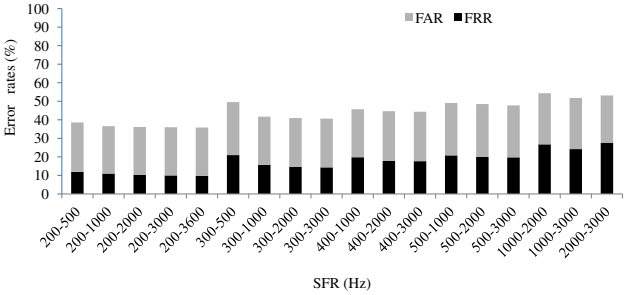

Fig. 11 Error rates of SFR (restaurant, SNR = 5 dB).

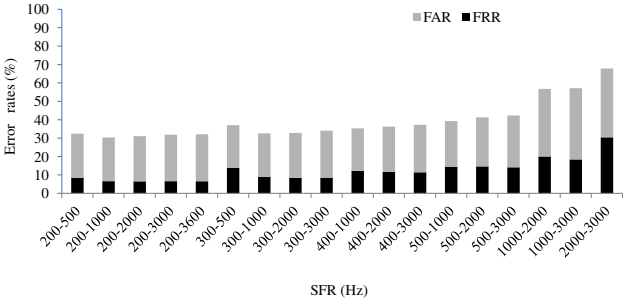

Fig. 12 Error rates of SFR (street, SNR = 5 dB).

combined with a lower limit of SFR below 400 Hz. The same trend was observed when SNR = 0 dB, although the FRRs are higher than and the FARs are lower than when SNR = 5 or 10 dB.

For babble, restaurant and street noises, the same trend was observed in car noise (SNR = 5 dB). For example, when the lower limit of SFR is increased from 200 to 500 Hz, the FRRs increase, but the FARs remain almost the same.

The results of the proposed method suggest that the upper limit frequency of 1,000–2,000 Hz performs well with all five types of noises used. The lower limit frequencies of more than 300 Hz tend to misrecognize speech as non speech portions. Among them, when combining the upper limit frequency of 1,000–2,000 Hz with the lower limit frequency of 200 Hz, FRRs and FARs are found to be lower than with other combinations. This also implies that although the error rates for subway noise are a bit different than for car noise, (Figs. 4–6), the important components of SFR are the same at all SNR levels. Such results are similar to [29] where it was reported that the speech frequencies below 1.9 kHz contain important information for speech intelligibility. Moreover, the upper limit frequency between 2,000 and 3,000 Hz can also be used, even though the FARs tend to be higher than 1,000–2,000 Hz in the case of subway (SNR = 0 dB).

3.4. Experiment 1.2: Experiment of Modulation Frequency Range (MFR)

In this section of Experiment 1, where we investigate the appropriate modulation frequency range of our proposed model, we used one of the best results we have found so far for the fixed SFR: 200–2,000 Hz. The same subway, car, babble, and exhibition noises are used as experimental data, as well as the same evaluation methods: FRR and FAR.

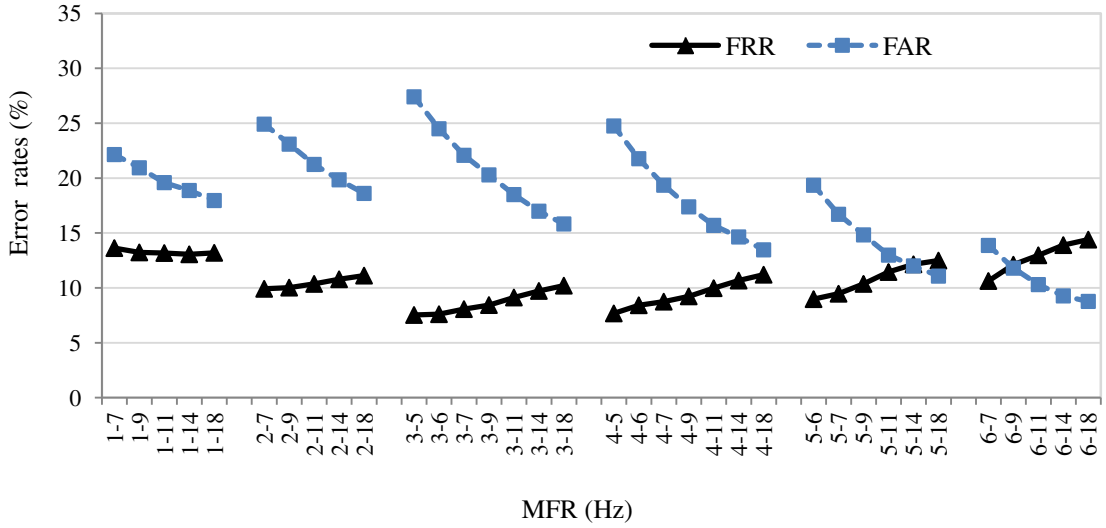

Fig. 13 Error rates of MFR (babble, SNR = 5 dB).

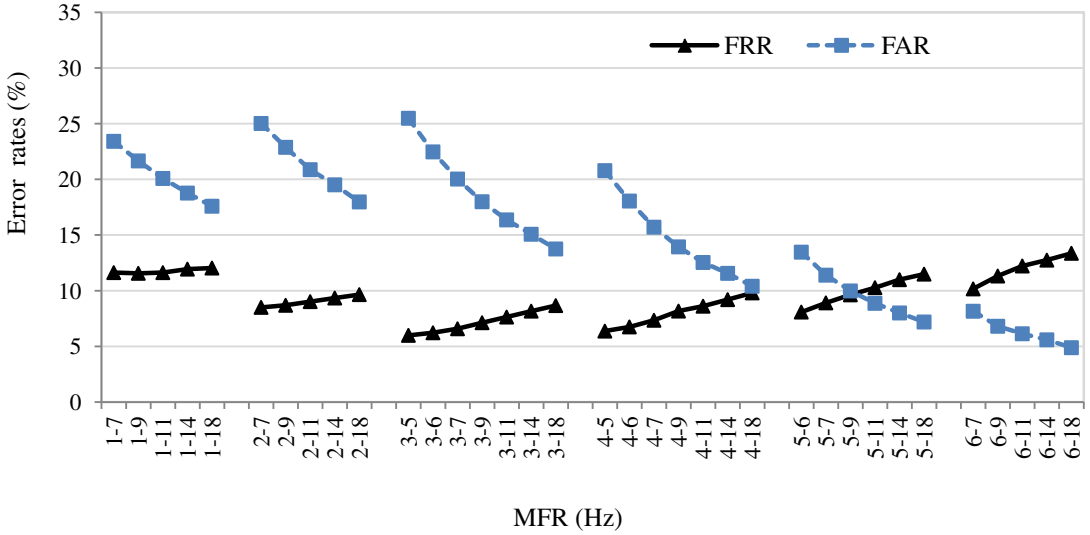

Fig. 14 Error rates of MFR (subway, SNR = 5 dB).

Figures 13–16 indicate the results of the classification of FRR and FAR errors as a function of MFR between 1–18 Hz for babble, subway, car and exhibition noises. For babble noise, when combining the lower limit of 3 Hz with the upper limit of MFR between 5–18 Hz, FRRs increased gradually from 7.5% to 10.2%, whereas FARs fell more rapidly from 27.4% to 15.8%. When combining the lower limit of 4 Hz with the upper limit between 5–18 Hz, FRRs increase slowly from 7.7% to 11.2% and FARs decrease rapidly from 24.7% to 13.5%. However, the FRR and FAR levels increased and decreased gradually in other MFR ranges. The FRR and FAR become closer to one another when MFR = 4–14, 4–18, 5–11, 5–14, 5–18, 6–7 or 6–9 Hz. Especially, when MFR = 5–11, 5–14, 5–18, and 6–9 Hz, the equal error rate is around 12%, which is the lowest value among all the ranges. However, the combinations of a lower limit of 1 Hz with upper limits between 7 and 18 Hz increased the FRRs up to 14%. Compared with MFR = 3–9 Hz, the MFR = 2–18 Hz has almost the same FAR value (approximately 19%) but the FRR value is increased by about 3%. While FRR of MFR = 2–18 Hz and 3–18 Hz shows almost the same error rate, the FAR of MFR = 2–18 Hz is higher than MFR = 3–18 Hz.

For subway and car noise, similar results are seen when the lower limit of 3–5 Hz is combined with the upper limit of 5–18 Hz. However, error rates of non-speech tend to be lower than that of speech, for example, when the lower limit higher than 5 Hz is combined with the upper limit 11–18 Hz; as well as in the case of MFR = 6–7 Hz or 6–9 Hz. The purpose of our study is to extract a speech period, thus, low speech error rates are preferred over low non-speech error rates.

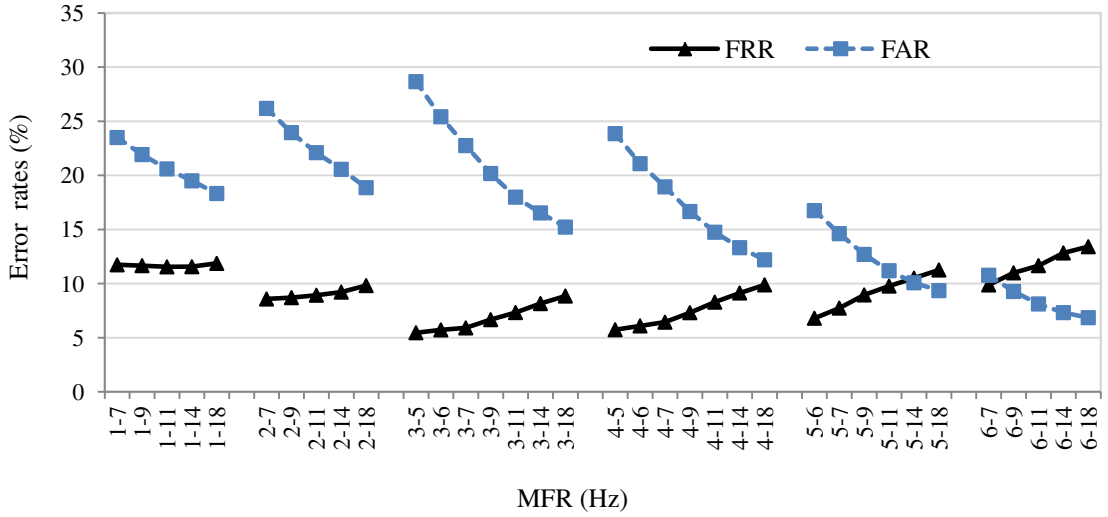

Fig. 15 Error rates of MFR (car, SNR = 5 dB).

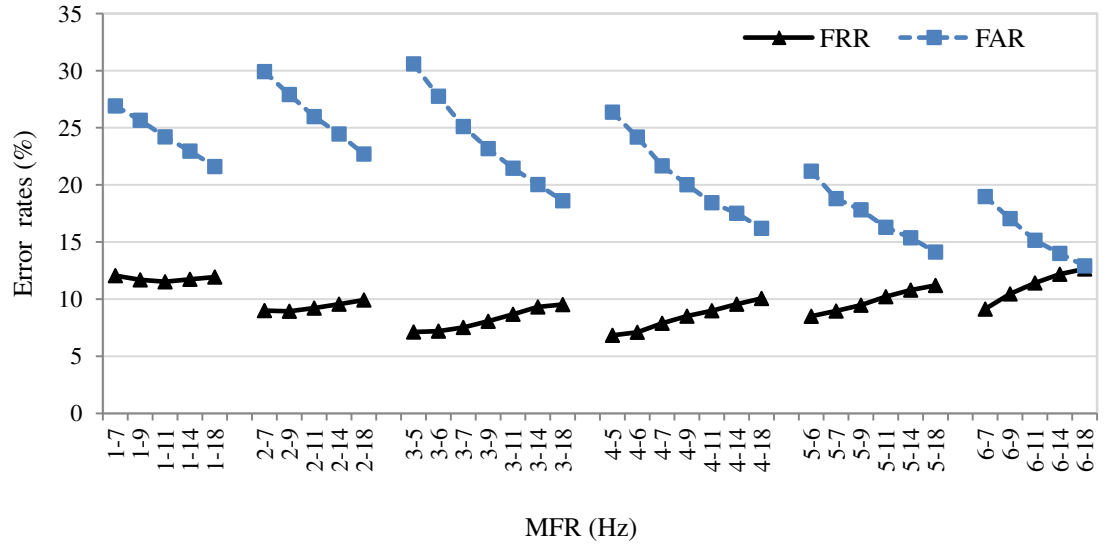

Fig. 16 Error rates of MFR (exhibition, SNR = 5 dB).

In the case of exhibition noise, the results are slightly different from the other noises when the lower limit is 6 Hz. The equal error rate appears when MFR = 6–18 Hz and the FRR value is highest among other MFRs.

According to these results, when we use the frequency components of the modulation spectrum between 3–9, 3–11, 3–14, 3–18, 4–9, 4–11, 4–14, 4–18, 5–7, 5–9, 5–11 and 5–14 Hz, the proposed method performs well for VAD. Although there are important modulation components less than 3 Hz which contribute to speech intelligibility, noises also have these modulation frequencies. Besides, previous studies indicate that the important MFR is between 1 and 16 Hz [15–17], so these outputs are considered reasonable and proper. In the next experiment, we used one of the MFRs for the feature vectors of the VAD.

4. EXPERIMENT 2

An additional experiment was conducted using the optimal SFR and MFR from Experiment 1 to compare speech detection performance with the other methods. In Experiment 1, noises are artificially added for the simulation. However, most speech data are generally recorded with real life background noises. In order to test speech data more closely resembling the speech recording environment, speech data with real noise (remote microphone) of the CENSREC-1-C corpus [23] was used for Experiment 2. Furthermore, we also compared our proposed method with other conventional methods in this experiment by using the CENSREC-1-C corpus.

4.1. Experiment 2.1

4.1.1. Experimental method and evaluation

The frame length is 112.5 ms, the frame shift is 1/3 of the frame length, and SFR is 200–2,000 Hz. There are several optimal MFRs that we obtained in Experiment 1. Since an MFR of 5–14 Hz has an equal error rate in babble noise, we chose it as the parameter in this experiment. We also used MFR = 3–18 Hz to compare its results with MFR = 5–14 Hz. We used the average value of the modulation indices of this MFR as the feature vector for VAD. Speech hit rates and non-speech hit rates were used to evaluate the results and are defined as:

$$\text{Speech hit rates } (\%) = 100 - \text{FRR}, \text{ and} \tag{7}$$

$$\text{Non-Speech hit rates } (\%) = 100 - \text{FAR}. \tag{8}$$

To investigate how well the modulation-spectrum-based features perform for VAD, we compare our results to the energy-based "Baseline" results in the CENSREC-1-C Corpus by means of ROC curves. When making the ROC curves, the speech and non-speech hit rates were calculated while changing the threshold value from 0.01 to 0.85 in 0.01 increasing steps. As in Experiment 1, we considered the 250 ms after the end point of speech to be a speech period.

4.1.2. Results and discussion

Figures 17 and 18 show the results of the ROC curves of the real noisy environment. Figure 17 shows the results for each SNR condition, and Fig. 18 shows the results for each noise type. The results of Figs. 17 and 18 are averaged by SNR levels and noise types, respectively. These results illustrate that our proposed method outperforms the Baseline in the ROC space, as the lines of our proposed method are located towards the upper right of the ROC space. As the accuracy rate gets closer to 100%, we have good speech and non-speech detection rates. According to Fig. 17, the results of MFR = 5–14 Hz improves the speech hit rate by approximately 17% and the non-speech hit rate by 11% for the low SNR level. For the high SNR level, the speech hit rate of MFR = 5–14 Hz increases by roughly 21% while the non-speech hit rate improves by 17% compared to the baseline. Also in Fig. 18, street noise has better results than restaurant noise. The speech/non-speech hit rates for street noise improve approximately 25% over the conventional method, whereas the speech hit rate for restaurant noise improves by only around 3% and the non-speech hit rate by 8%. The restaurant noise was recorded in school cafeterias where the background noise included speech, so the modulation spectrum of restaurant noise is similar to speech, while street noise is not (see Fig. 2). Thus, it is difficult to discriminate restaurant noise and speech with a modulation spectrum.

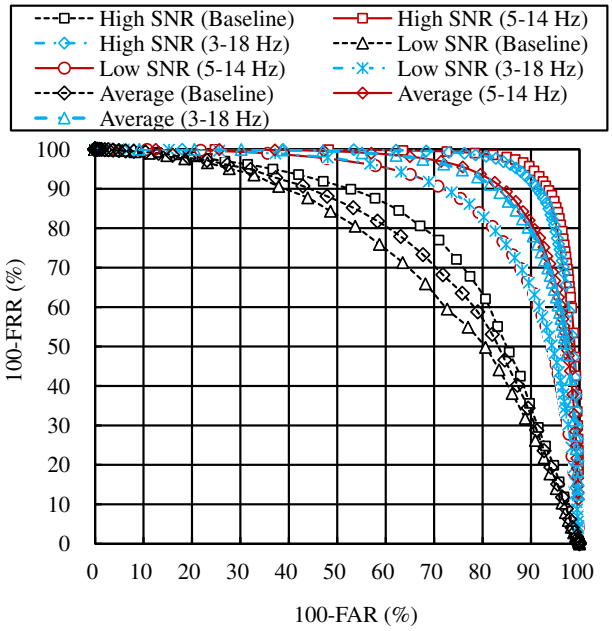

Fig. 17 ROC curves for different SNR in real environment data.

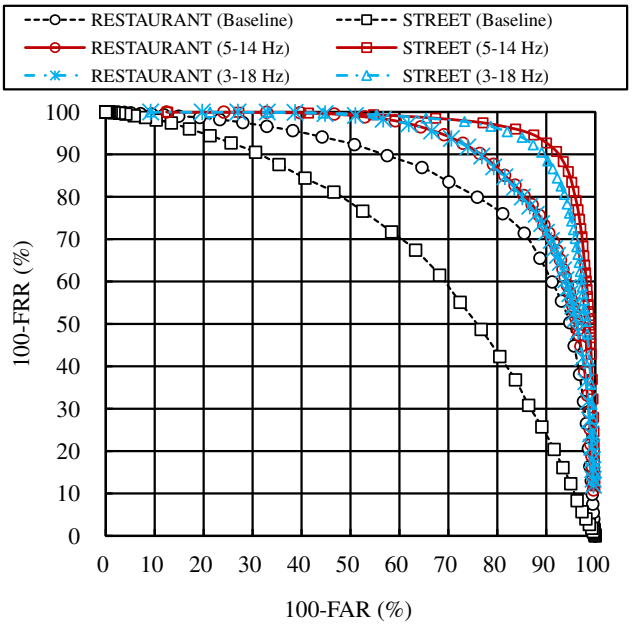

Fig. 18 ROC curves for different noises in real environment data.

However, the modulation spectrum still works well with the high SNR environment with different types of noise, including babble.

MFR = 3–18 Hz has almost the same results as MFR = 5–14 Hz except for a high SNR level (Fig. 17) and street noise (Fig. 18). These exceptions have a non-speech hit rate approximately 3% lower when MFR = 3–18 Hz than when MFR = 5–14 Hz. This is probably because when using MFR = 3–18 Hz, the FAR is higher than when MFR = 5–14 Hz, according to Experiment 1.2 (Figs. 13–16).

Table 1 VAD results of speech data with real noise.

		Rest. High	Rest. Low	St. High	St. low	Ave.	Rest. High	Rest. Low	St. High	St. low	Ave.
		Corr (%)					Acc (%)				
Conventional	Baseline	74.2	56.5	39.4	41.5	52.9	21.5	−43.5	−15.7	−33.9	−17.9
	AFE	43.8	46.7	79.1	71.9	60.4	−73.6	−94.2	−245.5	−167.0	−145.1
	G.729B	51.9	46.7	43.5	42.6	46.2	−204.4	−199.4	−72.2	−117.4	−148.3
	PARADE	70.7	57.1	87.3	80.6	73.9	24.4	−6.7	64.4	54.5	34.1
	Sohn	72.8	57.1	97.4	78.6	76.5	45.5	−6.4	94.5	57.4	47.8
Prop.	1–18 Hz	88.7	53.6	73.6	78.0	73.5	82.3	9.3	64.9	65.5	55.5
	3–9 Hz	94.2	64.3	96.2	86.4	85.3	86.4	23.2	94.8	76.5	70.2

(Rest.: Restaurant, St.: Street, Ave.: Average, Prop.: Proposed)

4.2. Experiment 2.2: Comparison with Other Conventional Methods

4.2.1. Experimental method and evaluation

We conducted two experiments using MFR = 3–9 Hz, the optimal parameter from Experiment 1, and MFR = 1–18 Hz, and compared the results. In general, the VAD method miss-detects the non speech portion as speech (FAR) for several frames at the end points of speech. However, the utterance evaluation in this experiment could address this problem by ignoring miss-detections within a margin of several hundred milliseconds around the end points of each speech portion. Moreover, MFR = 3–9 Hz has the lowest FRR among the optimal MFRs; therefore we have concluded this range will perform VAD well. The modulation indices within the MFR were averaged to obtain a single feature value per frame; if the value was greater than the threshold, we determined that frame to be speech, and if not, we considered it non-speech. The frame length is 112.5 ms, the frame shift is 1/3 of the frame length and SFR is 200–2,000 Hz.

We used utterance level performance to evaluate the results based on the utterance correct rates (Corr) and the utterance accuracy rates (Acc) defined as follows:

$$Corr\ (\%) = (N_c/N_u) \times 100 \text{ and} \tag{9}$$

$$Acc\ (\%) = ((N_c - N_f)/N_u) \times 100, \tag{10}$$

where N_c, N_f and N_u are defined as the number of correctly detected utterances, the number of incorrectly detected utterances, and the total number of speech utterances, respectively. When calculating multiple target data, we computed the N_c, N_f of every data point and then evaluated its mean value [23].

As in [23], the non-speech portion shorter than the 500 ms was defined as a speech portion. If the detected utterance was shorter than 100 ms, the detection was rejected. Both ends of the speech portions were extended by 300 ms according to the results in [23]. The optimal threshold, THR, was determined as in Experiment 1; and R was set to 45. To avoid the effect of threshold, we tested several values of r. As a result, the optimal r was set to 0 and −1 when using MFR = 3–9 Hz, MFR = 1–18 Hz, respectively.

4.2.2. Results and discussion

Table 1 shows the utterance-level evaluation results for the proposed method and other conventional methods. In the table, the row headings have the following representations: a) Baseline is the VAD technique of CENSREC-1-C (energy-based VAD with adaptive threshold) [23], b) AFE is the baseline ETSI ES 202 050 (advanced front-end) [6], c) G.729B is ITU-T G.729 Annex B [2], d) PARADE is the periodic to aperiodic component ratio based VAD [8,30], e) Sohn is the statistical model-based VAD method proposed by Sohn [5], and f) Proposed (1–18 Hz) and Proposed (3–9 Hz) represent our proposed method with different MFRs. In [31], the results of the conventional methods are provided.

Our purpose in this experiment is to compare our results with the standard methods, so we used AFE, G.729B and Sohn as a reference.

Following, we will compare our Proposed (3–9 Hz) method with the other conventional methods. For restaurant and street noises at all SNR levels, the Corr of 3–9 Hz has a significant improvement over baseline, AFE and G.729B. Comparing with PARADE and Sohn, there is significant improvement for every noise condition, except Sohn for Street noise with a High SNR level. Moreover, the average Corr of the modulation-spectrum based approach is better than for the other methods. On the other hand, the Acc of restaurant high, restaurant low and street high, street low are 86.4%, 23.2%, 94.8% and 76.5%, respectively for our proposed method. Among the conventional methods, the Acc of G.729B and AFE is very low which means there was a huge number of incorrectly detected utterances. The

Acc of both noises outperforms other methods similar to Corr. Additionally, the average of Acc of the proposed methods is 70.2% which has a significant improvement over G.729B, AFE, Baseline, PARADE and it is 22.4% improvement over Sohn. Thus, considering of the average of Corr and Acc, our proposed VAD yielded robust results when compared to the other five VAD methods analyzed.

Our Proposed (1–18 Hz) also has an improvement over the other methods except for the average of Corr for PARADE and Sohn. Also, the average of Acc of 1–18 Hz is higher in our method than in the other five methods. Still, our proposed method with an MFR of (3–9 Hz) outperforms Proposed (1–18 Hz).

5. CONCLUSION

This study presented a novel VAD algorithm for improving speech detection in noisy environments. The proposed VAD algorithm is based on the modulation spectrum of the whole speech data, and it used the modulation index as its feature.

Speech/non-speech detection using the CENSREC-1-C corpus was conducted to investigate the optimal speech frequency range and modulation frequency range. The modulation spectrum based VAD obtained the best performance in detecting speech periods when combining the lower limit of less than 300 Hz with an upper limit between 1,000 and 2,000 Hz for SFR, and when MFR = 3–9, 3–11, 3–14, 3–18, 4–9, 4–11, 4–14, 4–18, 5–7, 5–9, 5–11 and 5–14 Hz.

An analysis of the ROC curves was also conducted using the same corpus to assess the performance of the proposed algorithm by comparing it to the results provided with the corpus, and the best parameters from the first experiment are used as its parameter. The results witnessed an improvement in both speech and non-speech hit rates over the Baseline method. Moreover, when comparing with other conventional methods, our proposed methods performed well for both Corr and Acc.

In conclusion, the present study's two experiments suggest that the modulation spectrum can be applied to detect speech and non-speech periods over a wide range of SNRs (0, 5, 10, 15, 20 dB) for VAD in the future. However, further studies are required to clarify the ability of VAD with other sentence corpora and other types of additive noise, including music. Additionally, the proposed method itself can detect speech periods well, but we expect improvements as we combine it with other speech features.

ACKNOWLEDGMENT

This study work was partially supported by Sophia University Open Research Center from MEXT. The authors would like to thank Junko Yoshii, Fujiyama Inc., for her support.

REFERENCES

[1] L. R. Rabiner and M. R. Sambure, "An algorithm for determining the endpoints of isolated utterances," *Bell Syst. Tech. J.*, **54**, 297–315 (1975).

[2] ITU-T, "Annex B: A silence compression scheme for G.729 optimized for terminal conforming to recommendation V.70," *ITU-T Recommendation* (1996).

[3] Y. Fujikashi, A. Koga, T. Arai, N. Kanedera and J. Yoshii, "Linear-prediction based end-point detection of speech for captioning system," *Proc. Autumn Meet. Acoust. Soc. Jpn.*, pp. 33–34 (2005) (in Japanese).

[4] K. Ishizuka, M. Fujimoto and T. Nakatani, "Advance in voice activity detection," *J. Acoust. Soc. Jpn.* (*J*), **65**, 537–543 (2009) (in Japanese).

[5] J. Sohn, N. S. Kim and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, **6**(1), 1–3 (1999).

[6] ETSI ES 202 050 v.1.1.4, "Speech processing, Transmission and Quality aspects (STQ), Advanced Distributed Speech Recognition; Front-end feature extraction algorithm; Compression algorithms" (2006).

[7] J. Ramírez, J. C. Segura, C. Benítez, Á. de la Torre and A. Rubio, "Efficient voice activity detection algorithm using long-term speech information," *Speech Commun.*, **42**, 271–287 (2004).

[8] K. Ishizuka and T. Nakatani, "Study of noise robust voice activity detection based on periodic component to aperiodic component ratio," *Proc. SAPA*, pp. 65–70 (2006).

[9] M. Markaki and Y. Stylianou, "Discrimination of speech from nonspeech in broadcast news based on modulation frequency features," *Proc. ISCA Tutorial and Research Workshop* (*ITRW*) *on Speech Analysis and Processing for Knowledge Discovery* (2008).

[10] N. Kanedera, T. Arai, H. Hermansky and M. Pavel, "On the importance of various modulation frequencies for speech recognition," *Proc. Eurospeech*, pp. 1079–1082 (1997).

[11] N. Kanedera, H. Hermansky and T. Arai, "On properties of modulation spectrum for robust automatic speech recognition," *Proc. IEEE ICASSP, Seattle, WA*, pp. II-613–II-616 (1998).

[12] N. Kanedera, T. Arai, H. Hermansky and M. Pavel, "On the relative importance of various components of the modulation spectrum for automatic speech recognition," *Speech Commun.*, **28**, 43–55 (1999).

[13] A. Shadevsky and A. Petrovsky, "Bio-inspired voice activity detector based on the human speech properties in the modulation domain," in *Information Processing and Security Systems*, K. Saeed and J. Pejas, Eds., (Springer, New York, 2005), pp. 43–54.

[14] S. Greenberg and T. Arai, "What are the essential cues for understanding spoken language?," *IEICE Trans. Inf. Syst.*, **E87-D**, 1059–1070 (2004).

[15] R. Drullman, J. M. Festen and R. Plomp, "Effect of temporal envelope smearing on speech reception," *J. Acoust. Soc. Am.*, **95**, 1053–1064 (1994).

[16] T. Arai, M. Pavel, H. Hermansky and C. Avendano, "Intelligibility of speech with filtered time trajectories of spectral envelopes," *Proc. ICSLP*, pp. 2490–2493 (1996).

[17] T. Arai, M. Pavel, H. Hermansky and C. Avendano, "Syllable intelligibility for temporally filtered LPC cepstral trajectories," *J. Acoust. Soc. Am.*, **105**, 2783–2791 (1996).

[18] T. Arai and S. Greenberg, "The temporal properties of spoken Japanese are similar to those of English," *Proc. Eurospeech*, Vol. 2, pp. 1011–1014 (1997).

[19] S. Greenberg, "Understanding speech understanding: Towards

a unified theory of speech perception," *Proc. ESCA Workshop on the Auditory Basic of Speech Perception*, pp. 1–8 (1996).
[20] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," *Proc. ICASSP*, pp. 1331–1334 (1997).
[21] K. Pek, T. Arai, N. Kanedera and J. Yoshii, "Voice activity detection by using modulation filtering and its multi-language comparison," *Proc. Spring Meet. Acoust. Soc. Jpn.*, pp. 133–136 (2009) (in Japanese).
[22] T. Houtgast and H. J. M. Steeneken, "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," *J. Acoust. Soc. Am.*, **77**, 1069–1077 (1985).
[23] N. Kitaoka, T. Yamada, S. Tsuge, C. Miyajima, K. Yamamoto, T. Nishiura, M. Nakayama, Y. Denda, M. Fujimoto, T. Takiguchi, S. Tamura, S. Matsuda, T. Ogawa, S. Kuroiwa, K. Takeda and S. Nakamura, "CENSREC-1-C: An evaluation framework for voice activity detection under noisy environments," *Acoust. Sci. & Tech.*, **30**, 363–371 (2009).
[24] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition II: NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, **12**, 247–251 (1993).
[25] K. Pek, T. Arai, N. Kanedera and J. Yoshii, "Voice activity detection by using modulation spectrum in noise: Investigation on sound band frequency and modulation band frequency," *Proc. Autumn Meet. Acoust. Soc. Jpn.*, pp. 155–158 (2009) (in Japanese).
[26] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst. Man Cybern.*, **SMC-9**, 62–66 (1979).
[27] I. Maruyama, Y. Abe, E. Sawamura, T. Mitsuhashi, T. Ehara and K. Shirai, "Cognitive experiments of time lag for superimposing captions," *Proc. IEICE General Conf.*, p. 323 (1999) (in Japanese).
[28] A. Koga, K. Matsuura, T. Arai, N. Kanedera and J. Yoshii, "Study on end-point detection of speech and timing of caption for video contents," *Proc. Autumn Meet. Acoust. Soc. Jpn.*, pp. 445–446 (2007) (in Japanese).
[29] D. M. Jones, *"Noise," Stress and Fatigue in Human Performance*, R. Hockey, Ed. (Wiley, New York, 1983), Chap. 3, pp. 61–95.
[30] M. Fujimoto, K. Ishizuka and T. Nakatani, "A voice activity detection based on the adaptive integration of multiple speech features and a signal decision scheme," *Proc. ICASSP*, pp. 4441–4444 (2008).
[31] M. Fujimoto, S. Watanabe and T. Nakatani, "Voice activity detection using frame-wise model re-estimation method based on Gaussian pruning with weight Normalization," *Proc. Interspeech*, pp. 3102–3105 (2010).

Kimhuoch Pek received the B.A and M.A. degrees in electrical engineering from Sophia Univ., Tokyo, Japan, in 2007 and 2009, respectively. She is currently working in her PhD at the Graduate School of Science and Technology, Sophia Univ. Her research interests include acoustics, spoken language processing and signal processing.

Takayuki Arai received the B.E., M.E. and Ph.D. degrees in electrical engineering from Sophia Univ., Tokyo, Japan, in 1989, 1991 and 1994, respectively. In 1992–1993 and 1995–1996, he was with Oregon Graduate Institute of Science and Technology (Portland, OR, USA). In 1997–1998, he was with International Computer Science Institute (Berkeley, CA, USA). He is currently Professor of the Department of Information and Communication Sciences, Sophia Univ. In 2003–2004, he was a visiting scientist at Massachusetts Institute of Technology. His research interests include acoustics, speech communication (speech and hearing sciences), spoken language processing, and signal processing.

Noboru Kanedera received the B.E. degrees from the University of Electro-Communications in 1985 and M.E. degrees in division of engineering from the University of Tokyo in 1987. In 1996, he was a visiting scientist at Oregon Graduate Institute of Science and Technology (Portland, OR, USA). He is currently Professor of the Department of Electronics and Information Engineering, Ishikawa National College of Technology. He is a Doctor of Engineering. His research interests include speech recognition, natural language processing and neural network.