Journal of Computational Information Systems 10: 1 (2014) 1–8
Available at http://www.Jofcis.com

Enhanced Voice Activity Detection based on Automatic Segmentation and Event Classification ⋆

Yulong Wan*, Xianliang Wang, Ruohua Zhou, Yonghong Yan

Key Laboratory of Speech Acoustics and Content Understanding,
Institute of Acoustics, Chinese Academy of Sciences, Beijing, China

Abstract

Robust voice activity detection (VAD) is very important for automatic speech processing systems. Current VAD techniques commonly have a serious problem that the accuracy drastically reduces when used for noisy speech mixed with non-speech segments such as music and other stationary signals. This paper proposes a novel VAD using automatic segmentation and audio event classification, which shows good robustness against the non-speech sounds in real application environment. In our approach, we exploit harmonic structure information to segment the input stream and then classify the segments into speech and non-speech, finally combine the speech ones to form the result. Experiments in real environments were conducted to evaluate the proposed method by comparing it with typical methods. Results show that the proposed method led to considerable improvements, especially under low Signal Noise Ratio (SNR) conditions. Additionally, the recognition error rate was reduced by 25.91∼42.97% compared with energy based VAD for language recognition task in our test.

Keywords: voice activity detection; automatic segmentation; harmonic structure; noise robustness; audio event classification

1 Introduction

Voice activity detection (VAD) is the key technology for automatically distinguishing speech periods from non-speech in continuous audio streams. It is widely used as the first stage for various applications of speech technology, such as robust automatic speech recognition (ASR) [1], speech enhancement techniques, language recognition and speaker verification systems. According to the past research reports, a major cause of errors in these systems is the inaccurate VAD method, and developing a robust VAD for real environments with low signal-to-noise ratio (SNR) or mixed with non-speech signals is still a challenging task. An excellent VAD can not only reduce computation

⋆This work is partially supported by the National Natural Science Foundation of China (Nos. 10925419, 90920302, 61072124, 11074275, 11161140319, 91120001, 61271426), the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant Nos. XDA06030100, XDA06030500), the National 863 Program (No. 2012AA012503) and the CAS Priority Deployment Project (No. KGZD-EW-103-2).

*Corresponding author.
Email address: wanyulong@hccl.ioa.ac.cn (Yulong Wan).

January 2014

by eliminating unnecessary transmission and processing of non-speech segments, but also reduce potential mis-recognition errors in such segments.

In designing a robust VAD for real-world recordings, there are two aspects that must be taken into consideration. The first one is the robust features of speech on which the noise reduction can be applied. The second is how to construct the decision rules to classify the underlying segments into speech or non-speech. In this paper, both are considered.

During past decades, there have been many previous studies on robust features of speech for noise reduction, and a variety of related algorithms have been proposed. Classic methods were carried out with thresholds of the signal power, the number of zero crossings, fundamental frequency, linear prediction coefficients, and cepstral features [2], as these features are indeed effective under clean conditions, and can be used to detect complete speech periods in clean environments (where there is only target speech). However, the accuracy of these methods reduces remarkably when the background noise is much louder or more complex. More recently, in order to improve the detection accuracy of VAD in noisy environments, methods based on high order statistics [3, 4], adaptive sub-band energy sequence analysis [5] and harmonic detection [6] have been proposed. The decision rules of these methods are derived from the likelihood ratio test, and are known to work well for ambient noises. However, these approaches also have weaknesses as they can not remove non-speech segments such as music, rings, and faxes, etc., as these signals may have similar characteristics to speech such as periodicity or aperiodicity, and non-stationary properties. While in more real-world multimedia domains, speech is typically interspersed with segments of music and other background noise, as a result, standard speech recognizers attempting to perform recognition on all input frames will naturally produce high error rates with such a mixed input signal. Therefore, a pre-processing stage that can classify the signal segments into speech and non-speech is invaluable for improving recognition accuracy. Lu proposed a content-based method for audio classification and segmentation in [7], which employed support vector machines (SVMs) to classifying sub-segments into five classes, however it also had a limitation that the segments' length is fixed as one-second, without consideration of real segments' boundaries.

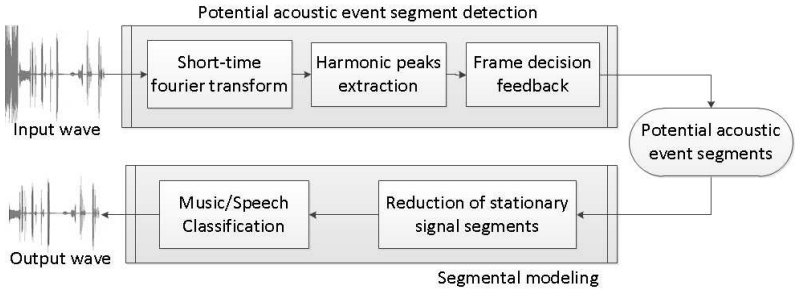

Fig. 1: Overall flow of the proposed system

To address these problems mentioned above, we adopt automatic audio segmentation and classification, and propose an enhanced VAD method that can provide the following capabilities: 1) automatic detection of potential acoustic event segments in the input audio stream, which could also be designed to provide additional interesting information, such as the division into speaker turns and the speaker identities (allowing, e.g., for an automatic indexing and retrieval of all occurrences of a same speaker); 2) classification of the segments and removal of the non-speech sounds in the continuously listening environment. Unlike most other approaches to VAD, ours proceeds in two stages: segmentation, followed by classification. The two stages are shown in the

first and second rows, respectively, in Figure 1. We describe each stage in turn in the sections below.

2 Potential Acoustic Event Segment Detection

Harmonic structure is widely regarded as discriminative speech feature, which is credible even under extremely noisy conditions, and very favorable for VAD. For example, Guo proposed a VAD algorithm based on harmonic detection in [5], Takashi proposed an improved VAD using static harmonic features in [8] and Chuangsuwanich proposed an approach based on harmonicity and modulation frequency in [6]. In this section, a novel noise-decision method based on harmonic structure is described.

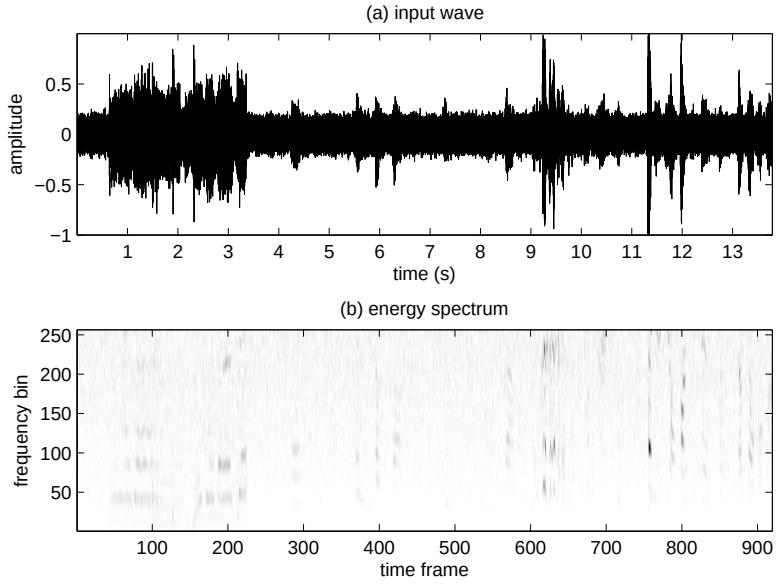

Fig. 2: Audio stream under low SNR and corresponding energy spectrum

The basic principle of our proposed algorithm depends on predefined threshold of the Mean of N-Largest Peaks (MNLP) in each frame of log-energy spectrum, and depending on that threshold the decision is made for each audio frame. The frame is termed as active frame marked by 1 if it is potential, otherwise is termed as inactive frame marked as 0.

The basic architecture of MNLP-based segmentation is illustrated in the first row of Figure 1. Essentially, it consists of an initial time-frequency analysis via Short-Time Fourier Transform (STFT) with each frame 15 ms long and non-overlapped. With the assumption that the speech and additional noise are independent of each other, the energy of each point can be given by $ES(k,t) = ES_{pe}(k,t) + ES_n(k,t)$, where $ES(k,t)$ denotes the energy of the t-th frame at k-th frequency bin and $ES_{pe}(k,t)$ and $ES_n(k,t)$ represent the potential event and noise part respectively. Actually, noise is also a non-stationary signal, and its energy sequence $ES_n(k,t)$ can not be estimated precisely. Additional noise can be classified to 5 classes: stable noise, slow-varying noise, impulse noise, fluctuant noise and step noise. All the kinds of noise are independent and additional, and the input noise is their sum. Based on the discussion in [5], the energy of noise can be expressed as $ES_n(k,t) = ES_{sn}(k,t) + ES_{nn}(k,t)$, where $ES_{sn}(k,t)$ denotes the energy of stationary noise (SN) composed of stable noise, slow varying noise and the stationary section of

step noise, whereas $ES_{nn}(k,t)$ is a non-stationary sequence made from other noise. Hence, in the total energy of $ES(k,t) = ES_{pe}(k,t) + ES_{sn}(k,t) + ES_{nn}(k,t)$, the $ES_{sn}(k,t)$ can be simply reduced by a median filter, and the $ES_{pe}(k,t)$ and $ES_{nn}(k,t)$ are both non-stationary sequences that are difficult to be discriminated, but can also be accomplished by the following steps. Here are the details for SN reduction, obtaining the MNLP and automatic segmentation method:

1. Remove the overall SN from the original Energy Spectrum (ES) as Eq. 1 and Eq. 2:

$$ES_{sn} = \frac{1}{N \times F} \sum_{t=1}^{N} \sum_{k=1}^{F} ES(k,t) \tag{1}$$

$$ES_{pe+nn}(k,t) = \begin{cases} 0, & \text{if } ES(k,t) <= ES_{sn} \\ ES(k,t) - ES_{sn}, & \text{otherwise} \end{cases} \tag{2}$$

where ES_{pe+nn} denotes the result ES of SN reduction, mixed up with the ES of potential event and non-stationary noise.

2. Convert the SN reduced ES to log-energy spectrum (LES) as follows in Eq. 3, and then normalize it to the range of $0 \sim 1$ as Eq. 4, which is shown in Fig. 3:

$$LES(k,t) = 20log_{10}(1 + ES_{pe+nn}(k,t)) \tag{3}$$

$$LES(k,t) = \frac{LES(k,t)}{\max\limits_{1 \le t \le N, 1 \le k \le F}(LES(k,t))} \tag{4}$$

where t is the frame index and k ($0 \le k \le N_{FFT}/2$) is the frequency bin index. As the most effective acoustic bandwidth of human speech is typically limited to the frequency range between $0 \sim 2$k Hz, and the sample rate of test data is 8k Hz, we just choose the frequency bin range between $(0, N_{FFT}/4)$.

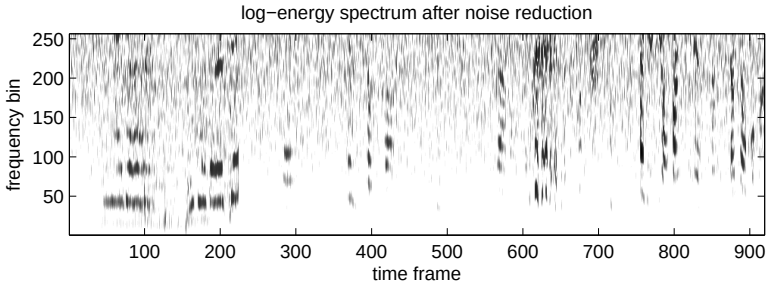

Fig. 3: Log-energy spectrum after stationary noise reduction

3. Pick up the N-largest peaks in each frame, and calculate the mean value of them as MNLP for the current frame following the Eq. 5.

$$MNLP_t = \frac{1}{N_{max}} \sum_{j=1}^{N_{max}} LES(k,t) \tag{5}$$

where $k = arg[max(LES(t))]$ are the frequency bin indexes of the N-largest peaks in current frame.

By comparing the MNLP with the predefined threshold, each frame can be termed as active frame marked by 1 or inactive frame marked by 0 as in Eq. 6:

$$Frame_t = \begin{cases} 1, & \text{if } MNLP_t \geq \sigma \\ 0, & \text{if } MNLP_t < \sigma \end{cases} \quad t = 1, 2, 3...N \tag{6}$$

where, σ is the predefined threshold, and we choose 0.16 in this paper.

4. Combine the nearby active frames to form potential segments. Considering that the content of audio stream is always continuous, the likelihood of audio class transfer is not too sudden nor frequent, i.e., the audio clip of the same class usually lasts for quite a few seconds. Under this assumption, we apply smoothing in labeling an audio sequence following the rules below:

$$if\ N_{s[i]} \leq N_{min}\ \&\ s[i-1] \neq s[i]\ \&\ s[i-1] = s[i+1] \qquad then\ s[i] = s[i-1]$$

where $s[i]$ stands for the index of the i-th segment, $N_{s[i]}$ denotes the number of frames in $s[i]$, and N_{min} denotes the minimum of the frame number in final segments, here we set it to 2. The rule implies that if the middle segment is too short and different from the other two while the other two are the same, the middle one is considered as misclassification. Fig. 4 shows the MNLP result and segmentation result for the audio stream sample in Fig. 2:

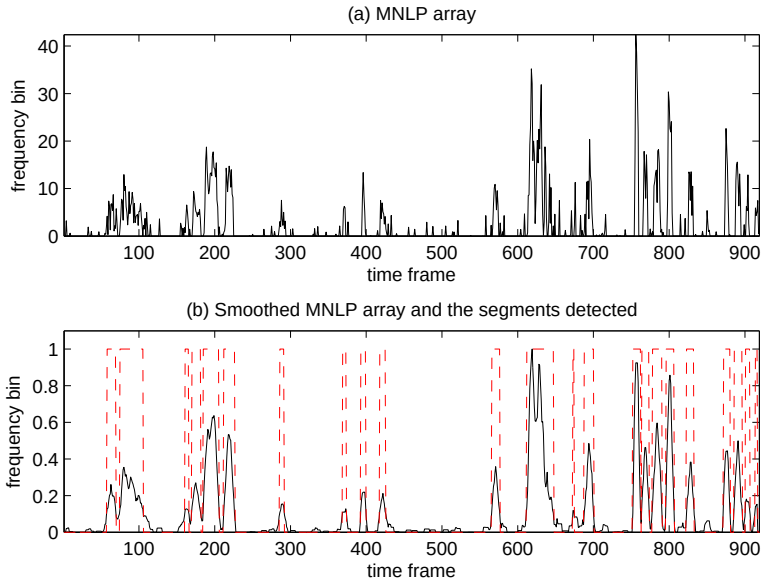

Fig. 4: MNLP result and segmentation result

3 Audio Segment Classification

For systems such as audio content indexing, automatic transcription of audio broadcasts, language recognition, etc., speech is mainly interspersed with segments of music, as a result, after acoustic event segments detection, the remaining segments can be seemed to only consist of music and speech. Thus, to detect speech segments, we just need a two-classes classifier.

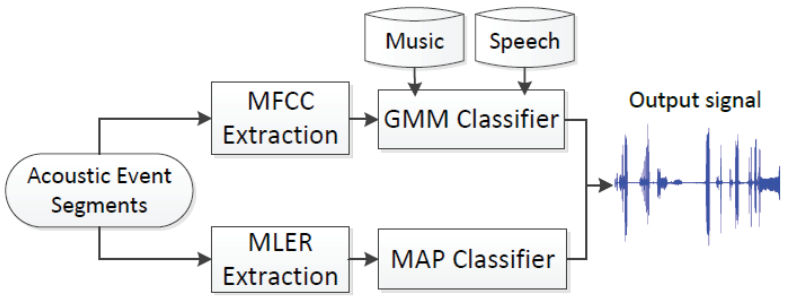

Fig. 5: Segment classification module

3.1 Music and speech segment classification

Here, we employed a traditional Gaussian Mixture Model (GMM) classifier and a Bayes Maximum A Posteriori (MAP) rule based on Modified Low Energy Ratio (MLER) feature, which was introduced in [9].

For each segment, we firstly extract its MLER feature as the definition in Eq. 7:

$$MLER_i = \frac{1}{2N_i}\sum_{n=1}^{N_i}[sign(lowthres_i - LES(n)) + 1] \tag{7}$$

$$lowthres_i = \delta \cdot \frac{\sum_{n=1}^{N_i} LES(n)}{N_i} \tag{8}$$

where N_i is the total number of frames in the i-th segment, $LES(n)$ is the short time log-energy of the n-th frame, δ is a control coefficient, which decides how low the $LES(n)$ needs to be so that the frame is considered as "low energy" and greatly affects the discrimination power of the feature for speech and music.

Then Bayes MAP decision rule is applied to decide the type of the segment, i.e., music or speech. Since the feature is only one-dimensional, the process can be simplified and modeled as:

$$S_i = \begin{cases} Music, \text{ if } MLER_i < \lambda \\ Speech, \text{ if } MLER_i \geq \lambda \end{cases} \quad i = 1, 2, 3, ..., N_s \tag{9}$$

Where S_i is the label of i-th segment, λ is a threshold which corresponds to the MLER value of the cross point of speech and music. More details of the MLER based music and speech classifier can be found in [9].

We also use traditional GMM classifier to classify the potential speech segments labeled by the MLER method. The features we used are Mel-frequency cepstral coefficients (MFCC) (12 MFCC coefficients, 50-ms window, 25-ms shift, and 24 bands in a Mel filter bank). Each feature vector is augmented with its first-order derivative which results in 24-dimensional feature vectors. GMM with M-mixed compositions can be expressed as:

$$p(v_i|\lambda) = \sum_{m=1}^{M} w_m p_m(v_i|\lambda_m) \tag{10}$$

where v_i is the 24-dimensional observation MFCC vector of i-th frame, M is the mixed number, here we use 32, $w_m(m = 1, 2, ..., M)$ are the mixed weights, and $\sum_{m=1}^{M} w_m = 1$. $p_m(v_i|\lambda_m)$ is a 24-dimensional Gaussian function.

The GMM training was performed with the traditional iterative Expectation-Maximization (EM) algorithm to ensure $p(v|\lambda)$ maximum. In our experiment, training process is iterative for 10 times.

In the classification stage, probability-score is used to classify the test segment: The mean of posteriori log-probability on an test segment is calculated for both music and speech models first. Then the audio segment is labeled according to speech or music that has the maximum score.

Finally, we combine all of the speech segments to form the final result audio stream.

4 Experiments

4.1 Segment classification performance

To evaluate the performance of proposed segment classification method, we conducted experiments based on a corpus collected from real life environment. All the data are presented as standard 16-bit, 8k Hz PCM digital data with the average duration of 10 minutes, from 30 seconds to 30 minutes. The SNR of these data varies significantly from 20 dB to 5 dB. We segmented and labeled these data manually and finally allocated them for training, development, and testing as follows in Table 1:

Table 1: Dataset allocation

	Train	Development	Test
Music	200	100	5377
Speech	200	100	4546

Table 2: Results of speech/music classification

		Precision	Recall	F-Measure
GMM	music	96.87%	96.58%	0.9672
	speech	95.97%	96.30%	0.9613
GSV-SVM	music	92.86%	96.06%	0.9443
	speech	95.14%	91.27%	0.9316
GMM-UBM	music	95.73%	95.15%	0.9544
	speech	94.30%	94.98%	0.9464
Proposed	music	95.59%	98.75%	0.9714
	speech	98.46%	94.61%	0.9650

The performance of our music and speech classification method is compared with GMM alone classifier, the GMM-UBM classifier proposed in [10] and Gaussian Super-vectors with Support Vector Machines classifiers (GSV-SVM) proposed in [11] using the same feature vectors. And the results are shown in the Table 2:

From the results above, we can see that with the same features, the proposed method gets better classification performance than the other three methods especially in recall of music and precision of speech. And this performance is much more fit for language recognition and speaker verification applications, as these systems require lower false-alarm rate.

4.2 VAD performance for real application

In order to evaluate the actual application effect, we carried out language recognition experiments using the proposed VAD on a dataset containing 1158 real-life telephone conversation recordings and on a subset chosen from NIST LRE 2011 evaluation database containing 4131 files. There were 9 target languages in all: English, Japanese, Korean, Mandarin, Cantonese, Hokkien, Russian, Uighur and Tibetan. We use the recognition Error Rate for evaluation and the result is compared to the same system using three other VADs, which we named VAD I, VAD II and VAD III separately. VAD I was carried out using energy based segmentation only; VAD II was VAD I followed by event classification method; VAD III using the proposed segmentation without event classification. And the comparison results are shown in Table 3:

Table 3: Language recognition results

Methods	Error Rate	
	NIST LRE 2011	Real-life Recordings
VAD I	9.344%	11.054%
VAD II	9.393%	8.722%
VAD II	6.851%	8.808%
Proposed	6.923%	6.304%

From the results, we can observed that our proposed VAD can reduced at most 25.91% error rate compared with energy based VAD for NIST database, and for real-life recordings with music and ringing segments, the error rate can be reduced by at most 42.97%.

5 Conclusion

In this paper, we propose an enhanced VAD algorithm based on automatic segmentation and audio event classification. In our algorithm, the harmonics structures of the input sound are captured, and used for noise reduction and segmentation. Then the segments are classified into speech and music using a fusion method of common GMM classifier and MNLP based discrimination. The comparison results show that our VAD significantly outperforms traditional energy-based VAD under non-stationary noise conditions, especially for the audio streams mixed with music and ringing signals. As the spectral-temporal analysis works on the standard Fourier spectrogram, the VAD can be easily integrated into conventional speech processing applications.

In the future, we will extend our algorithm to deal with more difficult tasks, by carefully re-designing and selecting features with a more complicated decision mechanism.

References

[1] SI, Yujing, et al. "Block Based Language Model for Target Domain Adaptation towards Web Corpus." Journal of Computational Information Systems 9.22 (2013): 9139-9146.

[2] Haigh, J. A., and J. S. Mason. "Robust voice activity detection using cepstral features." TENCON'93. Proceedings. Computer, Communication, Control and Power Engineering. 1993 IEEE Region 10 Conference on. IEEE, 1993.

[3] Cournapeau, David, and Tatsuya Kawahara. "Evaluation of real-time voice activity detection based on high order statistics." INTERSPEECH. 2007.

[4] Aronowitz, Hagai. "Segmental modeling for audio segmentation." Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on. Vol. 4. IEEE, 2007.

[5] Guo, Yanmeng, Qian Qian, and Yonghong Yan. "Robust voice activity detection based on adaptive sub-band energy sequence analysis and harmonic detection." INTERSPEECH. 2007.

[6] Chuangsuwanich, Ekapol, and James R. Glass. "Robust Voice Activity Detector for Real World Applications Using Harmonicity and Modulation Frequency." INTERSPEECH. 2011.

[7] Lu, Lie, Hong-Jiang Zhang, and Stan Z. Li. "Content-based audio classification and segmentation by using support vector machines." Multimedia systems 8.6 (2003): 482-492.

[8] Fukuda, Takashi, Osamu Ichikawa, and Masafumi Nishimura. "Improved voice activity detection using static harmonic features." Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on. IEEE, 2010.

[9] Wang, W. Q., W. Gao, and D. W. Ying. "A fast and robust speech/music discrimination approach." Information, Communications and Signal Processing, 2003 and Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint Conference of the Fourth International Conference on. Vol. 3. IEEE, 2003.

[10] Reynolds, Douglas A., Thomas F. Quatieri, and Robert B. Dunn. "Speaker verification using adapted Gaussian mixture models." Digital signal processing 10.1 (2000): 19-41.

[11] Campbell, William M., Douglas E. Sturim, and Douglas A. Reynolds. "Support vector machines using GMM supervectors for speaker verification." Signal Processing Letters, IEEE 13.5 (2006): 308-311.