A New VAD Framework Using Statistical Model and Human Knowledge Based Empirical Rule

Ji Wu, Xiao-lei Zhang, Wei Li

Multimedia Signal and Intelligent Information Processing Laboratory,
Tsinghua National Laboratory for Information Science and Technology,
Department of Electronic Engineering, Tsinghua University, Beijing, China
wuji_ee@tsinghua.edu.cn, huoshan6@126.com, liwei1976@gmail.com

Abstract

This paper presents a new voice activity detection (VAD) framework that is based on the empirical rules and statistical models. First, the VAD framework detects the candidate endpoints efficiently in the time domain with empirical rules which are based on the human knowledge and the nature of the speech continuousness, and then it confirms the candidate endpoints in the transform domain with different confirmation schemes for beginning-point and ending-point. Particularly in the transform domain, a new algorithm called sliding-window double-layer confirmation (SWDC) is proposed and employed to confirm the endpoint accurately, and sensitive data, which is used for GMM training, are proposed for our detection scheme. The experiments show that the proposed VAD framework achieves better performances in various environmental conditions.

Index Terms: empirical rules, GMM training, SWDC, VAD

1. Introduction

Voice activity detector (VAD) refers to the classical problem of distinguishing speech from background noise and has applications for a variety of speech communication systems, such as speech recognition, speech coding, noisy speech enhancement. The VAD strategy is becoming more and more complicated in order to be robust in real world environments. In the past few decades, many features and approaches were attempted including short-term energy, pitch detection, zero-crossing rate, energy-entropy feature, cepstral feature, teager energy, higher-order statistics, order statistics filter, multiband techniques, etc.

The research of statistical model was another field. Sohn [1] adopted the Gaussian statistical model that the discrete Fourier transform (DFT) coefficients of speech and noise processes were asymptotically independent Gaussian random variables, Gazor [2] further assumed that the discrete cosine transform (DCT) coefficients of the speech and noise processes followed Laplace and Gaussian distributions respectively, Chang [3] analyzed the Gaussian, Laplace and Gamma distributions in the DFT domain and integrated them with goodness-of-fit (GOF) test, Tahmasbi [4] supposed speech process, which was transformed by GARCH filter, having a variance gamma distribution, and Ramirez [5] employed revised multiple likelihood ratio test (MO-LRT) instead of single frame LRT [1].

Besides the algorithms mentioned above, many rules were attempted as well, which were based on the characteristics of the speech, the VAD detecting schemes themselves, the human knowledge, etc. Davis [6] designed a state machine based hangover scheme to lower the probability of false rejections, ETSI frame dropping (FD) VAD [7] was somewhat an assemble of rules that were based on the continuousness of speech, Ramirez [5] proposed the contextual multiple hypothesis which utilized the characteristics of the empirical minimum speech length, and Kuroiwa [8] designed a grammatical system where many human knowledge based grammars were developed.

The human knowledge based rules could not only distinguish the apparent noise from speech and also cover the trivial speech period easily missed, but they are less helpful in detecting the endpoints accurately; The statistical models could detect the voice activity exactly but sometimes they are less efficient when compared with other methods. And in the respect of machine learning methods related to VAD, the traditional schemes prefer to use as much training data as possible, but the effectiveness of the training data was not considered.

In this paper, we present a new VAD framework which combines the empirical rules and statistical models together after rational feature selection, and we propose a sensitive data based statistical model training method, which is used to improve the performance of our detecting scheme correspondingly. The rest of the paper is organized as follows. Section 2 presents the proposed VAD framework, followed by a presentation of the sliding-window double-layer confirmation (SWDC) algorithm for the endpoint detection and the speech sensitive data used for the matching training of statistical models. Section 3 discusses the performance of the proposed VAD under various noise conditions and compares its performance with that of 6 other algorithms. Finally, Section 4 summarizes the findings.

2. VAD framework and algorithm

2.1. VAD framework

For the framework, on one side, we employ the double threshold energy detection algorithm [9] in time domain with the short-term energy used as its feature; On the other side, we use the SWDC algorithm in transform domain where the mel-frequency cepstral coefficients (MFCC) are involved.

In this paper, we assume that there are frame based time series $[t_1, t_2, ..., t_N]$, whose short-term energy are $[E_1, E_2, ..., E_N]$, and MFCC features are $[X_1, X_2, ..., X_N]$ correspondingly.

1) Beginning-point (BP) detection scheme:

   a) Candidate BP detection in time domain: This module employs double threshold energy detection algorithm [9] to identify apparent noise, and transfer the candidate BP to the next module.

b) BP confirmation in transform domain: The GMM based SWDC algorithm is used to confirm the candidate BP. The algorithm will be presented in detail later.

2) Ending-point (EP) detection scheme:

a) EP range detection in time domain: The double threshold energy detection algorithm is employed to find one candidate EP, t_p, and we define its neighborhood $[t_p - \delta, t_p + \delta]$ as the scanning range for the next module, where δ is a constant.

b) The optimal EP detection in transform domain: The SWDC algorithm is used to scan the range given by the last module, and the optimal EP could be found after the sliding-window slides over the range.

Note that our framework is easily extended, any algorithm that is fast but maybe not robust enough, or any complex but accurate one can be introduced into the framework, as long as they complement each other to some extent.

2.2. Empirical rules based energy detection

The double threshold energy detection algorithm [9] is widely used by VAD. However, the algorithm will be in trouble when the SNR is low, so that we combine the algorithm with empirical rules, which are based on our experiences and the continuousness of speech process. The revised algorithm used in the proposed framework is presented briefly as follows:

1) In BP detection, the silence threshold and the low\high energy thresholds of the current nth frame are obtained by

$$E_{sil} = \frac{1}{3}\sum_{j=n}^{n+2} E_j \tag{1}$$

$$E_{low} = \alpha \cdot E_{sil} \quad , \quad E_{high} = \beta \cdot E_{sil} \tag{2}$$

where α, β are the energy threshold factors, which are related to signal-to-noise ratio (SNR). Given one signal segment starting from the current detecting position with a length of fixed frame number L_A, if there are several consecutive frames with a count L_B that the energy of each frame is higher than E_{low} and the ratio L_B/L_A is higher than an empirical threshold φ_{BP}^{low}, the first frame whose energy is higher than E_{low}, denoted as t_{low}, should be remembered; and then we detect the given segment starting from t_{low}, if there are another consecutive frames with a count L_C that the energy of them are higher than E_{high} and the ratio L_C/L_A is higher than another empirical threshold φ_{BP}^{high}, the candidate beginning-point is detected in the segment as t_{low} mentioned above.

2) In EP detection, we assume that the energy of the current frame is lower than E_{low}, and then in the subsequent signal segment with a length of fixed frame number L_D, there might be several frames with a count L_E that the energy of the frames is lower than E_{high}, if the ratio L_E/L_D is higher than an empirical threshold φ_{EP}, the candidate ending-point is detected as the current frame.

3) If the time span of the detected speech segment is shorter than the minimum predefined speech length Γ_{min}, the segment has little chance to be speech and should be discarded.

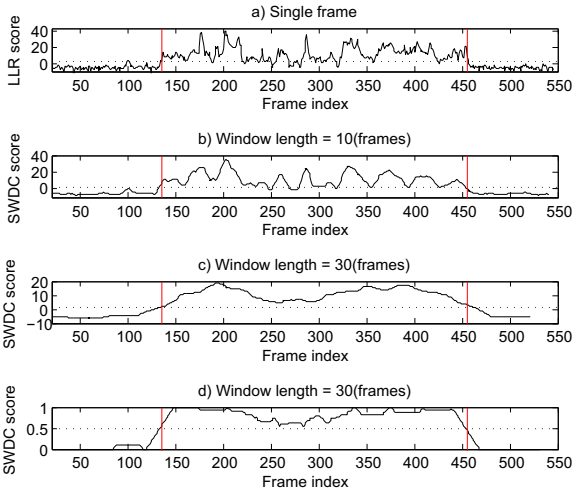

Figure 1: *SWDC scores (SNR = 15dB). The vertical solid lines are the endpoints of the utterance, the transverse dotted lines are the decision thresholds. a) LLR scores of single frame. b) SWDC scores adopting LLR scores on the first layer and window length = 10. c) SWDC scores adopting LLR scores on the first layer and window length = 30. d) SWDC scores adopting 0-1 scores on the first layer (threshold for 0-1 scores is 1.5) and window length = 30*

The revised algorithm is efficient to help covering trivial speech and rejecting noise wrongly detected, but it can't help much detecting the endpoints accurately, especially the EP. Therefore the SWDC algorithm is proposed to verify the candidate endpoints.

2.3. SWDC algorithm

In the VAD realization, SWDC is proposed for the candidate endpoint confirmation in transform domain. It has two layers presented as follows:

Assuming that the candidate endpoint t_d is to be confirmed, the sliding window is defined as a window that is centered on the t_d, with L frames on its backward and M frames on its forward, and has a length of $L + M + 1$ frames totally.

1) We extract a new feature within the sliding window for the second layer:

The GMMs are used to model the speech and the noise period separately with MFCCs as the features. Two hypotheses for each frame in the window are presented as H_i, where $i = \{0, 1\}$, indicating speech absence and presence respectively, then the probability density functions conditioned on H_0 and H_1 are given by

$$P(X_n|H_i) = \sum_{k=1}^{K} \pi_{i,k}\mathcal{N}(X_n|\mu_{i,k}, \Sigma_{i,k}) \tag{3}$$

where X_n is the MFCC feature of the frame t_n, $n = d - L, ..., d + M$, K is the mixture number of the GMM. $\pi_{i,k}$ are the mixing coefficients of the GMM under H_i, $\mathcal{N}(X_n|\mu_{i,k}, \Sigma_{i,k})$ is the kth component of the mixture under H_i and has its own mean $\mu_{i,k}$ and covariance $\Sigma_{i,k}$.

After the conditioned probability for each frame in the sliding window is calculated under two different hypotheses by

(3), the score of each frame, denoted as I_n, is calculated in the following two ways:

- Log likelihood ratio (LLR) score for each frame: The score of the nth frame in the window is given by

$$\begin{aligned} I_n &\triangleq \log\left(\Lambda(X_n)\right) \\ &= \log\left(P(X_n|H_1)\right) - \log\left(P(X_n|H_0)\right) \end{aligned} \tag{4}$$

  Figure 1 (a) shows the score curves for single frame.

- Zero-one (0-1) score for single frame: The hard decision on LLR score is made to achieve the 0-1 score

$$\log\left(\Lambda(X_n)\right) \underset{I_n=0}{\overset{I_n=1}{\gtrless}} \varepsilon \tag{5}$$

  where ε is the threshold for the 0-1 score, and could be assigned adaptively in the initialization of the proposed VAD by adding the average value of the LLR scores with a constant Δ.

Once the scores have been calculated, we could get a new feature, $\mathbf{I_d}$= $[I_{d-L} \dots I_d \dots I_{d+M}]^T$.

2) We verify the candidate endpoint t_d with the new feature: Given the new feature $\mathbf{I_d}$, many classifiers could be designed to confirm t_d.

$$\Lambda_d = f(\mathbf{I_d}) \tag{6}$$

where Λ_d is the SWDC score of t_d, $f(\cdot)$ denotes the function of the classifier. For example, given decision threshold η, the linear classifier could be presented as

$$\Lambda_d = \mathbf{g}_d^T \cdot \mathbf{I_d} = \sum_{n=d-L}^{d+M} g_n I_n \underset{H_d \in H_0}{\overset{H_d \in H_1}{\gtrless}} \eta \tag{7}$$

where $\mathbf{g} = [g_{d-L}, ..., g_{d+M}]^T$ is the linear weight and could be a time-variant or time-invariant vector. Other classifiers, such as SVM, could be attempted as well. Figure 1 (b)-(d) show the SWDC score curves with linear classifier differing in sliding window length and the detection scheme on the first layer.

In our realization of the VAD, the 0-1 score calculation scheme is adopted on the first layer, and the simple linear classifier is used on the second layer by using $\mathbf{g} = 1/(L+M+1) \cdot [1, ..., 1]^T$ with $L = M$. Because the SWDC is used for both BP and EP detection, η_{begin} and η_{end} in stead of η in (7) is given separately.

2.4. Sensitive data based GMM training

In our framework, statistical models are used to confirm the endpoints, so that the input of the SWDC algorithm is the neighborhood data of the endpoints. It's easy to understand that there will exist mismatching if we use all data for training. To deal with this issue, sensitive data based GMM training is proposed.

Sensitive data is the neighborhood data of the endpoints, and will be more matching with our detection scheme, when compared with other parts of the speech segment.

The expectation-maximum (EM) algorithm is employed to train the GMM, the speech part of the sensitive data is used for speech GMM and the noise part for noise GMM. The sensitive data based GMM not only could be trained with lighter load than the traditional method both in the memory and in the time complexity, but also could help to improve the performance.

3. Experiments

3.1. Databases

The TIMIT [10] speech corpus contains utterances from 8 different dialect regions in the USA. It consists of a training set of 326 male and 136 female speakers , and a testing set of 112 male and 56 female speakers, each utters 10 sentences, so that there are 4620 utterances in the training set and 1680 utterances in the testing set totally. All recorded speech signals are sampled at $fs = 16$kHz.

These TIMIT sets, after resampling from 16kHz to 8kHz, are distorted artificially with the NOISEX corpus [11]. Firstly, the original TIMIT and NOISEX corpora are filtered by intermediate reference system (IRS) [12] to simulate the phone handset, and then the SNR estimation algorithm based on active speech level [13] is used to add four different noise types (babble, factory, vehicle and white noise) at five SNR levels over a range of [5, 10, ..., 25 dB]. As [14] did, the TIMIT word transcription is used for VAD evaluation, and the inactive speech regions, which are smaller than 200ms are set to speech. The percentage of the speech process is 87.78%, which is much higher than the average level of true application environment, so that every utterance is artificially extended at the head and the tail respectively with some noise, whose length equals to 1/5 that of the utterance. The percentage of the speech is, afterwards, reduced to 62.83%, and the renewed corpora are more suitable for VAD evaluation.

3.2. Parameter settings

The frame length is 25ms with an overlap of 10ms, the slipping window length is 30 frames with a window shifting step size of 5 frames, the ε used in equation (5) is set in the VAD initialization for each utterance by adding the average LLR score of the first 20 frames with a constant Δ, and Δ is set to 1.5, the scanning range for ending-point mentioned before has a radius δ of 75 frames, the minimum legal speech length Γ_{min} is set to 35 frames, the other parameters related to SNR are show in Table 2

Table 2: *Parameters for VAD implementation.*

SNR (dB)	5	10	15	20	25
α	1.30	1.30			
β	1.90	2.50			
η_{begin}	0.27	0.45	0.55	0.60	0.65
η_{end}	0.2	0.25	0.40	0.50	0.55

In respect of GMM training, the sensitive data is defined as the neighborhood data of the endpoints with a radius of 50 frames in the training set.

For GMM training set, we extract 231 utterances randomly from every noise distorted corpus to form a noise-independent corpus and then train a pair of noise-independent models (NIM) with 5 mixtures. Note that the new noise-independent corpus has 4620 utterances totally, which has an equivalent utterance number with each former noise distorted corpus. We do this mainly because that the application environment is unknown. Our experiments proved that the performances of the proposed VAD changed slightly with different GMM mixture numbers and the application environments even when the noise type was known and the noise-dependent models (NDM) were used.

Table 1: *Performances of different VAD (%).*

	G.729B		AFE WF		AFE FD		Sohn		Ramirez		Tahmasbi		Proposed	
Noise type	RC	FA	RC	FA	RC	FA	RC	FA	RC	FA	RC	FA	RC	FA
Babble	81.58	44.57	95.35	25.65	99.99	79.89	85.78	20.93	90.93	16.40	84.28	20.56	96.23	13.67
Factory	78.81	42.03	94.15	19.66	99.98	74.68	84.41	21.35	90.11	16.02	85.64	21.39	96.10	10.45
Vehicle	74.77	37.62	91.46	19.66	99.98	71.55	86.23	10.68	93.10	5.60	85.80	14.74	95.94	4.81
White	72.32	38.07	90.95	5.75	99.94	70.81	87.03	8.18	93.47	4.56	87.06	11.39	95.86	4.29

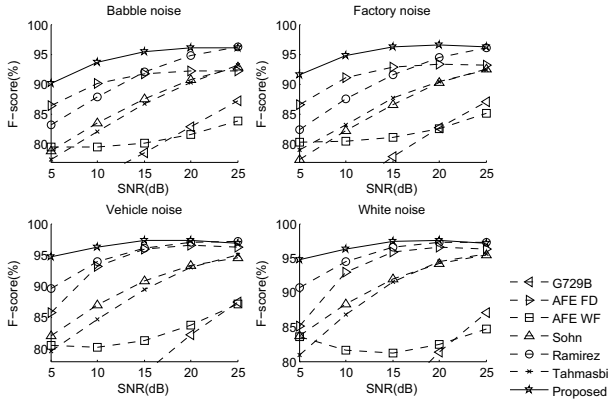

Figure 2: *Fscore values in different noise conditions.*

3.3. Results and analysis

For comparison, besides the proposed VAD with 5 mixture NIM, the G.729B VAD [15], the VAD from ETSI AFE ES 202 050 for DSR [7], which are the VAD for noise estimation (AFE WF VAD) and for frame dropping (AFE FD VAD), the Sohn VAD [1], the Ramirez VAD [5] and the Tahmasbi VAD [4] are also tested and compared against the proposed VAD. For simplicity, the mean values of the recall probability (RC) and the false alarm probability (FA) are averaged over different SNR levels under the same type of noise and are shown in Table 1.

As can be seen, firstly, the G.729B, the AFE WF and AFE FD VAD, which are the open sources, have relatively comparable performances with the Sohn, Ramirez, Tahmasbi VAD, this conclusion is identical with [14][1][5], etc. Secondly, the RCs of the proposed framework are more desirable than the others, while the FAs keep in a lower level.

To evaluate the performances in general, the harmonic mean $Fscore$ of the precision rate (PR) and the RC, which is employed here from [14], is calculated as follows

$$Fscore = \frac{2 \cdot RC \cdot PR}{RC + PR} \tag{8}$$

The higher the $Fscore$ is, the better the VAD performs. Figure 2 shows the results in terms of $Fscore$ in various noise conditions. It shows that the curves yielded by the proposed VAD are higher than the others, and is robust to SNR. Note that the proposed VAD also has a relatively high-efficiency due to its simple energy based algorithm in time domain.

4. Conclusions

This paper presents an efficient and robust VAD framework that is based on empirical rules and statistical models, and the framework can be easily extended. One new algorithm called SWDC is proposed and employed to confirm the endpoint accurately, and sensitive data based GMM training are proposed for our special detection scheme. The experiments show that the proposed VAD scheme achieves better performances in various environments.

5. References

[1] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, 1999.

[2] S. Gazor and W. Zhang, "A soft voice activity detector based on a Laplacian-Gaussian model," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 11, no. 5, pp. 498–505, 2003.

[3] J. H. Chang, N. S. Kim, and S. K. Mitra, "Voice activity detection based on multiple statistical models," *IEEE Trans. Signal Process.*, vol. 54, no. 6, pp. 1965–1976, 2006.

[4] R. Tahmasbi and S. Rezaei, "A soft voice activity detection using GARCH filter and variance Gamma distribution," *IEEE Trans. Audio, Speech and Language Proc.*, vol. 15, no. 4, pp. 1129–1134, 2007.

[5] J. Ramírez, J. C. Segura, J. M. Górriz, and L. García, "Improved voice activity detection using contextual multiple hypothesis testing for robust speech recognition," *IEEE Trans. Audio, Speech and Language Proc.*, vol. 15, no. 8, pp. 2177–2189, 2007.

[6] A. Davis, S. Nordholm, and R. Togneri, "Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold," *IEEE Trans. Audio, Speech and Language Proc.*, vol. 14, no. 2, pp. 412–424, 2006.

[7] ETSI, "Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms," *ETSI ES*, vol. 202, no. 050.

[8] S. Kuroiwa, M. Naito, S. Yamamoto, and N. Higuchi, "Robust speech detection method for telephone speech recognition system," *Speech Commun.*, vol. 27, pp. 135–148, 1999.

[9] L. R. Rabiner and M. R. Sambur, "An algorithm for determining the endpoints of isolated utterances," *Bell System Tech. J.*, vol. 54, no. 2, pp. 297–315, 1975.

[10] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, and N. Dahlgren, "DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM," *NTIS order number PB91-100354*, 1993.

[11] The Rice University, "Noisex-92 database," http://spib.rice.edu/spib.

[12] ITU-T Rec. P.48, *Specifications for an intermediate reference system*, ITU-T, March 1989.

[13] ITU-T Rec. P.56, *Objective measurement of active speech level*, ITU-T, 1993.

[14] T. V. Pham, C. T. Tang, and M. Stadtschnitzer, "Using Artificial Neural Network For Robust Voice Activity Detection Under Adverse Conditions," in *Int. Conf. Computing and Commun. Tech., RIVF '09.*, 2009, pp. 1–8.

[15] 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 Commun. Mag.*, vol. 35, no. 9, pp. 64–73, 1997.