Voice activity detection in noisy environment by using modulation spectrum: Modulation frequency range using sentence corpus *

☆Kimhuoch Pek[1] , Takayuki Arai[1] and Noboru Kanedera[2]
([1]Sophia University, [2]Ishikawa National College of Technology)

1 Introduction

Voice Activity Detection (VAD) determines if a sound segment is speech or non-speech, and 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. Usually, when the input speech signal contains noises, the accuracy of discrimination between speech and non-speech decreases.

In previous studies, researchers have used different strategies for detecting speech in noise. Sohn *et al.* [1] proposed a robust VAD technique based on a statistical model by yielding smoother estimates of the *a priori* SNR. Ramirez *et al.* [2] proposed the spectral divergence between speech and noise. Periodic to aperiodic component ratio (PARADE) of speech, proposed by Ishizuka *et al.* [3], covers a wide range of noise. The above methods work well in stationary noise but encounter problems with non-stationary noise.

To solve this difficulty, we have developed a VAD method for noisy environments based on the modulation spectrum and have successfully tested our proposed method with a digit corpus over several types of noise and SNR conditions [4]. Since the purpose of our study is to apply our proposed method to a subtitle-making system, we still need to test our method with sentence material that is more natural than that in the digit corpus.

2 Modulation Spectrum

The modulation spectrum is a type of frequency representation of the slowly varying temporal envelope components of speech [5, 6]. In the representation, the horizontal axis is the modulation frequency and the vertical axis is the modulation index.

Kanedera *et al.* [5] suggested that most of the information in the modulation frequency necessary for automatic speech recognition in a clean environment is found between 1 and 16 Hz. In an environment with no noise, or a clean environment, the modulation frequency components between 1 and 16 Hz are important for preserving the intelligibility of speech [6].

Our previous study [4] investigated the optimal ranges of speech and modulation frequencies for the proposed algorithm by using the simulated data in the CENSREC-1-C corpus which contains only Japanese digits. Results show that when we combine an upper limit frequency between 1000 and 2000 Hz with a lower limit frequency of less than 300 Hz as speech frequency ranges (SFR), error rates are lower than with other bands. Furthermore, when we use the modulation frequency ranges (MFR) 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 well.

In this study we used our previous study of the modulation spectrum-based VAD [4] to investigate the optimal modulation frequency range (MFR) with a sentence corpus. We will begin by comparing the optimal MFRs from our previous study [4] with other MFRs to see whether the sentence material has different optimal MFRs than the digit corpus.

3 Experiment

3.1 Experiment data and evaluation method

In this experiment, we used Japanese Newspaper Article Sentences (JNAS) [7]. The

*変調スペクトルを用いた雑音環境下における音声区間検出 : 文章読み上げコーパスにおける変調周波数範囲，ペク　キムホーチ，荒井　隆行 (上智大)，金寺　登 (石川高専).

speech signals were sampled at 16 kHz and quantized into 16 bits. The corpus contains speech recordings with orthographic transcriptions of 306 Japanese native speakers (153 males and 153 females). We used 5 types of Japanese sentences read by 12 males and 12 females, totaling 120 sentences, in this experiment. The average length of each datum is approximately 10 seconds. Three different noise types from the NOISEX-92 [8] corpus were artificially added to the JNAS data in this experiment: white noise (white), babble noise (babble) and noise on floor of car factory (factory1). The SNR was between 5 dB and 0 dB and clean speech data. The clean speech data was recorded in clean environment (SNR of infinity).

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

$$FRR = \frac{number of\ incorrectly\ detected\ speech\ frames}{number of\ hand\ labeled\ speech\ frames} \quad \text{and}$$

$$FAR = \frac{number of\ incorrectly\ detected\ non-speech\ frames}{number of\ hand\ labeled\ non-speech\ frames}$$

In this experiment, the MFR parameter started at 2 Hz, the optimal value obtained from [4]. We fixed SFR to 200–2000 Hz, the frame length was set to 112.5 ms (L = 9) and the frame shift was set to 1/3. The optimal threshold is determined by using the threshold selection method in Otsu [9]. Since we also want to investigate how accurately the proposed method can detect a speech portion, we do not use any decision rule (such as the 250-ms rule) after the detected end point of speech in this experiment.

3.2 Results

Figures 1–4 show the results for FRR and FAR errors as a function of MFR between 2–36 Hz for clean, white noise, babble noise and factory1 noise when SNR = 5 and 0 dB.

For clean results (Fig. 1), when combining the lower limit of between 2–5 Hz with the upper MFR limit of between 14–36 Hz, FRRs decrease gradually from 13.0% to 12.2%, whereas FARs fall more rapidly from 31.6% to 18.8%. Equal error rates are obtained when combining the lower limit of between 9–18 Hz with the upper MFR limit of between 14–36 Hz and MFR = 23–23 Hz. FARs tend to be lower than FRRs when a lower limit higher than 23 Hz is combined with an upper limit of 29–36 Hz.

Similar results were found for white with SNR = 5 dB or 0 dB (Fig. 2). White noise is stationary noise; with modulation spectrum contours that are different from speech [4]. Therefore, even in the low SNR level, the accuracy is not significantly different from clean speech.

For babble noise (Fig. 3) with SNR = 5 dB, while the FRR stays at approximately 12%, the FAR decreases 17.7% from 32.2% (MFR = 2–14 Hz) to 14.5% (MFR = 36–36 Hz). However, the FAR increases when the lower MFR limit is higher than 14 Hz for SNR = 0 dB. When combining the lower limit of between 5–9 Hz with the upper MFR limit of between 14–36 Hz, the FAR is around 26%, which is the lowest rate for SNR = 0 dB.

For the factory noise with SNR = 5 dB, when we combine the lower limit of 2–9 Hz with the upper MFR limit between 14–36 Hz, FRRs remain at approximately 12%, whereas FARs decline rapidly from 31.4% to 12.2%. The FARs become lower than the FRRs in other ranges. According to Fig. 4, when SNR = 0 dB, the FARs tend to stay the same (about 12%) when the lower limit is over 9 Hz. When MFR = 36–36 Hz, the FAR value climbs rapidly to 22%.

These results reveal that the method detects speech portions well when we combine a lower limit of between 3–36 Hz with an upper MFR limit of between 14–36 Hz. Among the above optimal MFRs value for sentence corpus, the combination of a lower limit between 5–14 Hz and an upper MFR limit between 14–36 Hz yields the best VAD performance for our method.

3.3 Discussion

Previous studies [7, 8] suggest that most of the information in the modulation frequency necessary for automatic speech recognition in a clean environment is found in the 1 to 16 Hz range. Testing our method with the digit corpus in [4], we found it performs VAD well when the MFR = 3–9, 3–11, 3–14, 3–18, 4–9, 4–11, 4–14,

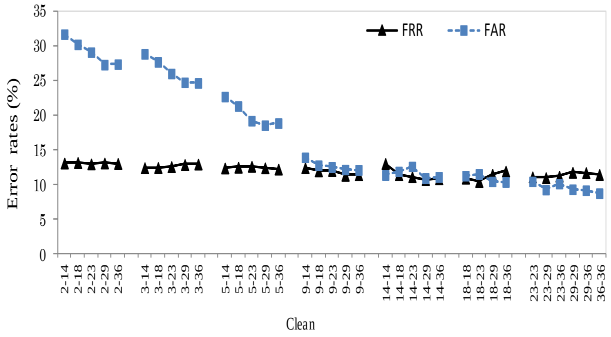

Fig. 1 Error rates (clean speech)

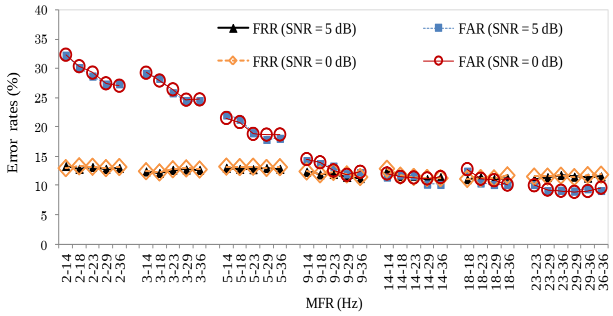

Fig. 2 Error rates (white)

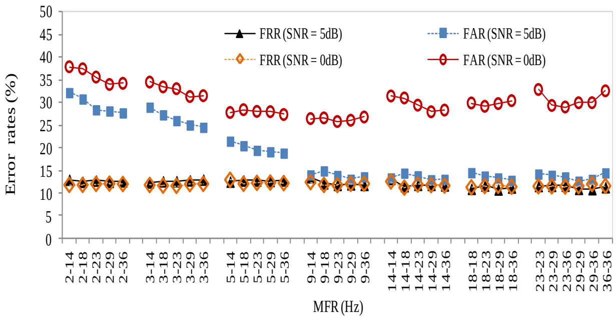

Fig. 3 Error rates (babble)

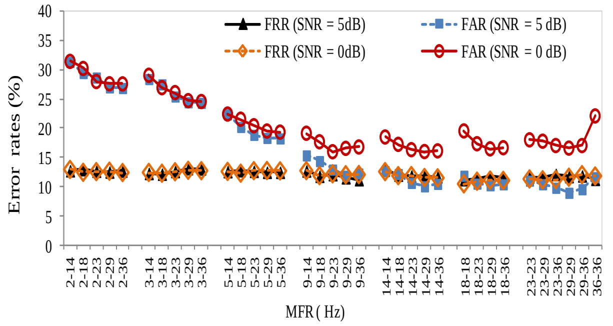

Fig. 4 Error rates (factory1)

4–18, 5–7, 5–9, 5–11 and 5–14 Hz. In this experiment, when using the sentence corpus, we found the most accurate speech detection performance when the lower limit of the MFR was between 5–14 Hz. Thus, the results for the digit and sentence corpora are slightly different. A higher optimal lower limit of MFR is preferred with the sentence corpus than with the digit corpus. This can be seen in Figs. 1–4, where, though the miss-detection of speech as non-speech (FRR) is fairly constant throughout all ranges, the miss-detection of non-speech as speech (FAR) decreases significantly.

Figures 5 and 6 show a speech waveform, a feature value contour, correct labels and detection results for clean speech (JNAS) (Fig. 5) and babble noise as additive noise (SNR = 0 dB) (Fig. 6) with different MFRs, respectively. Fig. 5 (a) and (b) tend to miss-detect non-speech as speech approximately 15% more than in Fig. 5 (c), (d) and (e). The trend of miss-detection happens between speech portions. The main reason is that the feature contour when MFR = 3–14 Hz or 4–14 Hz is not able to drop to near zero in the non-speech portion between speech periods. These results could be improved by using MFR = 5–14, 9–14 or 14–14 Hz. However, according to Figs. 6 (d) and (e), when MFR = 14–14 Hz, non speech could more easily be detected as speech when the noise energy is equal to or higher than speech energy.

Figure 7 indicates the speech waveform, feature value contour, correct labels and results for the clean speech data of CENSREC-1-C. In this case, the proposed method was able to drop to a value close to zero between digit utterances (non speech periods) even when MFR = 3–14 Hz. This is different from the JNAS results in Figures 5 and 6. We think this is due to the length between speech portions. While the time between the digit utterances in CENSREC-1-C is approximately two seconds, the interval between the sentences in JNAS is only about 400 milliseconds. The optimal MFRs seem to depend on the length of this interval between utterances. When the time between utterances is longer, the lower limit of MFR higher than 5 Hz is appropriate to use. Note, however, that when using the lower limit of MFR over 14 Hz, the error rates of non-speech detection raised from about 25% to 30% for babble noise with SNR = 0 dB. Therefore, it is not always best to use this lower limit of MFR. In addition, the data of JNAS corpus is more natural than digits corpus

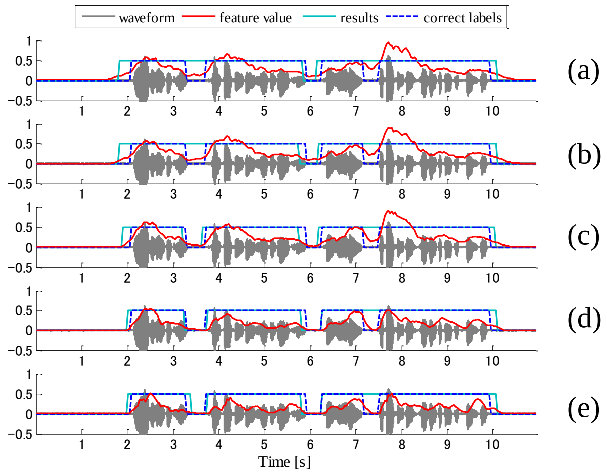

Fig. 5 An example of clean speech from the JNAS corpus.

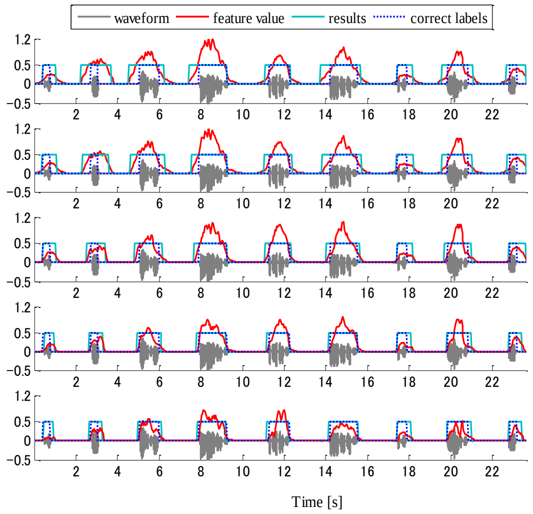

Fig. 7 An example of clean speech from the CENSREC-1-C corpus.
In Fig. 5 to 7: (a) MFR = 2–14 Hz, (b) MFR =3–14 Hz, (c) MFR = 5–14 Hz, (d) MFR = 9–14 Hz and (e) MFR = 14–14 Hz.

and the intervals between utterances are not that long as the intervals in CENSREC-1-C corpus. On the basis of this observation, the optimal MFRs could be in between the results of the two corpora.

4 Conclusions

This study investigate the performance of modulation spectrum based VAD in a sentence corpus. The results of the experiment show that the optimal lower limit of MFRs for the sentence experiment tended to be higher than for the digit experiment. However, we found that the common optimal lower limit of these two corpora is

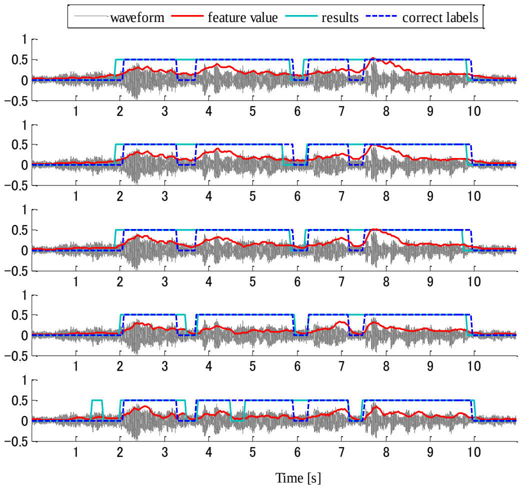

Fig. 6 An example of clean speech from the JNAS corpus with babble noise added when SNR = 0 dB.

approximately around 5 to 9 Hz and upper limit is approximately 11 to 23 Hz when performing the VAD at a low SNR level (5 or 0 dB). For further studies, we need to clarify the effectiveness of VAD with other types of sentence corpora and additive noise.

Acknowledgements

This study was partially supported by Sophia University Open Research Center from MEXT.

References

[1] J. Sohn *et al.*, *IEEE SP Letters*, vol. 6, no. 1, 1–3, 1999.

[2] J. Ramírez *et al.*, *Speech Commun*, vol. 42, no. 3–4, 271–287, 2004.

[3] K. Ishizuka *et al.*, *Proc. of SAPA '06*, 65–70, 2006.

[4] K. Pek *et al.*, *Proc. Autumn Meet. Acoust. Soc. Jpn.*, 155-158, 2009.

[5] N. Kanedera *et al.*, *Proc. Eurospeech*, pp. 1079-1082, 1997.

[6] T. Arai *et al.*, *Proc. ICSLP*, pp. 2490-2493, 1996.

[7] K. Itou *et al.*, *J. Acoust. Soc. Jpn.*, 199-206, (1999).

[8] A. Varga *et al.*, *Speech Commun.*, 12(3), 247-251 (1993).

[9] N. Otsu, *IEEE Trans. Syst. Man, Cybern.*, SMC-9, 62-66, 1979.