INTERSPEECH 2009 BRIGHTON

# Robust Speech Recognition Using VAD-measure-embedded Decoder

*Tasuku Oonishi*[1], *Paul R. Dixon*[1], *Koji Iwano*[2], *Sadaoki Furui*[1]

[1]Department of Computer Science, Tokyo Institute of Technology
2-12-1, Ookayama, Meguro-ku, Tokyo, Japan
[2]Faculty of Environmental and Information Studies, Tokyo City University
3-3-1 Ushikubo-nishi, Tsuzuki-ku, Yokohama, Japan
{oonishi,dixonp}@furui.cs.titech.ac.jp,iwano@tcu.ac.jp,furui@cs.titech.ac.jp

## Abstract

In a speech recognition system a Voice Activity Detector (VAD) is a crucial component for not only maintaining accuracy but also for reducing computational consumption. Front-end approaches which drop non-speech frames typically attempt to detect speech frames by utilizing speech/non-speech classification information such as the zero crossing rate or statistical models. These approaches discard the speech/non-speech classification information after voice detection. This paper proposes an approach that uses the speech/non-speech information to adjust the score of the recognition hypotheses. Experimental results show that our approach can improve the accuracy significantly and reduce computational consumption by combining the front-end method.

**Index Terms**: speech recognition, voice activity detection, decoder

## 1. Introduction

When a speech recognition system is deployed into a practical environment such as controlling a car navigation system or an automatic telephone attendant, the recognizer will frequently have to process input signals which contain long non-speech regions and environmental noises. It is essential to be able to robustly detect and remove these non-speech regions, because the introduction of the non-speech harms the recognition accuracy and background noise often leads to an increase in insertion errors.

A Voice Activity Detector (VAD) is a device which will attempt to detect the speech and non-speech region in the given input signal. Robust VAD techniques have been extensively utilized because they are not only important in a speech recognition system, but also a very fundamental part of most modern speech communication systems. Often the non-speech regions do not contain useful information, therefore removing the non-speech regions allows for a reduction in the storage, transmission and computational requirements.

Several extremely popular methods for VAD are front-end based approaches which exploit characteristics of the input signal such as the energy and Zero Crossing Rate (ZCR) [1]. Other front-end based detector uses models such as Gaussian Mixture Models (GMMs) to model the speech and non-speech statistics and performs a likelihood ratio test [2].

The other major approach [3] utilizes recognition results to segment input signal. After segmentation GMM based VAD method is utilized to pick up a speech regions and decoder recognizes again by using only the speech regions.

These approaches typically take intervals from the input signal and perform detection by calculating a Speech/Non-Speech (SNS) score. If the score exceeds some pre-determined threshold, then the input interval is judged as speech and is passed along the recognition pipeline. If the score is below the threshold, the input interval is marked as non-speech and dropped. This approach to VAD is a hard decision and the score assigned to speech interval is not used during decoding.

When the input signal contains a speech region whose SNS score is high the system should correctly recognize this input as a word. Likewise a silence region whose SNS score is low should be correctly recognized as a silence. However, in noisy conditions silence frames are often incorrectly labeled as speech frames and passed to the decoder, the SNS scores are discarded and the decoder will not have the ability to recover, therefore the noise frames are often incorrectly recognized as words and lead to insertion errors.

In this paper we introduce a novel approach to VAD that makes use of SNS score on a frame by frame basis to bias the hypothesis scores in the decoding phase. The basic idea is to add a confidence measure of non-speech to a frame acoustic score if the state belongs to a silence or short-pause model. Alternatively, a confidence measure of speech is added if the state belongs to a phonetic model. By using the proposed method the SNS score is used as continuous value rather than a hard binary decision and the decoder uses the value to perform better discriminations of words and silences.

The rest of the paper is structured as follows; The next section describes in more detail two common approaches to VAD which will serve as the baselines in our experiments. In section 3 we introduce our technique. In section 4 we demonstrate the effectiveness of the proposed algorithm on an in-car speech recognition task. The paper finishes with conclusions and future work.

## 2. Front-end based VAD

In this paper we use two front-end based approaches as baseline comparisons. The first is a VAD that exploits the ZCR as described in [1]. The second baseline is a GMM based likelihood ratio scheme as detailed in [2].

### 2.1. Zero crossing rate based VAD

In conditions which have a high Signal to Noise Ratio (SNR) a VAD based on the energy level of the signal can operate at satisfactory accuracy [4]. However, the detector will often incorrectly label on low-power speech signals and as the SNR falls the performance can suffer quite significantly [4]. The ZCR based approach exploits another characteristic of speech

signals. Speech signals will change phase frequently and this fact can be combined with the energy level to more accurately label an input signal frame by frame using the following:

1. Does the power exceeds pre-determined threshold: $TH_{power}$
2. The sign of adjacent sample is opposite

If the number of samples which satisfies the above both two conditions exceeds the pre-determined threshold:$TH_{ZCR}$, we judge the interval as speech, if not we judge the time interval as non-speech. The main advantages of this method are simple to implement and can yield good results.

### 2.2. GMM based VAD

Another common front-end based approach to VAD is to use GMMs to learn the acoustic characteristics of the input signal. Typically this involves using labeled data to train separate GMMs to model the speech and non-speech data. Then given the $i^{th}$ observation vector $X^i$, the frame can be labeled as speech if the likelihood ratio below exceeds pre-determined threshold ($TH_{GMM}$) and the frame can be labeled as non-speech if the likelihood ratio does not exceed the threshold.

$$L_{GMM} = \log \frac{p(X^i|H_1)}{p(X^i|H_0)} \tag{1}$$

Here, $H_1$ is hypothesis of speech and $H_0$ is hypothesis of non-speech. This method has been demonstrated to detect speech robustly even in the case of low SNR.

## 3. Proposed method

In this paper we propose a VAD scheme that utilizes an SNS score to adjust the score of the recognition hypothesis. In this section we describe details of our approach.

### 3.1. Acoustic likelihoods in the proposed method

The first step is to compute for each frame the SNS likelihoods $p(X^i|H_1)$ and $p(X^i|H_0)$ using the speech and non-speech models. In this work we used the GMM likelihoods for $p(X^i|H_1)$ and $p(X^i|H_0)$. For each frame $i$ these likelihoods are then used to calculate confidence measures according to:

$$C^i_{H_1} = \frac{p(X^i|H_1)}{p(X^i|H_1) + p(X^i|H_0)} \tag{2}$$

$$C^i_{H_0} = \frac{p(X^i|H_0)}{p(X^i|H_1) + p(X^i|H_0)} \tag{3}$$

Here, $C^i_{H_1}$ is the confidence measure of the $i^{th}$ frame contains speech, and $C^i_{H_0}$ is the confidence measure of the frame does not contain speech. The denominator terms ensure the confidence measures exist in the range 0 to 1.

We then use the confidence measures to bias the acoustic model scores.

If the hypothesis belongs to a phone model, the acoustic model score is biased using:

$$\log\ \hat{p}_{am}(X^i|\theta) = \log\ p_{am}(X^i|\theta) + \alpha\ \log\ \bar{C}^i_{H_1} \tag{4}$$

$$\bar{C}^i_{H_1} = \frac{\sum_{i-n}^{i+n} C^i_{H_1}}{2n+1} \tag{5}$$

Otherwise the hypothesis must belong to a silence model and the acoustic model score is biased using:

$$\log\ \hat{p}_{am}(X^i|\theta) = \log\ p_{am}(X^i|\theta) + \alpha\ \log\ \bar{C}^i_{H_0} \tag{6}$$

$$\bar{C}^i_{H_0} = \frac{\sum_{i-n}^{i+n} C^i_{H_0}}{2n+1} \tag{7}$$

Here, $X^i$ is the $i^{th}$ feature vector, $\theta$ is hypothesis, $p_{am}(X^i|\theta)$ is the acoustic model score, $\alpha$ is a scaling factor and $n$ is a smoothing parameter for computing $C^i_{H_1}$ and $C^i_{H_0}$ over a window. The smoothing parameter $n$ means that the proposed method yields latency of $n$ frames.

If the parameter $\alpha$ is set to 0, $\log\ \hat{p}_{am}(X^i|\theta)$ and $\log\ p_{am}(X^i|\theta)$ become equal. In the interval whose $C^i_{H_1}$ (or $C^i_{H_0}$) becomes 1, the acoustic model scores in word and silence models become $\log\ p_{am}(X^i|\theta)$ and $-\infty$(or $-\infty$ and $\log\ p_{am}(X^i|\theta)$). This means that only a word (or silence) is recognized in this interval. In the speech interval a word is correct and in the non-speech interval a silence is correct. Therefore this method does not affect the score of the best hypothesis if the $C^i_{H_1}$ and $C^i_{H_0}$ have high confidence.

### 3.2. Combination with front-end VAD

As previously discussed one of the merits of utilizing a front-end VAD is that non-speech frames can be discarded early in the recognition pipeline, which means computational cost is saved because not all input needs to be recognized. To reduce the computational burden we also employ frame discarding and only reject frames in which we have a high confidence of non-speech. The combination of our scoring technique with frame discarding is a powerful approach allowing we can achieve higher recognition accuracy and simultaneously reducing the computational costs.

## 4. Experiments

We evaluated our approach using the Drivers Japanese Speech Corpus in a Car Environment (DJSC) corpus [5]. This is a hands-free command and control task composed of utterances recorded in a car driving on a motorway. The test set consists of 40 speakers equally split between male and female speakers. Each participant provided 41 commands in an utterance continuously that would operate navigation whilst driving. The commands within each utterance are separated by one to two seconds non-speech regions which capture the background noise conditions. The recordings were performed at 16 kHz using a microphone mounted in the position of the navigation device. The acoustic models were trained on 52 hours of speech data from the Japanese Newspaper Article Sentences (JNAS [6]) corpus. The training material is gender balanced containing 130 male speakers giving 25 hours of speech and 130 female speakers providing another 27 hours of speech. From the processed data the acoustic models were EM trained and this process yielded a set of three states left-to-right tri-phone HMM with 2000 states. Each state output density was a 16 component GMM with diagonal covariance. In the evaluation the training and testing data were processed as follows. The raw speech waveforms were converted to a sequence of 38 dimensional feature vectors with 10 ms frame rate and 25 ms windows size. Each feature vector was composed of 12 Mel-frequency cepstral coefficients (MFCCs) with deltas and delta-deltas, augmented with log delta and delta-delta energy terms. The lan-

guage model was a network grammar and the vocabulary size was 83 words to cover all of the commands. The network had a path which corresponded to each of the valid commands that looped through the initial state to allow continuous recognition of the utterance stream.

The GMMs for the VAD each had four Gaussian components. The speech GMM was trained using the data from 967 lectures of the Corpus of Spontaneous Japanese (CSJ) [7] and the non-speech GMM was trained with data from car noise from Japan Electronic Industry Development Association (JEIDA). In the recognition evaluations we used $\mathrm{T}^3$ Decoder [8] currently under development at Tokyo Institute of Technology.

### 4.1. Word accuracy results

We first show in Figure 1 the recognition accuracy when using the different VAD methods.

- *baseline* represents the result without any VAD.
- *ZCR* is the result when using ZCR and energy for VAD (front-end VAD).
- *GMM* is the result using the GMM based likelihood ratio detector (front-end VAD).
- *proposed* is the result of our proposed VAD.
- *manual* corresponds to the result when using the corpus labels to remove non-speech region.

The scores $C^i_{H_1}$ and $C^i_{H_0}$ were also calculated using the GMMs which were used with GMM based likelihood VAD. The parameters of each of the VADs were optimized manually. (In ZCR VAD threshold of power $TH_{power}$ was set to 0, threshold of zero crossing rate $TH_{ZCR}$ was set to 10 and the frame length was 25ms. In GMM VAD threshold of likelihood ratio $L_{GMM}$ was set to -5. In the proposed method the scaling factor $\alpha$ was set to 10 and smoothing parameter $n$ was set to 15.) The results show that without any VAD the word recognition accuracy was 43.1% (deletion:23.7%, insertion:5.0%, substitution:28.2%). The ZCR method achieved 46.5% (deletion:24.6%, insertion:4.3%, substitution:24.6%) and GMM method achieved 45.8% (deletion:27.6%, insertion:3.7%, substitution:22.9%) word accuracy. These corresponded to an absolute word improvement of 3.4% and 2.7% respectively. The proposed technique achieved a 53.1% (deletion:25.6%, insertion:1.8%, substitution:19.5%) word accuracy and this corresponded to an absolute 10% improvement over the non-VAD baseline, this was the highest word accuracy we obtained in the evaluations and this result shows the effectiveness of our method. Using the labels from the corpus a 60.4% (deletion:19.3%, insertion:2.1%, substitution:18.2%) word accuracy was achieved. This indicates that there is still room to improve the proposed method.

### 4.2. Robustness results

Next we illustrate the performance of the proposed method for various parameter settings. Figure 2 shows the relationship between word accuracy and the scaling factor $\alpha$. The $X$ of $n$ $X$ in this figure expresses the smoothing parameter, for example $n$ 12 means that $C^i_{H_1}$ and $C^i_{H_0}$ are smoothed over 12 frames on either side of the current frame $X^i$ (25 frames in total). This figure shows that not only the accuracy is higher as the smoothing factor $n$ is increased, the selection of $\alpha$ parameter also becomes more robust. For these evaluations we achieved the highest word accuracy at a smoothing frame of 15 and an $\alpha$ value of 10. This corresponds to the best case recognition latency of 150 ms.

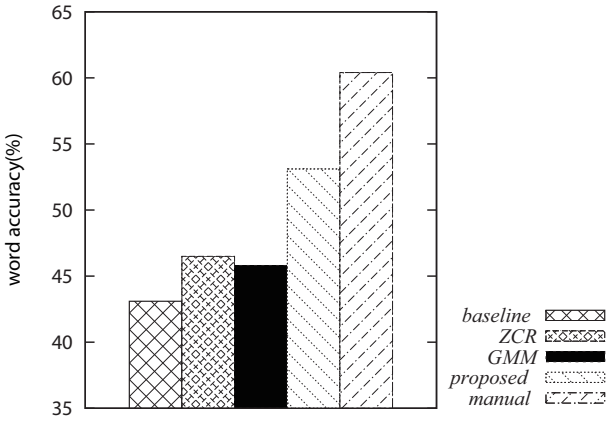


Figure 1: *The effect of VAD.*

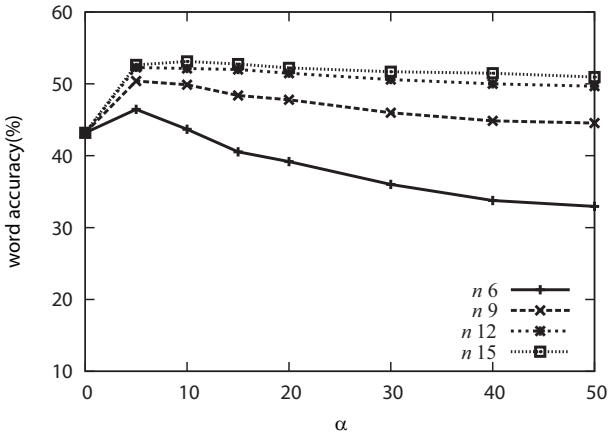


Figure 2: *The relationship between word accuracy and parameters in the proposed method.*

### 4.3. Combination results

In this section we demonstrate accuracy of the decoder when using a combination of the GMM based front-end VAD with the proposed method. Figure 3 shows the word recognition accuracy as the threshold value in the front-end GMM based detector is varied. The horizontal axis is $TH_{GMM}$ for the GMM based detector and the vertical axis is word accuracy. The threshold was varied from -15 to 5 with the figure showing the best improvement occurring for the value of -5. This is because the front-end rejects the frame with low scores and this rejection reduced recognition errors from the non-speech interval.

After the threshold is increased past 0 there is a rapid reduction in accuracy. This is because blocks of speech are rejected and this manifests as an increase in deletion errors.

### 4.4. Computational cost

In this section we show how the frame dropping front-end VAD can lead to a reduction in computational cost. The horizontal axis in Figure 4 is $TH_{GMM}$ for the GMM based detector

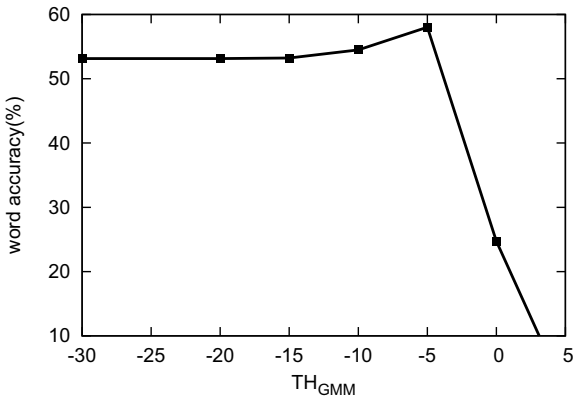


Figure 3: *The effect of combination of proposed and front-end VAD.*

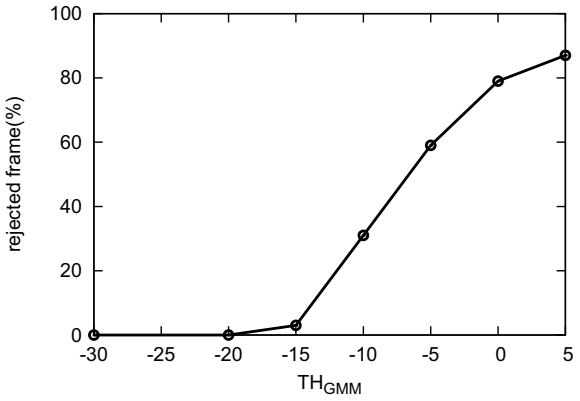


Figure 4: *The reduction of computation cost.*

and the vertical axis is the amount of frames dropped by the front-end VAD. The threshold value of -5 achieved the highest word accuracy and for this setting the front-end rejected 60% of the frames. The front-end VAD rejects a lot of frames past the threshold value of -5 and rejection rate beyonds 70% which is ratio of speech and silence in the test sound data. In this case the word accuracy degrades rapidly because a lot of speech frames are rejected. Therefore it is important to reject frames safely in front-end if the SNS score is low. Otherwise the frames should be passed to the proposed VAD method. By combining techniques in this manner we can reduce the computation cost by 60% without any degradation of accuracy.

## 5. Conclusion

In this paper we have described an approach to VAD method that utilizes the speech/non-speech score to adjust recognition hypotheses. The experimental results show that by using the method we can achieve a large improvement in word recognition under real world noisy conditions. We have also shown that the proposed method in combination with a front-end VAD can maintain the best accuracy whilst reducing computational cost by up to 60%.

In future work we are going to perform further evaluations on the robustness of the proposed method in various noise and SNR conditions and investigate ways to adjust parameters automatically on-line. Although the combination method has the ability to perform very well for certain parameter settings, future work is needed to ascertain if this is a general result and if on-line parameter adjustments are necessary to fully harness the method in changing environmental conditions.

## 6. Acknowledgements

This work was supported by the METI Project “Development of Fundamental Speech Recognition Technology”.

## 7. References

[1] A. Benyassine, E. Shlomot, H.-Y. Su, D. Massaloux, C. Lamblin and J.-P Petit, “ITU-T Recommendation G.729 Annex B: a silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications,” IEEE Communications Magazine 35 (9): pp.64-73, 1997.

[2] R. Singh, M. Seltzer, B. Raj and M. Stern, “Speech in Noisy Environments: robust automatic segmentation, feature extraction, and hypothesis combination,” Proc. ICASSP, vol 1, pp.273-276, 2001.

[3] H. Sakai, T. Cincarek, H. Kawanami, H. Saruwatari, K. Shikano and A. Lee, “Evaluation of hands-free speech recognition algorithm using decoding speech activity detection based on acoustic and language models,” IEICE Technical Reports, vol.107, no.107, pp.13-18, 2007.

[4] J. Ramírez, J. M. Górriz and J. C. Segura, “Voice Activity Detection: Fundamentals and Speech Recognition System Robustness,” Robust Speech Recognition and Understanding, I-Tech Education and Publishing, pp.1-22, 2007.

[5] K. Hiraki, T. Shinozaki, K. Iwano, A. Betkowska, K. Shinoda and S. Furui, “Initial evaluation of the driver’s Japanese speech corpus in a car environment,” IEICE Technical Reports, Asian Workshop on speech science and Technology, SP-2007-202, pp.93-98, 2008.

[6] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano and S. Itahashi, “JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research,” J.Acoust. Soc. Jpn.(E), vol.20, no.3, pp.199-206, 1999.

[7] K. Maekawa, “Corpus of spontaneous japanese: Its design and evaluation,” Proc. ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition, pp.7-12, 2003.

[8] P.R. Dixon, D.A. Caseiro, T. Oonishi and S. Furui, “The TITECH large vocabulary WFST speech recognition system,” Proc. IEEE Workshop on ASRU, pp.443-448, 2007.