Voice Activity Detection Applied to Hands-Free Spoken Dialogue Robot based on Decoding using Acoustic and Language Model

Hiroyuki Sakai, Tobias Cincarek
Hiromichi Kawanami, Hiroshi Saruwatari, Kiyohiro Shikano
Graduate School of Information Science,
Nara Institute of Science and Technology
8916–5, Takayama-Cho, Ikoma-City, Nara, 630–0192, Japan
Email: hiroyuki-s@is.naist.jp

Akinobu Lee
Nagoya Institute of Technology, Japan
Email: ri@nitech.ac.jp

***Abstract*—Speech recognition and speech-based dialogue are means for realizing communication between humans and robots. In case of conventional system setup a headset or a directional microphone is used to collect speech with high signal-to-noise ratio (SNR). However, the user must wear a microphone or has to approach the system closely for interaction. Therefore it's preferable to develop a hands-free speech recognition system which enables the user to speak to the system from a distant point. To collect speech from distant speakers a microphone array is usually employed. However, the SNR will degrade in a real environment because of the presence of various kinds of background noise besides the user's utterance. This will most often decrease speech recognition performance and no reliable speech dialogue would be possible. Voice Activity Detection (VAD) is a method to detect the user utterance part in the input signal. If VAD fails, all following processing steps including speech recognition and dialogue will not work. Conventional VAD based on amplitude level and zero cross count is difficult to apply to hands-free speech recognition, because speech detection will most often fail due to low SNR.**

This paper proposes a VAD method based on the acoustic model (AM) for background noise and the speech recognition algorithm applied to hands-free speech recognition. There will always be non-speech segments at the beginning and end of each user utterance. The proposed VAD approach compares the likelihood of phoneme and silence segments in the top recognition hypotheses during decoding. We implemented the proposed method for the open-source speech recognition engine Julius. Experimental results for various SNRs conditions show that the proposed method attains a higher VAD accuracy and higher recognition rate than conventional VAD.

I. Introduction

Recently, automatic speech recognition (ASR) technology has been applied to real environment applications such as speech guidance system, robots, car navigation systems, portable speech translators, etc. Speech is the easiest, most natural and effective way for humans to communicate. Therefore speech should also be considered as an effective method for natural communication of humans with robots. In a conventional speech recognition system, a headset or a directional microphone is employed to collect speech with high SNR. However, the user must wear the microphone or has to approach the system closely for interaction. Therefore the introduction of a hands-free ASR system that can be used without these burdens is expected for robots, etc. We have been operating the speech oriented guidance system "Takemaru-kun"[1] (Fig.1) and "Kita-chan, Kita-robot"[2] (Fig.2) in real-environment for several years. Presently, each system uses a directional microphone. We are going to replace it with a microphone array to realize a hands-free interface.

In a hands-free automatic speech recognition system, SNR of the input signal will be worse than in a conventional systems due to the background noise from the real environment and that the user is not standing in front of the microphone. Low SNR is very likely to cause degradation in voice activity detection (VAD) and speech recognition performance when using a conventional VAD[3] approach. The purpose of VAD is to detect the user utterance part in the input signal. VAD is very important because if VAD fails, all following processing steps including speech recognition and dialogue will also fail. Amplitude level (AL) and zero cross count (ZC) are employed for VAD in a conventional speech recognition system. Speech recognition is only carried out when the amplitude level of the input signal exceeds a certain threshold. The lower the SNR, the more likely that this conventional approach to speech detection fails.

The employment of a "Push and Talk" interface would be a different approach to realize a hand-free ASR system which is also effective in noisy environments. However, the requirement for the user to push a button once before and once after talking is very inconvenient and would even require extra equipment. Therefore "Push and Talk" is no decent alternative to a real hands-free ASR system. Speech/non-speech discrimination using frame-base GMMs[4], optimization of discrimination using Adaboost[5] and the Speech Starter interface[6] which is based on the detection of lip movements are further

ROBOCOMM 2007, 14th – 16th Oct 2007, Athens, Greece.
Copyright © 2011 – 2012 ICST
ISBN 978-963-9799-08-0
DOI 10.4108/ICST.ROBOCOMM2007.2088

Fig. 1. *Real-Environment Speech Information Guidance System "Takemaru-kun" : System installed at the local community center since November 2002.*

Fig. 2. *Real-Environment Speech Train Information Guidance System "Kita-chan(back), Kitarobo(front)" : System installed at the local railway station since March 2006.*

examples of methods which can be employed for speech detection. However, when considering the employment in a real environment these methods are not enough to realize practical VAD considering both performance and real-time capabilities.

In this paper, we propose a VAD method for hands-free speech recognition which is based on the speech recognition algorithm. For decoding the spoken text from the input signal an acoustic model (AM) and a language model (LM) are employed. The acoustic model is employed to model the characteristics of each phoneme of the target language and non-speech segments such as the background noise. The language models determines which sentences should be recognized by the system. As the conventional approach, the proposed method also employs a statistical model (GMM) to reject unwanted speech inputs, e.g. laughter, coughing and many other kinds of noise. Therefore, robust VAD can be expected. The effectiveness of the proposed method is evaluated in this paper.

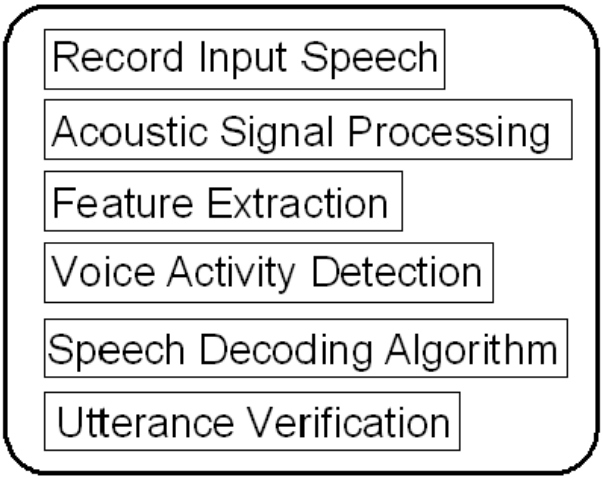

Fig. 3. *Structure of Standard Speech Recognition System*

II. Standard Speech Recognition System

The structure of a standard speech recognition and the processing steps of the system from speech input to utterance verification are shown in Figure 3. Input speech is collected with a microphone and digitized by the computer hardware. Additionally, acoustic signal processing can be applied to improve SNR, e.g. by using noise suppression and blind source separation (BSS)[7]. After that, acoustic feature extraction is carried out. As next step, VAD is employed to detect the user utterance part in the input signal. Non-speech input, e.g. laughing, coughing and other kinds of noise can be rejected after segmentation by VAD based on utterance classification using a statistical model, e.g. a Gaussian mixture model (GMM)[8]. In parallel to voice activity detection, the search for recognition hypotheses is carried out most often using a HMM-based acoustic model and a statistical n-gram language model. The acoustic model models the acoustic characteristics of phonemes and the background noise. The language model models the connection of words and the structure of sentences. Recognition works by calculating the similarity between the segments of the input signal and a dynamically generated graph of recognition hypotheses using both acoustic and language model. Finally, utterance verification, i.e. classifying the input either as speech or as noise, is conducted. Most often GMMs are employed to represent speech and many kinds of noise. In case of a speech-oriented dialogue system, the system also analyzes a user's intention using the recognition result and generates a system answer to respond to a user's request (dialogue management). In the following the conventional and proposed VAD method are explained in detail.

III. The Approach

A. Conventional Method

VAD is carried out as preprocessing before actual speech decoding. Although a few conventional VAD methods exist we restrict their treatment to one example of a real-time

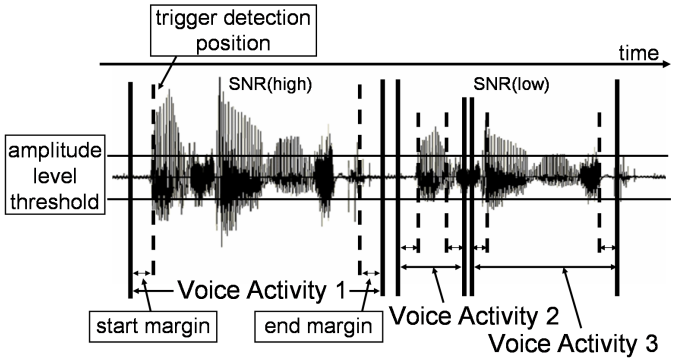

Fig. 4. *Summary of Conventional VAD Method*

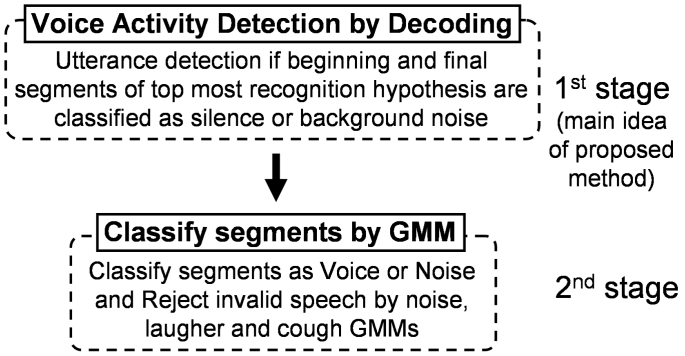

Fig. 5. *Block Diagram : Proposed Voice Activity Detection*

VAD method because this work considers the introduction of VAD into a real environment speech dialogue system. Consequently, VAD based on amplitude level (AL) and zero cross count (ZC) is considered here. The method is designed to make use of the difference of the amplitude level between the user utterance part and the background noise, i.e. the signal-to-noise ratio (SNR). If SNR is high, it is easy to set a threshold to separate the utterance from the background noise with high performance, because the difference between the utterance amplitude level and the background noise level is large. However, for ASR under hands-free condition, adjustment of the amplitude threshold is very difficult due to a low SNR. Therefore, performance of user utterance detection is likely to degrade. Fig. 4 gives an overview to the conventional VAD method. The waveform of the same utterance with two different SNRs is shown in Fig. 4 (left: high SNR, right: low SNR). This is an example where the low SNR utterance can be detected separately using the same threshold as for the high SNR utterance. In order to prevent cutting of the start and end of the utterance, start margin and end margin are employed before the beginning and after the end of the detection points.

B. Proposed Method

We propose a two-stage VAD for a hands-free ASR application. Fig. 5 gives an overview to the proposed method. Segments of the input containing only laughter,

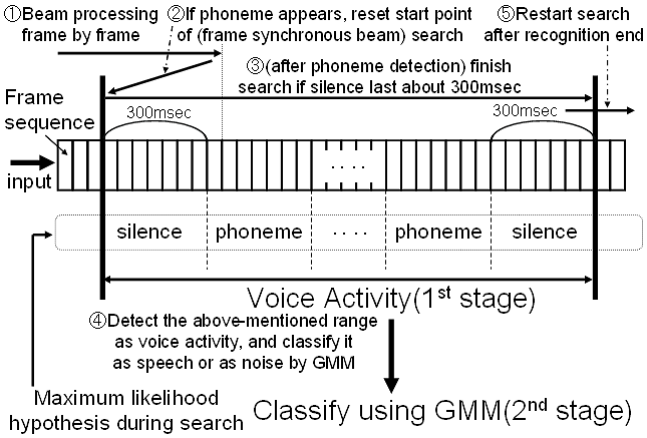

Fig. 6. *Flow of VAD by decoding in First stage*

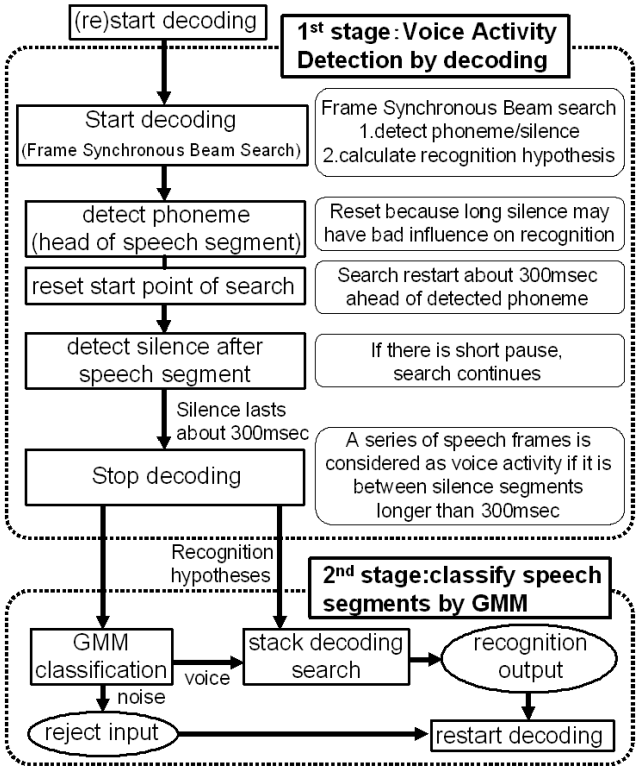

Fig. 7. *Concept of Recognition Flow in Proposed Method*

coughing, car horn honk, clapping, etc. are defined as "noise" segments. A noise segment has a temporarily high amplitude level but it is not part of the utterance. Segments containing user speech are defined as "voice" or "phoneme" segments. Otherwise, a segment is defined as "silence (non-utterance)", i.e. it contains only background noise but no user speech.

The following explains the two stages of the proposed VAD. Additionally, Fig. 6 shows the flow of processing of the proposed VAD which is carried out in parallel with the decoding algorithm. Fig.7 shows the main steps of VAD and the decoding algorithm.

Proposed VAD:

- First Stage

Voice Activity Detection and Preliminary Decoding. There will always be non-speech segments (silence) before and after user utterances. Frame Synchronous Beam Search (FSBS) is carried out using language model (LM) and acoustic model (AM). FSBS classifies segments as phonemes, silence or noise. A frame or a series of frames which are recognized as part of phonemes or series of phonemes are considered as "voice activity". Otherwise, a sequence of silence frames of about 300 to 400 msec in duration is considered as noise (cf. Fig. 6). There are two processing steps. The first is VAD based on decoding using AM and LM. The other is to search for one or more recognition hypotheses.

- Second Stage
  Voice activity segments from the first stage are reclassified either as voice or noise based on likelihood scores of GMMs for voice and noise. If the classification result is noise, the input is rejected and VAD is restarted. If the result is voice, the final recognition hypothesis is calculated. The GMM-based classification has the positive effect of preventing useless computation for noise segment which are rejected in any case.

There are two differences between the conventional method and the proposed method. Firstly, the proposed method is not considering amplitude level (AL) and zero cross count (ZC). Acoustic and linguistic models are employed instead. Secondly, VAD is carried out by decoding and is part of the actual speech recognition algorithm. This is different to the conventional approach where VAD is carried out independently from the recognition process. While in the conventional method the search for recognition hypotheses is only carried out for segments actually cut by initial VAD, the complete input stream is processed by the recognition algorithm in the proposed approach. Since background noise and other noise segments should be rejected, the start position for decoding has to be reset whenever a longer background noise segment is detected. After a reset, the search is restarted at the beginning of the background noise segment detected last. A fatal processing delay does not occur because the computational complexity of frame synchronous beam search (FSBS) is adjustable by changing the beamwidth, etc. Consequently, the proposed VAD method works in real-time.

The acoustic model (AM) employed for decoding has been adapted to the target environment background noise by MLLR[9] using speech data collected in the target environment. This improves detection performance over a conventional acoustic model because the environmental background noise can be recognized as silence effectively. Moreover GMM-based noise classification makes it possible to reject various kinds of noise. We integrated the proposed VAD method into the open-source speech recognition decoder Julius[10][11].

C. Overview of Julius

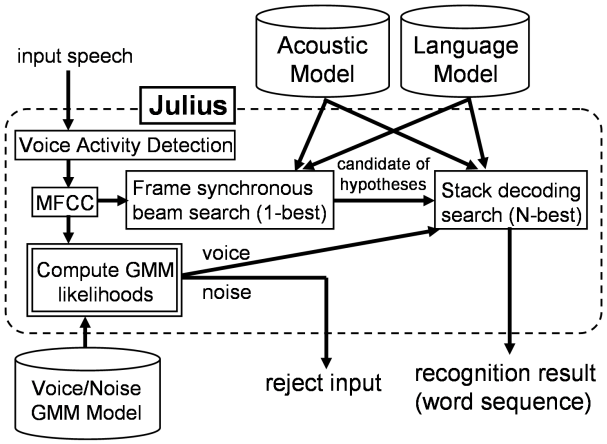

Fig. 8. *Outline of System Organization of Julius*

Julius is a high-performance, real-time, two-pass large vocabulary continuous speech recognition (LVCSR) engine for research and practical system development. Julius is open-source software, may be distributed freely and even included in commercial software. It works on various software platforms such as Windows, FreeBSD, Linux and other Unix derivatives. Julius can be employed for different languages by changing the acoustic (AM) and language models (LM). Therefore, it can be applied to a wide range of applications and is usable for arbitrary target languages. The system's organization and the flow of processing during recognition are shown in Fig. 8. Julius' conventional VAD method uses amplitude level (AL) and zero cross count (ZC). Mel-frequency cepstrum coefficients (MFCC) are extracted as acoustic features. The search algorithm for recognition hypotheses employs HMM-based acoustic model and a statistical n-gram language model. It is a two pass algorithm. In the first pass, frame synchronous beam search (FSBS) is carried out using a bi-gram language model to determine a preliminary recognition hypothesis. After the first pass is finished, GMM-based classification of the input into voice and noise is carried out. If the classification result is voice, the second pass is carried out with stack decoding using a tri-gram language model. If the classification result is noise, the second pass is skipped and the input rejected.

D. Julius with the Proposed Method

The system organization of Julius that integrated the proposed method is shown in Fig.9. MFCC features for each input speech frames is extracted. The proposed VAD method is carried out using synchronous beam search (FSBS) based on acoustic model (AM) and language model (LM). The search also determines a preliminary recognition hypothesis. Next, the voice activity segments are reclassified using the voice and noise GMMs. If the classification result is voice stack decoding using the tri-gram language model is carried out. If the classification result is noise, the input

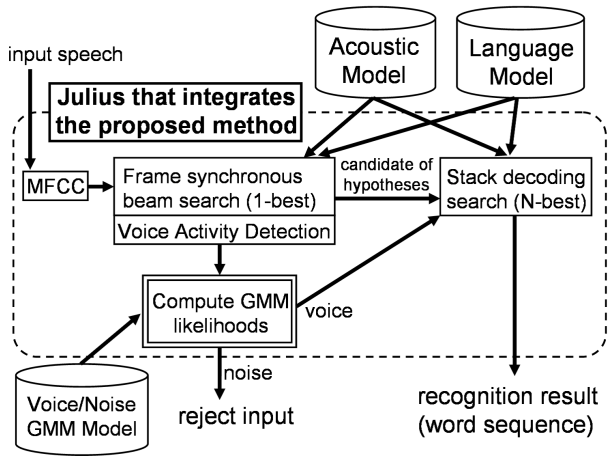

Fig. 9. *Julius that integrated Proposed Method*

is rejected, and frame synchronous beam search (FSBS) is restarted.

IV. Evaluation Experiments

In the following the experimental setup for evaluating the recognition performances of the proposed and conventional VAD method is explained. For the conventional method several thresholds for amplitude level (AL) and zero cross count (ZC) are considered.

A. Experimental Conditions

Julius as shown in Fig. 8 is employed for evaluating conventional VAD. Fig.9 shows Julius for evaluating the proposed VAD method. The Kita-chan system (cf. Fig.1) was employed to collect the evaluation speech data for three different SNR conditions which are controlled by changing distance between the speaker and the microphone. Kita-chan is a speech-oriented guidance system operated at a railway station near the author's university in Nara, Japan. The acoustic model (AM) was constructed using the data collected during two years by second speech-oriented guidance system, Takemaru-kun (cf. Fig.2). The number of utterances for training an initial acoustic models was 23,417 for adults and 120,671 for children. After that MLLR adaptation was carried out to adapt the AM to the environment of the railway station. The background noise level at the railway station was between 56 and 63 dB(A). The adaptation data was collected during six weeks of regular system operation in the Kita-chan environment. The amount of adaptation data is 6,661 utterances for adults and 9,472 utterances for children. Moreover, cepstrum mean normalization (CMN) was employed to reduce acoustic mismatch due to speaker characteristics. Further experimental conditions are given in Table I.

B. Feature of input data

Examples of input speech waveforms for conditions 1 to 3 are shown in Figs. 10-12, respectively. In Condition 1, the SNR is high with about 50 dB. The difference between

TABLE I
Experimental Condition

Input Speech (2&3 are Hands-Free)	Condition 1	Close talk
	Condition 2	Collected from about 1m distance
	Condition 3	Collected from about 1.5m distance
Threshold in Convention	amplitude	100〜1000
	zero cross	0〜100 (times/sec) (default=60)
Acoustic Model(AM)		2000 states, PTM, Gaussian
Acoustic Features		12MFCC, 12ΔMFCC, ΔE
AM Training		Baum-Welch, 3 Iterations
AM Adaptation		MLLR-MAP, 3 Iterations, 256 Classes
Language Model(LM)		3-gram, Kneser-Ney smoothing
		Vocabulary size is 40k.
Task	guidance of railway station, train information, sightseeing, institutions, local area information, news, weather forecast, greetings and conversation	
Evaluation Data	1 speaker, 204 Utterances, 1024 words OOV = 0% (no unknown word).	

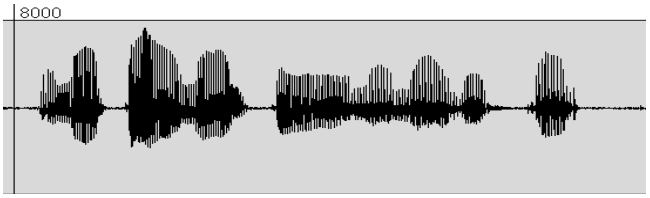

Fig. 10. *Input Wave in Condition 1 (SNR = 50dB, close talk)*

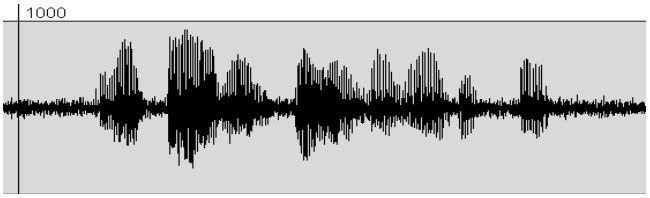

Fig. 11. *Input Wave in Condition 2 (SNR = 10dB, 1m distance)*

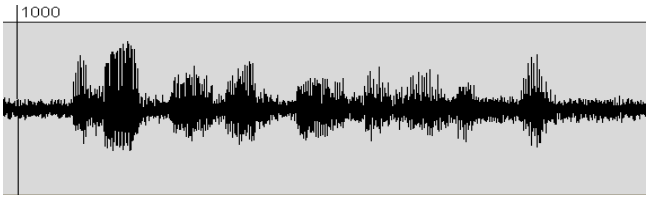

Fig. 12. *Input Wave in Condition 3 (SNR = 6dB, 1.5m distance)*

the amplitude of the utterance and the background noise is large. In Condition 2, SNR is about 10dB and the level difference between the utterance and the background noise is quite small. The part which corresponds to the start and the end of the utterance has almost been buried in the background noise. In Condition 3, SNR is only about 6dB which is worse than Condition 2.

V. Experimental Result

Figures 13 to 15 show word accuracy in Conditions 1-3. The results of the proposed method are shown as planes for various thresholds of the amplitude level (AL) and zero cross count (ZC). The effectiveness of the proposed method is clear from the experiment result. In hands-free condition (2 and 3), the recognition performance of

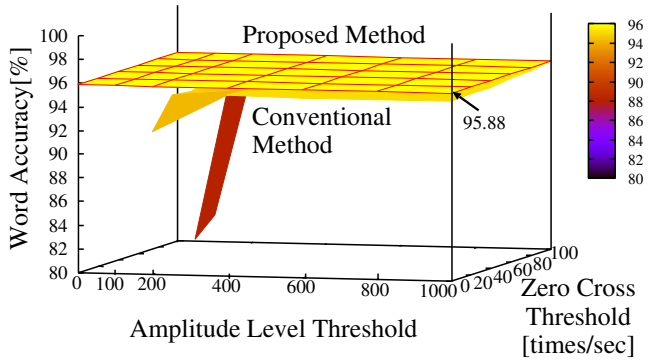

Fig. 13. *Result of Condition 1 (50dB SNR, close talk)*

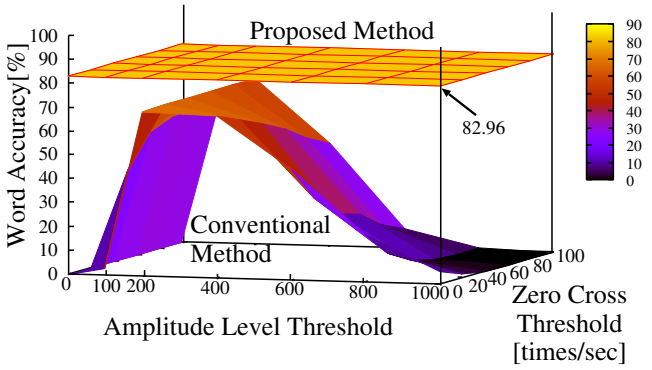

Fig. 14. *Result of Condition 2 (10dB SNR, 1m distance)*

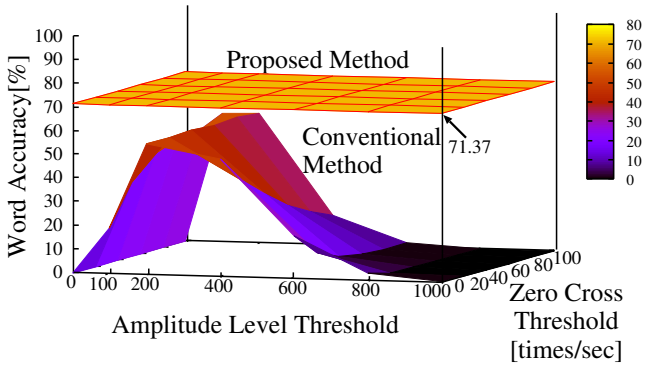

Fig. 15. *Result of Condition 3 (6dB SNR, 1.5m distance)*

the conventional VAD method depends greatly on the AL threshold. Moreover, the performance is lower than the proposed method. The dependency on ZC is small. It has only a small influence on the recognition performance. The results show that the performance of the proposed method degrades to some extent in hands-free condition, but it is much higher than when using the conventional method. Moreover, the experimental results show that the proposed method is also better than the conventional method even in case of condition 1. The results for ASR performance are given by the word correct rate (cf. Figs. 16 and 17), since it more related related to the system's response performance than word accuracy and the investigation in this paper considers introduction of hands-free speech recognition to a speech dialogue systems. Fig. 16 shows ASR performance

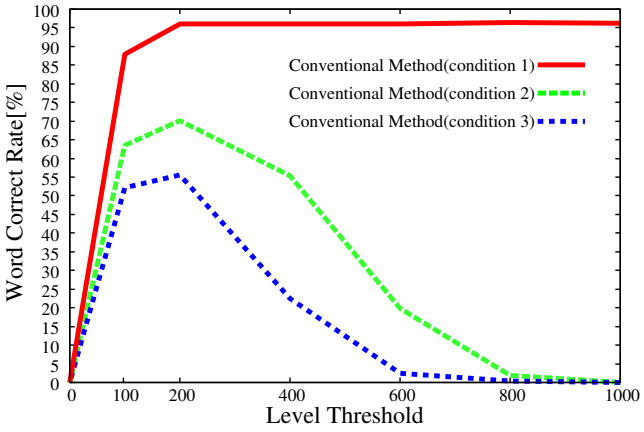

Fig. 16. *Recognition rate of conventional method (ZC threshold 60)*

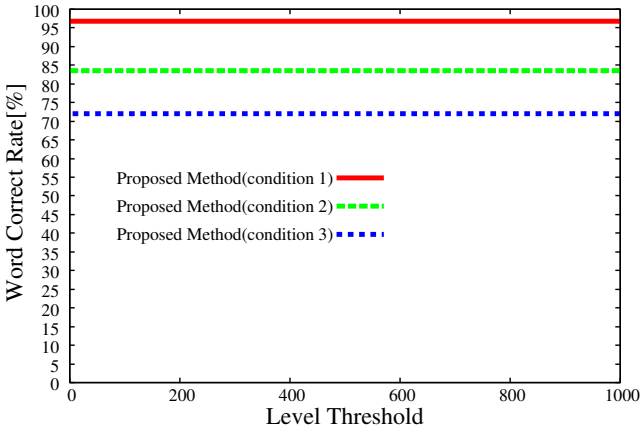

Fig. 17. *Recognition rate of proposed method (ZC threshold 60)*

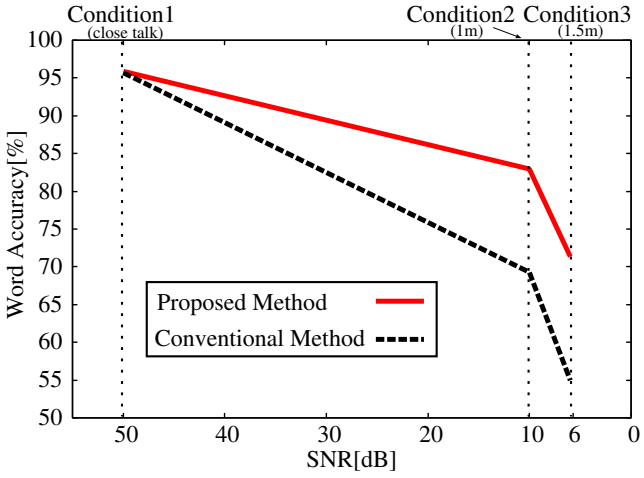

Fig. 18. *Dependency of Recognition rate on SNR*

of the the conventional method. Fig. 17 shows results for the proposed method. Furthermore, the change of speech recognition rate depending on conditions 1-3 is shown in Figs. 18. It is clear that the proposed method improves the performance in hands-free condition remarkably.

Finally, it is confirmed whether VAD itself is effective. As indicators to evaluate VAD performance, the number of "False Rejections" and the number of "False Acceptions"

are defined as follows:

- False Rejection
  Utterance detection fails, i.e. a segment which is actually voice is rejected as noise or background noise. For such segments speech recognition is not carried out in the end. For the conventional VAD method, detection of the user utterance fails if AL threshold is higher than the utterance AL.
- False Acceptance
  A segment which is not part of the user utterance is detected as part of a user utterance, e.g. a background noise segment is recognized as part of a user's utterance. In the conventional method, background noise is very likely to be classified as part of the utterance if the AL threshold is lower than the background noise.

Figure.19 shows the number of false rejections, Fig. 20 shows the number of false acceptions. Fig. 19 and Fig. 20 both show the typical recognition result for conditions 1 and 3. For ZC a threshold of 60 is selected because the result is not very dependent on the ZC threshold. In the proposed VAD method, the number of errors occurring is constant, and performance improves in comparison to the conventional method. Some errors may occur without affecting SNR because the proposed method processes the complete input stream and many kinds of noises exist in a real environment. The conventional method needs an appropriate threshold setting. However, it is difficult to realize that for low SNR.

VI. Conclusion

In this paper, we proposed a method for VAD using acoustic model and language model by speech decoding for real-environment hands-free speech recognition. Moreover, we integrated the proposed method into the open-source speech recognition decoder Julius. We evaluated the proposed method and compared its performance to the conventional VAD method which is based on amplitude level and zero cross count. Experimental results show that for high SNR, both the conventional and the proposed method work well. However, when SNR is low due to a hands-free setup the conventional VAD causes more misdetections. The proposed VAD method outperformed the conventional method w.r.t to both VAD and ASR performance.

The proposed method requires to adapt the acoustic model to the target environment in order to recognize the background noise as silence effectively. Adaptation is possible with a small amount of data collected in the target environment. Consequently, the proposed method can be easily applied for various kinds of real-environment ASR applications. Apart from that, the system will always have a high computational load with the proposed method since it has to process the complete input stream. In future work, we are going to employ face detection to detect whether a user is present to control termination and restart of the recognition engine. This will decrease the computational load of the system and may further improve ASR performance.

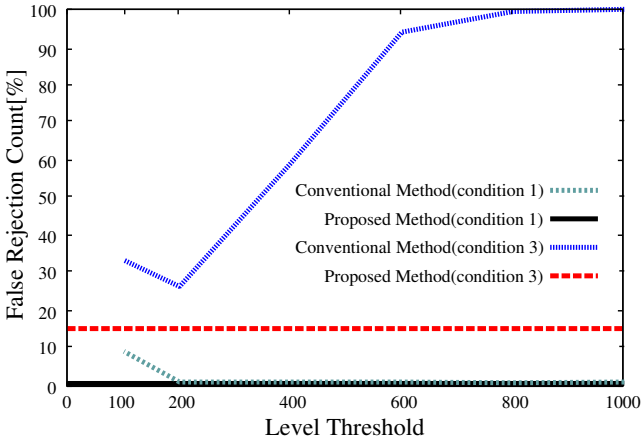

Fig. 19. *False Rejection Count*

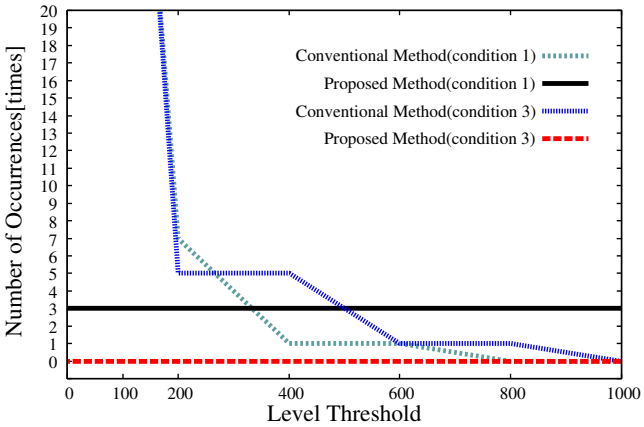

Fig. 20. *Number of False Acceptance*

Moreover, the speaker's position can be obtained from image processing and employed as initial value for direction of arrival estimation which is necessary for many signal processing techniques. On the other hand, there is also new potential by recognizing the complete input stream. For example, sound of running water, blowing of the wind, blaze and whistling kettle, etc. could be detected in order to asses the environment in which the system is currently used.

Acknowledgment

This work is partly supported by MEXT e-society leading project.

References

[1] R. Nisimura, A. Lee, H. Saruwatari, and K. Shikano, "Public Speech-oriented Guidance System with Adult and Child Discrimination Capability", proc.of ICASSP, pp.433–436, 2004.

[2] H. Kawanami, M. Kida, N. Hayakawa, T. Cincarek, T. Kitamura, T. Kato and K. Shikano, "Spoken Guidance Systems Kita-chan and Kita-chan robot.Their Development and Operation in a Railway Station" , tech.rep., IEICE, SP2006–14, 2006.

[3] L. R. Rabiner, M. R. Sambur: "An Algorithm for Determining the Endpoints of Isolated Utterances", BSTJ, vol.54, No.2, pp.297–315, 1975.

[4] Norbert Binder, Konstantin Markov, Rainer Gruhn, Satoshi Nakamura, "Speech/Non-Speech Separation with GMMs" , Proc.of ASJ Fall Meeting, Vol1, pp.141–142, 2001

[5] Oh-Wook Kwon, Te-Won Lee, “Optimizing speech/non-speech classifier design using AdaBoost”, Acoustic, Speech, and Signal Processing, 2003.Proceedings.(ICASSP;03).2003 IEEE International Conference on Volume 1, Issue, 6–10 April 2003 Page(s):l436-l.439 vol.1

[6] K. Kitayama, M. Goto, K. Itou, T. Kobayashi, “Speech Starter : “SWITCH” on Speech”, IPSJ SIG Technical Report, 2003–SLP–46–12, Vol.2003, No.58, pp.67–72, May 2003.

[7] S. Haykin, Ed., “Unsupervised Adaptive Filtering (Volume I, Blind Source Separation)”, John Wiley & Sons, 2000.

[8] G. McLachlan, “Finite Mixture Models”, Wiley-Interscience, 2000.

[9] M. Yamada, A. Baba, S. Yoshizawa, Y. Mera, A. Lee, H. Saruwatari, K. Shikano, “Unsupervised Acoustic Model Adaptation Algorithm Using MLLR in Noisy Environment”

[10] Open-Source Large Vocabulary CSR Engine Julius developed by A.LEE
introductory web pages at: `http://julius.sourceforge.jp/`

[11] A. Lee, T. Kawahara and K. Shikano. “Julius - an open source real-time large vocabulary recognition engine”, Proc Eurospeech2001, pp1691–1694, 2001