Voice activity detection in noisy environments using modulation spectrum and its performance in a subtitle-making system *

☆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) is a very important preprocessing scheme in speech communication, speech recognition, speech coding, and speech enhancement under noisy environments. Recently, VAD systems have also played an important role in captioning video contents.

In general, when translating videos or movies, the start and end points of each speech portion are hand-labeled by translators. This extra work for translators, in addition to the translation, adds time and expense.

In [1], a system was introduced that analyzes the characteristics of speech, determines the end points of each speech portion, and automatically creates a time code. Unfortunately, this system only performs well in low noise conditions whereas most speech data are recorded in environments containing noise or background music. Therefore, a robust VAD algorithm that works in noisy environments is needed.

In previous studies, researchers have used different strategies for detecting speech in noise. Sohn *et al.* [2] proposed a robust VAD technique based on a statistical model which requires prior knowledge of the noise. European telecommunications standard has been recommended ES 202 050 for VAD based on energy values [3]. The spectral divergence proposed by Ramirez *et al.* [4], with a periodic to aperiodic component ratio of speech, covers a wide range of noise [5, 6]. These methods work well in stationary noise but have problems with non-stationary noise. In this paper, we use a VAD algorithm based on the modulation spectrum to detect end points of speech portions in background noise with an SNR of less than 30 dB. We use the algorithm in a subtitle-making system and compare its performance with hand-labeling.

2 Modulation Spectrum

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

Kanedera *et al.* [7] 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 [8].

In this paper we used our previous study of the modulation spectrum [9] to investigate VAD performance with the real, noisy speech of a moving picture. Then, we input the VAD results into a subtitle-making system and asked users to evaluate the system. Fig. 1 shows the contour of the modulation spectrum used to detect speech portions in our experiment.

3 Experiment

The aim of the experiment is to determine whether the modulation spectrum reduces the time needed for hand- labeling and whether users will accept it. First, we conducted an experiment to detect speech portions by using the modulation spectrum. In this experiment, we use a speech frequency range of 200–2000 Hz and a

* 変調スペクトルを用いた雑音下における音声区間検出：字幕付与システムでの性能評価，ペク キムホーチ，荒井　隆行 (上智大)，金寺　登 (石川高専).

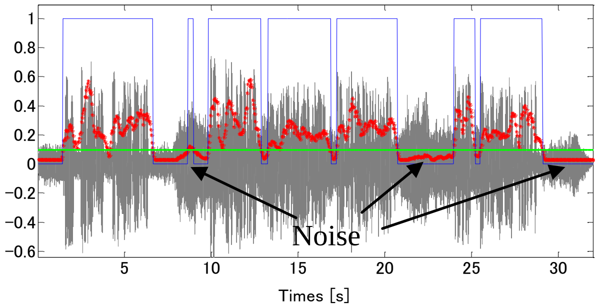


Fig. 1 An Example of the modulation spectrum contour of a noisy speech. The straight line is the threshold and the blue boxes are the VAD results.

modulation frequency range of 5–14 Hz, the optimal parameter from [9], to detect the speech portions of a moving picture. The frame length is 112.5 ms; the frame shift is 1/3 of the frame length. The optimal threshold was determined by using the threshold selection method in Otsu [10]. Finally, we put the results into a subtitle-making system and ask users to evaluate it.

3.1 Participants

The forty native Japanese speaking participants in this experiment are divided into two groups: Those who have background making subtitles or editing moving pictures and those who don't. The first group has experience editing moving pictures either as a hobby or work. There were 17 participants in the experienced group (5 female, 12 male) and 23 participants in the non-experienced group (11 female, 12 male). Their ages range from 19 to 34 years old. None of the participants reported any hearing or computer operating problems.

3.2 Procedure

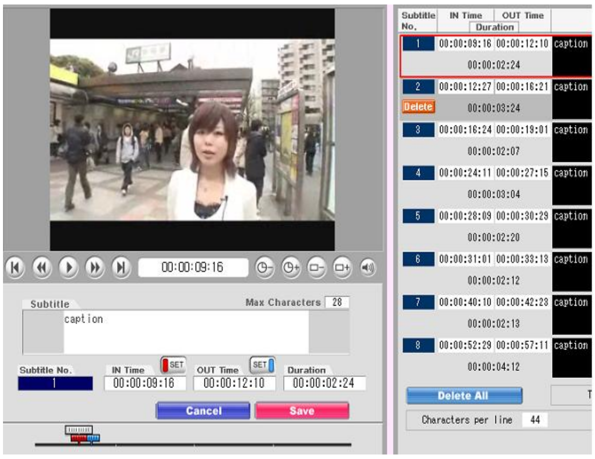
The VAD results mark the start and end points of speech where a caption or subtitle is to be input by participants. Conventionally, a translator or user must extract the time label (the start and end points of speech) by hand before putting a caption into the time label. This work adds extra time and expense. We used the VAD results of our proposed method to extract the time labels automatically, and participants only needed to add captions. The hope was to save time and reduce the burden on the users.

We used moving picture data recorded in real noisy environments near the street from Harajyuku to Shibuya station, one of the most crowded streets in Tokyo, having an average SNR of 4 dB. There was one Japanese female speaker. The length of one experimental segment is approximately 1 minute and 20 seconds. Our method detected 13 pairs of start and end points where captions needed to be input. Since the VAD results might be incorrectly divided one sentence into two time labels in the middle of a sentence, or two consecutive sentences could be detected as one sentence, they needed to be checked with a caption script. (The type of caption to be placed was predetermined within each correct pair.)

In our experiment, VAD produced 11 correct and 2 incorrect pairs of time labels. One of the two errors was noise that was miss-detected as speech, and the other error was a single speech period containing two consecutive correct pairs of time labels.

We put the VAD time labels into some subtitle-making software [11] that edits time labels and enters/edits captions by participants, as shown in Figure 2. In the right column, the box refers to the time labels determined by our VAD method.

Participants spent about 15 minutes getting acquainted with the software and then the experiment began. We asked participants to click on a box, select a time label, and enter a caption. A laptop computer was used to perform this task.


00:00:09:16

Subtitle — Max Characters 28
caption

Subtitle No. 1 | IN Time 00:00:09:16 | OUT Time 00:00:12:10 | Duration 00:00:02:24

Subtitle No.	IN Time	OUT Time	
	Duration		
1	00:00:09:16	00:00:12:10	caption
	00:00:02:24		
2	00:00:12:27	00:00:16:21	caption
Delete	00:00:03:24		
3	00:00:16:24	00:00:19:01	caption
	00:00:02:07		
4	00:00:24:11	00:00:27:15	caption
	00:00:03:04		
5	00:00:28:09	00:00:30:29	caption
	00:00:02:20		
6	00:00:31:01	00:00:33:13	caption
	00:00:02:12		
7	00:00:40:10	00:00:42:23	caption
	00:00:02:13		
8	00:00:52:29	00:00:57:11	caption
	00:00:04:12		

Characters per line 44

Fig. 2 The interface for the experiment

Participants also labeled the data the conventional way, placing the time labels by hand and adding captions. Finally, the participants in both methods checked whether the labels were correct or not by comparing the contents of the captions to the moving picture.

We measured the time required to complete both tasks and asked participants to respond to questions about the work. In order to reduce bias between the VAD method and the manual method, we took a counter balance by switching the order in which the two tasks were performed (the VAD method then the manual method, and vice versa).

3.3 Results and discussion

Figures 3 and 4 indicate the results of the processing time and the questionnaire, respectively. The average processing time in the two methods by the experienced-group and the non-experienced-group are shown in Fig. 3.

Figure 4 shows the results for when we asked the participants, "How did you feel when the time labels were provided in advance?" It was a multiple choice question, with four options, as shown in Figure 4.

Figure 3 shows that both groups took less time to complete the task when the time labels were provided in advance. The improvements were approximately 30% (four minutes) for both groups. The experienced group finished in less time than the non-experienced group in both conditions. Even though all participants practiced with the software before the task, the experienced group still performed more quickly than the non-experienced group because of their skill level.

94.1% of the experienced group and 82.6% of the non-experienced group favored the automatic VAD label method. One non-experienced participant (4.3%) preferred labeling by hand, but no experienced participant preferred hand labeling. This participant preferred a longer endpoint (250ms+) than was provided, and found it necessary to edit every caption. This person seemed annoyed by the editing task.

In reference to whether participants preferred the automatic VAD labels over hand labeling, 5.9% of experienced-group and 13.0% of non-experienced-group chose the answer "It is hard to say". However, none of the participants in either group said that providing time labels in advance was not helpful for them. Participants who felt negative about VAD said if the between groups experimental data were longer, they might have preferred VAD.

Figures 5 and 6 show the results for when participants answered the question, "How did you feel about the placement of the start and end points when time labels were provided in advance?" Although there were some participants in both groups who corrected time labels that we provided, more than 82% of the users felt the labels were accurate. Due to the difficulty in deciding where to put the start point of speech, more than 91% of the participants preferred having the VAD label in advance of the speech, even if they had to correct it (Fig. 5).

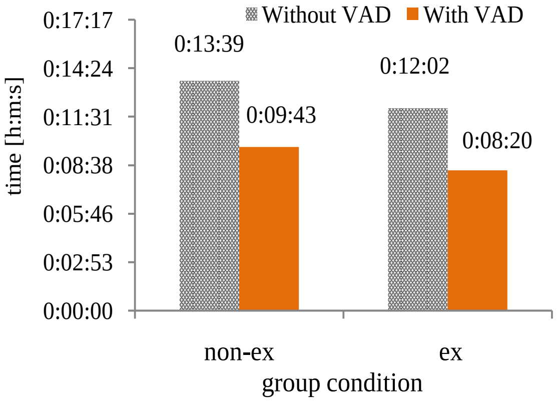


Fig. 3 Processing time between groups

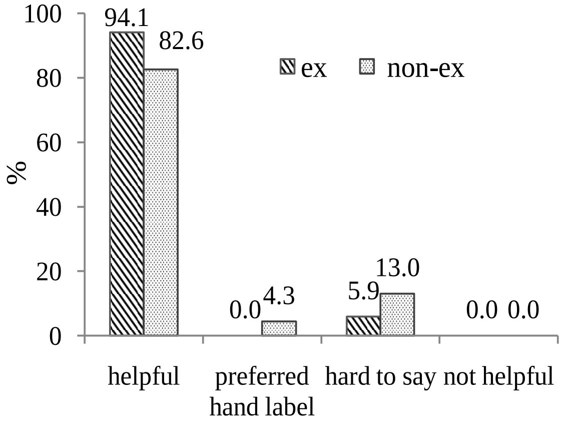


Fig. 4 Evaluation of the proposed method

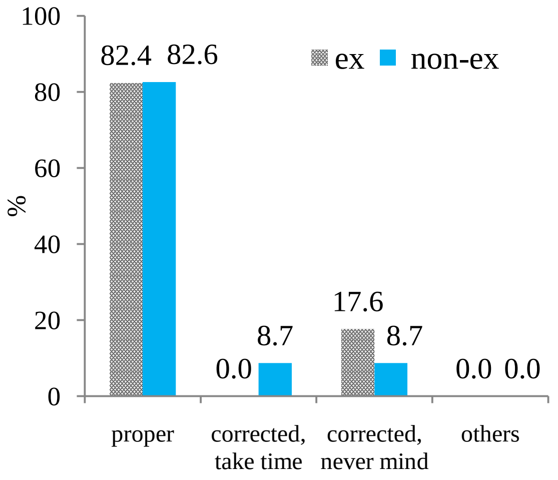


Fig. 5 Evaluation of starting points of VAD

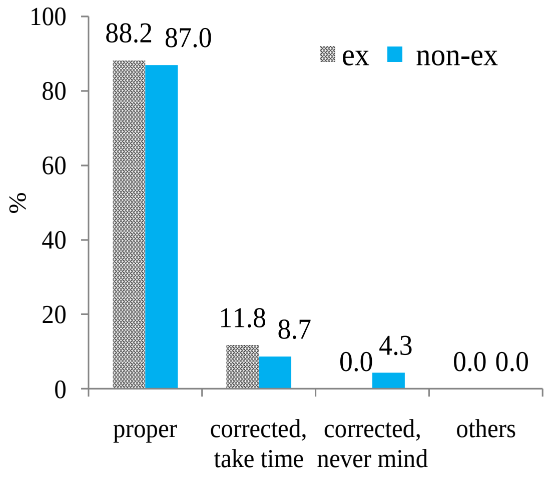


Fig. 6 Evaluation of end points of VAD

From these results, we can say that our method for making subtitles will be generally accepted by users, even though there were some participants who wanted to determine the start and end points by hand, and some participants could not decide which method they preferred. Some participants seemed annoyed when editing the errors in the time labels detected by our system. In the software used in this experiment, there was only a delete button, but no function to combine two pairs of time labels into one or to divide one speech portion into two. This may have made the participants feel some angst while putting captions into the predetermined time labels. Some modifications to the editing software may result in more users in support of our method.

4 Conclusions

This study introduced a VAD algorithm based on the modulation spectrum and its application for a subtitle-making system. The results of the experiment show a time savings over hand-labeling. Although there were a few participants who preferred hand-labeling to our method, the results showed a much greater percentage of participants who supported the VAD predetermined time labels for subtitles. However, the results of the present experiment do not provide strong evidence, because the experimental data were too short. We also need to improve the interface of the software for editing subtitles.

Acknowledgements

This study 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] Y. Fujikashi *et al.*, *Proc. Autumn Meet. Acoust. Soc. Jpn.*, 33-34, 2005.

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

[3] ETSI ES 202 050 v.1.1.4, 2006.

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

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

[6] M. Fujimoto *et al.*, *Proc. of ICASSP ' 08*, 4441-4444, 2008.

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

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

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

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

[11] http:// www.fujiyama1.com/