interspeech 2008 incorporating SST 08

Phone-duration-dependent Long-term Dynamic Features for a Stochastic Model-based Voice Activity Detection

Takashi Fukuda, Osamu Ichikawa, and Masafumi Nishimura

Tokyo Research Laboratory, IBM Japan, Ltd.

{fukuda1, ichikaw, nisimura}@jp.ibm.com

Abstract

Accurate voice activity detection (VAD) is important for robust automatic speech recognition (ASR) systems. This paper proposes noise-robust VAD using long-term temporal information in speech. Long-term temporal information has been an ASR focus recently, but has not been investigated sufficiently for VAD. This paper describes an attempt to incorporate long-term temporal information into a feature parameter set by extracting conventional dynamic features from long-term cepstrum sequences. The proposed method includes the temporal contexts of phonemes by using long-term features and allows distinguishing between speech and non-speech intervals. The long-term features calculated over the average phoneme duration provide noise robustness. In an experiment on the Japanese digit corpus, the proposed method led to considerable improvements over conventional methods including the G.729 Annex B and the ETSI AFE-VAD under low SNR conditions, and had 71.1% error reduction on average as compared to the ETSI AFE-VAD.

Index Terms: voice activity detection, dynamic feature, long-term temporal information, average phoneme duration

10.21437/Interspeech.2008-311

1. Introduction

It is crucial for automatic speech recognition (ASR) systems to accurately detect speech present and speech absent segments in an utterance. Such a technology is called voice activity detection (VAD). Many approaches have been investigated over the years for creating a robust voice activity detector. In a standard ASR system, basically only the speech present segments found by the VAD system are passed to the ASR system and the ASR process are enabled for those segments, and the speech absent intervals are discarded. Therefore if the VAD system fails to detect a speech present segment due to noise, etc., the ASR system cannot respond correctly to the speaker's command. On the other hand, if the VAD system misdetects noise as a speaker's command, the ASR system can malfunction. Therefore there is a need for a robust VAD that works even in adverse conditions.

Energy-based criteria and zero crossing information have been widely used in conventional VAD systems, but they are not reliable against ambient noises [1]. In recent years, VAD systems based on statistical models were proposed and shown to work well in noisy environments [2, 3]. A Gaussian mixture model (GMM) is often used as the statistical model for VAD. These methods achieve noise robustness by designing the GMMs using data recorded in the same kind of situation where the VAD system will be deployed. This is a very simple and reliable method to configure a robust VAD system, but the performance degrades sharply under low signal to noise ratio (SNR) conditions. To overcome this problem, feature-based approaches have been investigated [4, 5, 6]. Li et al. proposed a method to leverage the higher order statistics of speech and non-speech signals [5]. Yamamoto et al. introduced a discriminative feature extraction technique for a VAD system to improve the speech and non-speech classification [6].

There have been several attempts to use long-term temporal information in ASR [7, 8]. In their VAD approach, Ramires et al. proposed a framework to exploit long-term spectral divergence between speech and non-speech intervals [9]. In contrast, Poeppel used psychological test to show that human beings use two types of temporal information extracted from both short (20 to 40 ms) and long (150 to 250 ms) temporal windows to understand spoken language [10]. These previous studies suggest that long-term temporal information can lead to robust VAD systems, but there is still room for improvement.

Our work applies long-term temporal information to a feature parameter set for stochastic-model-based VAD in a simple way and improve the VAD performance in low SNR conditions. In the proposed method, long-term dynamic features are obtained from enlarged delta window lengths in the linear regression calculations. The long-term dynamic features have significant effect on performance in the low SNR conditions and greatly reduced the false alarms when the delta window lengths are larger than the average phoneme duration in each utterance.

This paper is organized as follows. Section 2 shows the effectiveness of dynamic features as a long-term temporal information representation, and Section 3 includes an outline of the proposed method, experimental results, and provides some discussion. Finally, Section 4 presents our conclusions.

2. Long-term dynamic feature

2.1. Dynamic feature extraction

Dynamic features are widely used in standard ASR systems. They capture the variations of spectrum envelopes along the time axis. In ASR systems, the dynamic features are generally calculated with a short window length consisting of several consecutive frames and are combined with static cepstral coefficients. Based on this established practice in ASR, the dynamic features are often used for the VAD system, too. The first-order derivative of the cepstral sequence is called Δcepstrum. The Δcepstrum d_t at time t is estimated as

$$d_t = \sum_{k=1}^{K} \{k \cdot (c_{t+k} - c_{t-k})\} / 2\sum_{k=1}^{K} k^2 , \qquad (1)$$

where c_t is the cepstrum coefficient at time t. In Equation (1), $2K$+1 consecutive frames are used to extract Δcepstrum. A window of length $2K$ +1 frames for Δcepstrum is often called a delta window. In a standard ASR system, the value of K is set to from two to four, based on the frame size, the frame

rate, and other parameters. Thus similar values are used in a conventional VAD system. However, the usefulness of the long-term speech information was little investigated in VAD systems, although longer time intervals provide more useful VAD information. A straightforward approach to incorporate more temporal context into the feature parameters is to simply widen the delta window by increasing the window length. Below, we discuss the effectiveness of dynamic features for VAD when long deltas (or large windows) are used. The Δcepstrum parameters obtained from short and long delta window lengths are referred to short-term and long-term dynamic features, respectively.

In the discrimination stage of VAD, the log likelihood ratio computed from speech and non-speech GMMs is

$$L(x) = \log P(x \mid \Lambda_{Sp}) - \log P(x \mid \Lambda_{Sil}) \qquad (2)$$

where Λ_{Sp} and Λ_{Sil} indicate speech and non-speech (or silence) GMM, respectively. A GMM-based VAD decides on speech and non-speech frames by comparing $L(x)$ to an appropriate threshold.

2.2. Distribution of log likelihood ratio

This section describes the effectiveness of the long-term temporal speech information in relation to the log likelihood ratios. Fig. 1 shows distributions of the log likelihood ratio $L(x)$ when the speech present and the speech absent frames of signals recorded in a car-environment are input to the VAD system. In the figure, static MFCC, short-term Δcepstrum (K=3) alone, and long-term Δcepstrum (K=9) alone are compared as feature parameters. The left side of the figure illustrates distributions in high SNR conditions with clean, 20 dB, 15 dB, and 10 dB signals, while the right side of the figure shows low SNR conditions of 5 dB, 0 dB, and -5 dB. The VAD system can minimize detection errors if the overlap of the speech present and the speech absent distribution is small. As can be seen in the figure, the long-term Δcepstrum has small overlaps for the two distributions of high and low SNR conditions compared to the MFCC and short-term Δcepstrum. This suggests that the long-term speech information allows more precise detection of speech present and speech absent frames. The intersection point of the speech present and speech absent distribution is taken as a good threshold for the discrimination stage of VAD.

2.3. Long-term dynamic feature extraction as a filtering process

Here we discuss the long-term dynamic feature extraction in terms of a filtering process of a modulation spectrum. Drullman et al. showed that high-pass filtering above 4 Hz or low-pass filtering below 16 Hz for the modulation spectrum did not reduce the speech intelligibility [11, 12]. Kanedera et al. found that most of the useful linguistic information of speech lies in the modulation frequencies ranging from 1 to 16 Hz and especially from 2 to 10 Hz [13]. Similar to these findings, a linear regression calculation used in the (short-term) dynamic feature extraction was also regarded as the filtering process which emphasized a beneficial modulation spectrum for ASR. Fig. 2 shows the frequency responses of linear regression filtering with 7 frames (K=3), which extracts short-term dynamic features and with 17 frames (K=8) which extracts long-term dynamic features. In the figure, the attenuation amplitude of the filtering process is adjusted to compare two different window lengths. This shows that the

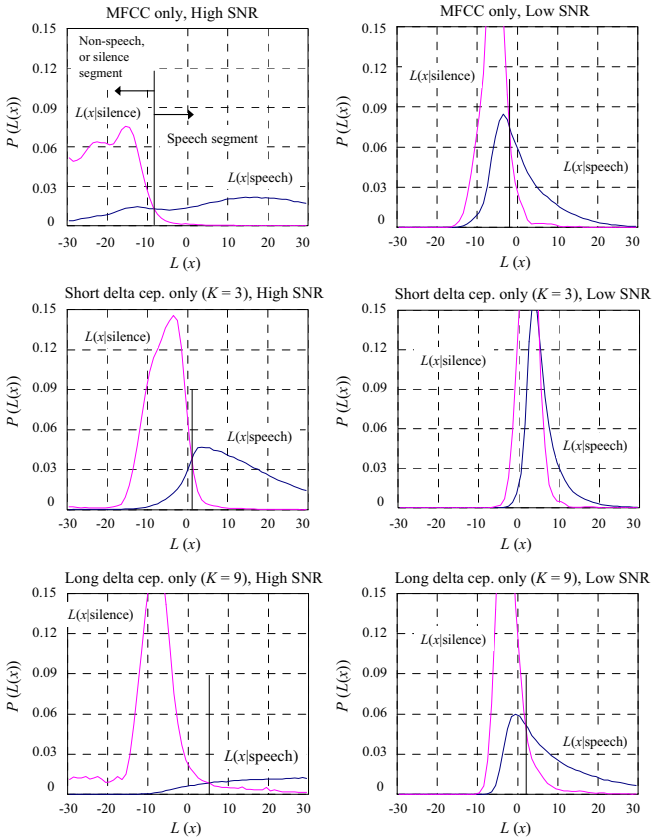

Figure 1: *Differences of log likelihood ratios for each feature parameter.*

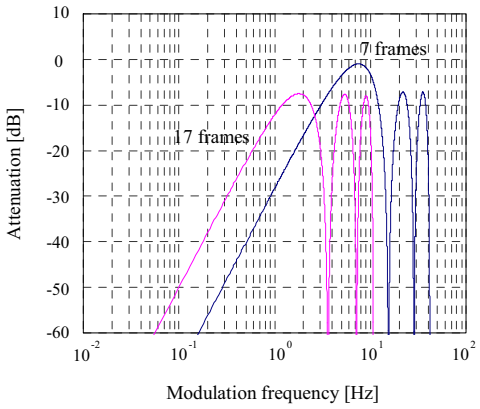

Figure 2: *Frequency responses of linear regression filtering with 7 and 17 frames.*

modulation frequencies around 10 Hz are emphasized by short-term linear regression filtering with 7 frames while those around 2 Hz are emphasized by long-term linear regression filtering with 17 frames. In [10], Poeppel suggested that both short-term variations and long-period spectral information were important for listening to spoken words. In a similar manner, we showed that the combination of short-term and long-term temporal information in speech improved the ASR performance [14]. For the VAD system, the long-term dynamic features focus on slowly changing spectral variations including formants in vowels that show the robustness against noises as compared to consonants, and are thus expected to allow for noise-robust VAD.

3. Experiment

3.1. Experimental Setup

The CENSREC-1-C Japanese connected digit corpus for VAD [15] from the IPSJ-SIG SLP noisy speech recognition evaluation working group in Japan was used in our experiment. The CENSREC-1-C database tasks involve continuous digit strings recorded in several situations including an automotive environment. In the experiments, first, the speech data in the automobile was used. Driving noise was added to the clean speech in 5dB increments from 20dB to -5dB. There are 6,986 sentences uttered by 52 male and 52 female speakers in the test set.

The input speech was sampled at 8 kHz and each 25 ms speech segment was pre-emphasized using a filter $H(z) = 1 - 0.97z^{-1}$ every 10 ms. A 256-point FFT for each frame was then applied after the framed speech signal was Hamming-windowed. The resultant FFT power spectrum was integrated into the outputs of band-pass filters with 24 channel mel-scaled center frequencies. Then the outputs of the band-pass filters were converted into 12 static cepstrums (MFCC) by using a discrete cosine transform. Finally, Δcepstrum was extracted with Equation (1). Each feature parameter involves a power and/or Δpower coefficient. An Aurora-2J corpus which is provided by the same working group was used to train speech and non-speech GMMs for VAD. There are 1,668 sentences uttered by 55 male and 55 female speakers. The number of mixtures for each GMM was set to 32. The appropriate threshold was selected at the discrimination stage in each delta window length. The VAD system was evaluated in terms of detecting voiced segments correctly.

3.2. Experimental Result

In this section, we discuss the effectiveness of Δcepstrum when long deltas or large windows are used. Figure 3 shows experimental results. The Δcepstrum alone, not including the MFCC was investigated in the experiment. The Δcepstrum drastically improved the performance as the delta window length was increased, though it showed poor performance for less than K=3. The Δcepstrum yielded best performance when the delta window length was ten (K=10), which showed an accuracy of 82.7%.

Table 1 compares the long-term dynamic feature to other feature vectors for GMM-based VAD and conventional methods including the standard G.729 Annex B [16], an energy-based VAD provided by the European Telecommunication Standards Institute (ETSI) [17], and the long-term spectral divergence (LTSD) based VAD [9]. The numbers in brackets indicate the number of dimensions in each feature vector. In the table, short-term Δcepstrum is extracted with K=3 and long-term Δcepstrum is obtained with K=10 (which showed the greatest accuracy). The G.729 VAD and the LTSD-VAD suffered poor speech detection performance in low SNR condition. In contrast, the ETSI-VAD led to considerable improvements over the G.729 and the LTSD-VAD in the low SNR condition but lower performance under the high SNR condition. In a comparison with 13 dimension feature vectors for the GMM-based VAD, the long-term Δcepstrum significantly outperformed static MFCC, the short-term Δcepstrum, and the conventional methods both in high and low SNR conditions. It had 71.8% and 47.7% error reductions on average in correct rate and accuracy in comparison with MFCC alone.

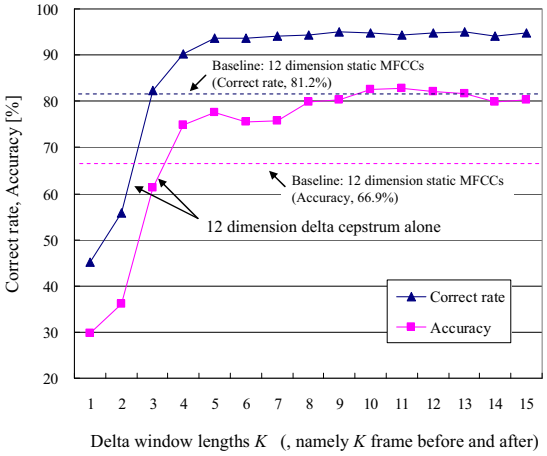

Figure 3: *Performance for different delta window lengths.*

Table 1. *Performance in each feature parameter set.*

Conventional methods (top), Feature parameters (middle, bottom)	Correct rate [%]		
	High SNR	Low SNR	Average
G.729 Annex B	92.9	49.5	74.3
ETSI AFE-VAD (ES 202 050)	87.3	79.4	83.4
Long-term spectral divergence (LTSD)	96.5	56.4	79.3
MFCC (13 dim.)	94.6	63.2	81.2
Short-term ΔCep. (13 dim. K=3)	93.9	66.8	82.3
Long-term ΔCep. (13 dim. K=10)	99.7	88.1	94.7
MFCC + ShortΔCep. (26 dim. K=3)	99.1	82.9	92.2
MFCC + LongΔCep. (26 dim. K=10)	99.7	89.1	95.2
Conventional methods (top), Feature parameters (middle, bottom)	Accuracy [%]		
	High SNR	Low SNR	Average
G.729 Annex B	88.6	20.1	59.2
ETSI AFE-VAD (ES 202 050)	69.8	32.0	50.9
Long-term spectral divergence (LTSD)	88.7	22.6	60.3
MFCC (13 dim.)	90.5	35.5	66.9
Short-term ΔCep. (13 dim. K=3)	86.5	27.5	61.2
Long-term ΔCep. (13 dim. K=10)	97.8	62.4	82.7
MFCC + ShortΔCep. (26 dim. K=3)	96.3	58.3	80.0
MFCC + LongΔCep. (26 dim. K=10)	97.2	68.2	84.8

Next we considered 26-dimension feature vectors. MFCC with conventional short-term Δcepstrum showed higher performance than MFCC alone. The long-term Δcepstrum with 13-dimension feature vector, in contrast, performed better than MFCC with short-term Δcepstrum in spite of lower number of dimensions. Also, the long-term Δcepstrum combined with MFCCs showed significant improvements. This means that long-term temporal speech information suppresses false alarms of the VAD system and detects speech present intervals more accurately.

Performances of the proposed VAD in other noisy environments including subway, crowd of people (babble), car, and exhibition noises at different SNRs were finally investigated. Table 2 shows experimental results. The proposed VAD system using long-term Δcepstrum achieved better performance as compared with using short-term Δcepstrum in all of these cases. However detecting and recognizing only the target speaker's voice is difficult in an environment such as the 'babble' case which contains other people's voices as noise sources. In this situation, combining other techniques such as a microphone array system with VAD seems promising.

Table 2. *Performance in each noise environment.*

Feature Parameter	Correct rate [%]			
	Subway	Babble	Car	Exhibition
MFCC + ShortΔCep. (26 dim.)	92.2	85.4	92.2	93.8
MFCC + LongΔCep. (26 dim.)	95.8	88.0	95.2	96.1
Feature Parameter	Accuracy [%]			
	Subway	Babble	Car	Exhibition
MFCC + ShortΔCep. (26 dim.)	77.1	65.9	80.0	78.5
MFCC + LongΔCep. (26 dim.)	83.7	67.4	84.8	82.0

3.3. Discussion

Here we consider the relationship between speech rate and the window length for dynamic feature extraction. The average phoneme duration for all of the test data was 98.6 ms, according to the hand-labeled voiced segment information in CENSREC-1-C. As shown in Figure 3, the proposed dynamic feature-based VAD system showed comparatively high performance for K more than 5, which indicates the 100 ms window length of cepstral sequences in the 10 ms frame shift, and approximately corresponds to the average phoneme duration. Based on this result, another experiment was conducted for Δcepstrum with a subset of the test data consisting of utterances in which the average phoneme duration is below 80 ms, and one considering of utterances in which the average phoneme duration is above 120 ms. Figure 4 shows the experimental results. This shows that the long-term Δcepstrum approached an upper limit of performance at K=4, which means the 80 ms window length, and at K=6 which indicates the 120 ms window length. The proper delta window lengths for both test sets corresponded to their phoneme duration. The lengths of the phonemes vary, but the long-term features of interest span several phonemes. Therefore the delta windows to extract these long-term features should be adjusted over average phoneme length.

4. Conclusions

This paper described GMM-based VAD using long-term temporal information in the speech. The proposed system extracts temporal information by calculating dynamic features with enlarged delta window lengths. The long-term dynamic features showed better log likelihood ratio distributions than the standard MFCC. In experiments on a Japanese digit corpus, the proposed method achieved significant improvements when using the long-term dynamic features calculated over more than the average phoneme duration in the utterances.

In future work, we will investigate the robustness using spontaneous speech corpora and evaluate the ASR system combined with the proposed VAD system.

Acknowledgements

The present study was conducted using the CENSREC-1-C database developed by the IPSJ-SIG SLP Noisy Speech Recognition Evaluation Working Group.

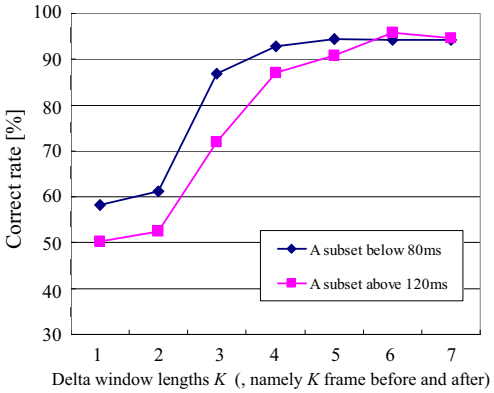

Figure 4: *Relationships between speech rate and delta window length.*

References

[1] S. V. Gerven and F. Xie, "A comparative study of speech detection methods," *Proc. Eurospeech '97*, vol. III, pp.1095-1098 (1997).

[2] J. Sohn, N. S. Kim, W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, Vol. 6, pp. 1-3 (1999).

[3] Y. D. Cho and A. Kondoz, "Analysis and improvement of a statistical model-based voice activity detector," *IEEE Signal Processing Letters*, Vol. 8, No. 10, pp.276-278 (2001).

[4] A. Martin, D. Charlet, and M. Manuuary, "Robust speech / non-speech detection using LDA applied to MFCC" *Proc. ICASSP '01*, vol. I, pp.237-240 (2001).

[5] K. Li, M. N. S. Swamy, and M. O. Ahmad, "An improved voice activity detection using higher order statistics," *IEEE Trans. Speech and Audio Processing*, Vol. 13, No. 5, pp. 965-974 (2005).

[6] K. Yamamoto, F. Jabloun, K. Reinhard, and A. Kawamura, "Robust endpoint detection for speech recognition based on discriminative feature extraction," *Proc. ICASSP '06*, vol. I, pp.805-808 (2006).

[7] H. Hermansky and S. Sharma, "TRAPS – Classifiers of Temporal Patterns", *Proc. ICASSP '99*, Vol. I, pp. 289-292 (1999).

[8] B. Chen, Q. Zhu, and N. Morgan, "Learning long-term temporal features in LVCSR using neural networks", *Proc. ICSLP*, pp. 612-615 (2004).

[9] J. Ramirez, J. C. Segura, C. Benitez, A. Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech Communication*, Vol. 42, pp. 271-287 (2004).

[10] D. Poeppel, "The analysis of speech in different temporal integration windows: cerebral lateralization as asymmetric sampling in time," *Speech Communication*, Vol. 41, pp. 245-255 (2003).

[11] R. Drullman, J. M. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech perception," *J. Acoust. Soc. Amer.*, Vol. 95, pp. 1053-1064 (1994).

[12] R. Drullman, J. M. Festen, and R. Plomp, "Effect of reducing slow temporal modulations on speech perception," *J. Acoust. Soc. Amer.*, Vol. 95, pp. 2670-2680 (1994).

[13] N. Kanedera, H. Hermansky, and T. Arai, "Desired characteristics of modulation spectrum for robust automatic speech recognition," *Proc. ICASSP'98*, pp.613-616 (1998).

[14] T. Fukuda, O. Ichikawa, and M. Nishimura, "Short- and Long-term Dynamic Features for Robust Speech Recognition," *Interspeech 2008.*

[15] N. Kitaoka et. al, "Development of evaluation framework for voice activity detection under noisy environment," *IPSJ sig. technical reports*, 2006-SLP-63, pp. 1-6 (2006), *in Japanese*.

[16] ITU-T recommendation G.729-Annex B, "A silence compression scheme for G.729 optimized for terminals conforming to recommendation V.70," (1996).

[17] ETSI ES 202 050 recommendation, "Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithm" (2002).