FUZZY CLUSTERING AND BAYESIAN INFORMATION CRITERION BASED THRESHOLD ESTIMATION FOR ROBUST VOICE ACTIVITY DETECTION

Ye Tian, Ji Wu, Zuoying Wang, and Dajin Lu

Department of Electronics Engineering
Tsinghua University, Beijing 100084, P. R. China
tianye@thsp.ee.tsinghua.edu.cn

ABSTRACT

In previous Voice Activity Detection (VAD) approaches that using threshold, consistent accuracy cannot be achieved since the mean-value based and the histogram based threshold estimation algorithms are not robust. They strongly depend on the percentage of voice and background noise in the estimate interval. In this paper, fuzzy clustering and Bayesian Information Criterion are proposed to estimate the thresholds for VAD. Compared to previous algorithms, the new algorithm is more robust and heuristic-rules-free. It is insensitive to the estimated interval, and can maintain fast tracking speed of environment change when combining with online update. Experiment shows it works very well with energy feature in both stationary and non-stationary environments.

1. INTRODUCTION

Voice Activity Detection (VAD) is an important front-end of speech recognition, speech coding and speech communication. There is a strong need of robust VAD algorithm with more and more speech recognition systems employed in real application.

The threshold based VAD algorithm extract some measured features from the input signal and compare to thresholds. Voice active decision is made if the measured values exceed the threshold [1]. Features widely used are the time-domain energy [2], the entropy [3], the high frequency and the low frequency energy [4], the NP parameter [5], and et al.

Unfortunately, consistent accuracy cannot be achieved since the threshold estimation algorithms are not robust. They often use background noise interval to estimate the threshold, and the first several seconds of recording are assumed to contain no voice. In mean-value based algorithm, the threshold level is determined by the mean-value of background noise intervals. If the interval also contains voice, the threshold will be arbitrary. Histogram based algorithm use the Probability Distribution Function (PDF) [2] and the background level is estimated from the peak of the histogram. It is more robust then mean-value based global algorithm, because it is insensitive to a small quantity of voice data. However, it depends on the percentage of voice and background noise in the interval. Fig. 1 shows a energy histogram, in which the peak value occurs at the mean value of voice, but not at that of background noise.

Robust threshold estimation algorithm should work with any estimated intervals, without caring of the percentage of voice and background noise in the interval. Only this kind of algorithm can be well used for threshold online updating to track the environment change. From Fig. 1, we can see that the histogram shows two local peaks, corresponding to voice and background noise, respectively. Accurate threshold estimation can be obtained by discovering local peaks. However, local peaks detection in the histogram is not an easy way. Even complex rules cannot work very well. Moreover, enough data is needed to obtain accurate histograms.

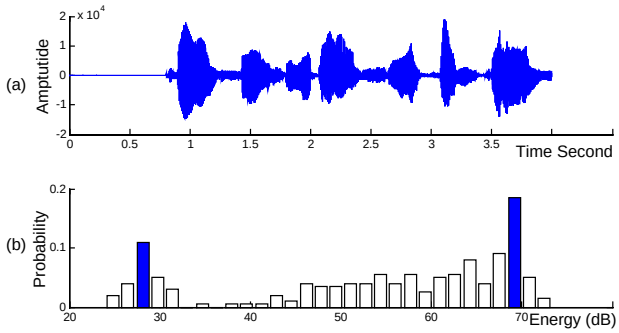

Figure 1. Energy histogram (a) the waveform of a background noise-voice interval, (b) The histogram of energy in dB

In this paper we propose a new threshold estimate algorithm based on fuzzy clustering and Bayesian Information Criterion. This algorithm can be applied on any feature that has discrimination between voice and background noise. For an interval of waveform, features of each frame are extracted. These features are organized into clusters by fuzzy c-Means clustering. The Bayesian Information Criterion is used to determine the best cluster number. There are two clusters for interval contains both voice and background noise, and only one cluster for background noise interval. Centers of these clusters describe the mean of voice and background noise. They can be used to determine the thresholds for VAD.

The advantage of using fuzzy clustering is that it is unsupervised pattern recognition technology to handle the case where class labeling of the training patterns is not available. It discovers similarities and differences among patterns automatically. It is heuristic-rules-free and small data is needed to get good estimation. The Bayesian Information Criterion, which has theoretical advantage, is used to determine the best cluster number. Compared to the mean-value based and based threshold estimation algorithm, our algorithm is more robust. It

is insensitive to the estimated interval, and can maintain fast tracking speed of environment change when combining with online update.

The paper is organized as follows: section 2 gives a brief introduction to fuzzy clustering; section 3 shows the selection of cluster number; their application in robust VAD is given in section 4; section 5 is the experiment results.

2. FUZZY CLUSTERING

The fuzzy clustering schemes have been the subject of intensive research during the past two decades. In the probabilistic schemes each vector belongs *exclusively* to a single cluster, while in the fuzzy clustering schemes a vector *simultaneously belongs* to more than one cluster. [6]

Assume that there are N data, $x_1, x_2, \ldots, x_N$. C is the expected cluster number and $m_1, m_2, \ldots m_C$ are the center of the clusters. The fuzzy clustering algorithm is derived by minimizing the cost function of the form

$$J = \sum_{j=1}^{C}\sum_{i=1}^{N}\left[\mu_j(x_i)\right]^b \left\|x_i - m_j\right\|^2 , \qquad (1)$$

where $\mu_j(x_i)$ is the grade of membership of x_i in the j -th cluster, and b is the *fuzzifier* parameter. In the fuzzy c-Means cluster algorithm, $\mu_j(x_i)$ subjects to the constrains,

$$\sum_{j=1}^{C}\mu_j(x_i) = 1, \quad i = 1,2...N , \qquad (2)$$

Minimization of J subject to the constrains (2), leads to the following function:

$$m_j = \frac{\sum_{i=1}^{N}\left[\mu_j(x_i)\right]^b x_i}{\sum_{i=1}^{N}\left[\mu_j(x_i)\right]^b}, \quad j = 1,2,\ldots,C , \qquad (3)$$

$$\mu_j(x_i) = \frac{\left(1/\left\|x_i - m_j\right\|^2\right)^{1/(b-1)}}{\sum_{k=1}^{C}\left(1/\left\|x_i - m_k\right\|^2\right)^{1/(b-1)}}, \quad \begin{matrix} i = 1,2,\ldots,N \\ j = 1,2,\ldots,C \end{matrix} . \qquad (4)$$

Fuzzy c-Means clustering

Given the cluster number C and fuzzifier parameter b .

Choose initial estimate of cluster centers $m_1, m_2, \ldots m_C$

REPEAT

Determine $\mu_j(x_i)$ according to Equ (4).

Update the cluster center m_j according to Equ (3).

UNTIL no change in the cluster centers.

3. THE CLUSTER NUMBER SELECTION

For an interval of waveform, we must judge if it is background noise only, or it also contains voice. In this paper the Bayesian Information Criterion (BIC) is proposed to determine the best cluster number.

BIC is well known as a model selection criterion in the statistics literature. It can be also purposed to choose the number of the cluster [7]. BIC penalizes likelihood criterion by the

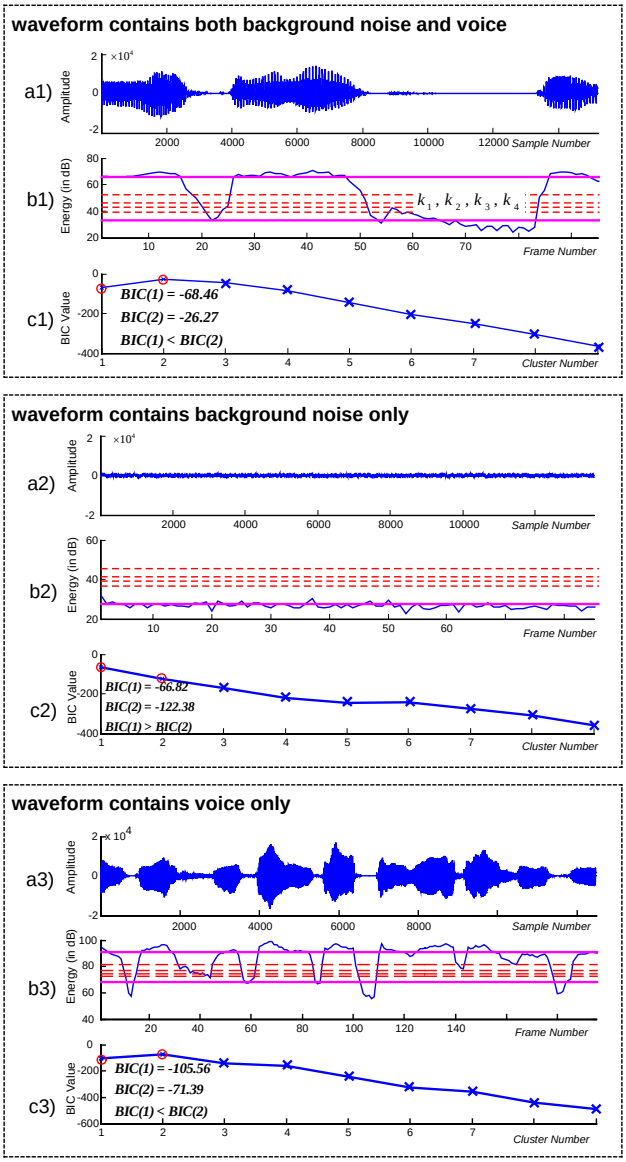

Figure 2. illustration of VAD based on fuzzy clustering and BIC (a1)(a2)(a3) the waveform of a voice and background-noise interval, background noise interval, and voice interval, respectively. (b1)(b2)(b3) The energy of the waveform in dB. The solid lines show the centers of the two clusters, the dashed lines show the four thresholds calculated from the centers, (c1)(c2)(c3) the BIC value versus the cluster number.

model complexity: the number of parameters in the model. The BIC value is defined as

$$BIC(M) = \log L(X,\Phi) - \lambda_p \frac{1}{2}\#\Phi \times \log(N) , \qquad (5)$$

where $X = \{x_1, x_2, \ldots, x_N\}$ is the data set we are modeling, $\Phi = \{\phi_1, \phi_2, \ldots, \phi_C\}$ is the parametric models. $L(X,\Phi)$ is the likelihood function between data X and model Φ . $\#\Phi$ is the number of parameters of model Φ . N is the number of samples. λ_p is the penalty weight.

According to BIC, the best model number is the one with maximized BIC value. If voice and the background noise are modeled as multi-variance Gaussian distribution $N(\mu_i, \Sigma_i)$, where μ_i is the mean vector and Σ_i is the full covariance matrix, one can show that the BIC value at cluster number C is

$$BIC(C) = \sum_{i=1}^{C}\left\{-\frac{1}{2}N_i \log|\Sigma_i|\right\} - \frac{\log(N)}{2}\lambda_p C\left[d + \frac{d(d+1)}{2}\right], \quad (6)$$

where N is the total sample number. N_i is the number of sample in the i-th cluster. d is the dimension of the feature space.

For our VAD purpose, the best cluster number C_{best} can be present as

$$C_{best} = \begin{cases} 1 & BIC(1) > BIC(2) \\ 2 & else \end{cases}. \quad (7)$$

4. APPLICATION OF FUZZY CLUSTERING AND BAYESIAN INFORMATION CRITERION IN VAD

In this section, we will show how the fuzzy clustering and BIC can be used in VAD. Note that fuzzy clustering can be used on any features that have discrimination between voice and background noise. The energy is one of the most widely used features for threshold based VAD [2]. It is calculated on a frame-by-frame basis with a typical frame-length of 10 ms. In this paper, we will use energy as our VAD feature.

Threshold estimation algorithm

Step1: Calculate the energy of each frame in the threshold estimate interval.

Step2: Given cluster number $C = 2$, making fuzzy clustering on the energy of frames.

Step3: Use Equ (7) to determine the best cluster number C_{best}.

Step4: IF $C_{best} = 1$,

The set of energy of threshold id given by

$$k_i = m_{center} + \alpha_i, \quad (8)$$

where the coefficients α determine the relation of the thresholds. In our experiments, $\alpha_1 = 5, \alpha_2 = 8, \alpha_3 = 6, \alpha_4 = 10$.

ELSE

The mean of energy of voice and background noise are given by

$$m_{Speech} = \max\{m_1, m_2\}, \quad (9)$$

$$m_{Silence} = \min\{m_1, m_2\}, \quad (10)$$

The set of energy of threshold id given by

$$k_i = (m_{speech} - m_{silence})\beta_i + m_{silence}, \quad (11)$$

where the coefficients β determine the relation of the thresholds. In our experiments, $\beta_1 = 0.1, \beta_2 = 0.3, \beta_3 = 0.2, \beta_4 = 0.6$.

END

Step 5: Make VAD on the determined thresholds.

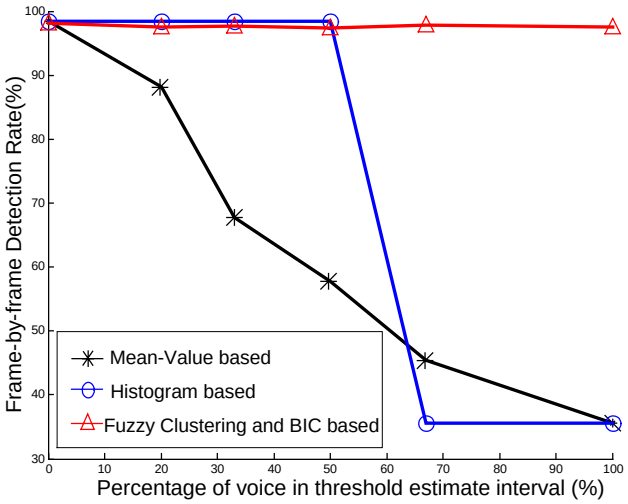

Figure 3. VAD performance of different estimate intervals

The visual result is shown in Fig.2. The four thresholds k_1, k_2, k_3, k_4 are used to find the voice-like burst during the recording interval, are the same meaning as those defined in [2]. As shown in figure, there is only one cluster for background noise interval because $BIC(1) > BIC(2)$. There are two clusters for interval contains both voice and background noise. For voice interval, we also find two clusters. The thresholds determined by the centers of cluster are quite well for VAD.

The thresholds can be obtained by initial recording interval and fixed for the following detection in stationary noise. However, when the environment is non-stationary, we need to track the change of environment fast and accurately. It is an easy case for fuzzy clustering, because thresholds can be estimate from all the intervals, without caring of the percentage of voice and background noise. Small updating time interval leads to higher background noise tracking speed, while more computation cost. The computation cost and background-tracking speed can be balanced by varying update time interval.

5. EXPERIMENTS

The proposed algorithm is compared to the mean-value based and the histogram based threshold estimation algorithms in both stationary and non-stationary environment.

5.1 Threshold estimation with different intervals

The first experiment is evaluated on the data recorded from dictation machine, and the background noise is stationary. The correct active and inactive regions of the voice signal are marked manually. To test the robustness of VAD, we use different intervals for threshold estimate. Each interval contains different percent of voice, varying from 0% to 100%. We use the threshold obtained from these intervals to make VAD on the recording data. The results are shown in Fig 3. The frame-by-frame correct detection rate is defined as the percentage of the correct classification frames relative to the total frames, including both background noise and voice frames.

Tabel 1. Performance of three types of VADs in non-stationary environment
(The correct sentence number of the file is **26**)

Thresholds estimation algorithms		VAD detected sentence number	Frame-by-frame detection rate		Speech recognizer correct-understood sentence number (rate)
			Background noise	Voice	
Fixed threshold estimated from the initial 2 seconds interval	Mean-value based	19	48.7%	99.7%	16 (61.5%)
	Histogram based	19	48.7%	99.3%	16 (61.5%)
	Fuzzy Clustering and BIC based	19	50.5%	99.3%	17 (65.4%)
Online threshold updating using the recent 2 seconds waveform	Mean-value based	21	100.0%	38.8%	12 (46.2%)
	Histogram based	27	78.5%	76.5%	18 (69.2%)
	Fuzzy Clustering and BIC based	**27**	**94.5%**	**96.3%**	**23 (88.5%)**

From the figure, we can see that the performance of mean-value based algorithm decrease with the increase of the voice percent. More voice will boost up the threshold, which will lead to the voice truncation in the following detection. When using voice-only interval to estimate the threshold, there is no voice detected because the threshold is too high, and the frame-by-frame detection rate decreases to the percent of background noise in the recording data. Histogram based algorithm is more robust. It keeps consistent performance when the voice is less than half of the estimated interval. However, if more than half of the estimated interval is voice frame, the peak of histogram will shift to the voice average energy, as shown in Fig 1. Therefore, no voice can be detected with the arbitrary thresholds.

The fuzzy-clustering based algorithm is insensitive to the estimated interval. From background noise-only interval to voice-only interval, it shows consistent performance. We can use any interval for threshold estimate without caring of the percentage of voice and background noise of the interval.

5.2 VAD in non-stationary environment

In the second experiment, we test the performance of the three types of VADs in non-stationary noise environments. The data recorded by a telephony dialogue system. The data file contains totally 26 sentences, each by different user from different place. The background is quite non-stationary. Some callers are in quite environment, while others are in noisy environment.

We use three types of VADs as front-end of dialogue understanding system. The results are shown in Table 1. The detected sentence number given by VAD shows the segmentation ability of VAD. If more than one sentence are only segmented as one sentence, the dialogue system will only give one response. Frame-by-frame detection is the percentage of the correct classification frames relative to the total frames. Both the detected sentence number and frame-by-frame detection rate are important to dialogue system, and they determine the final correct-understood sentence number.

From the table, we can see that for non-stationary environment, the fixed thresholds that estimated from initial several seconds are not enough. All VADs with fixed thresholds show poor performance. This performance is not acceptable for application.

Online threshold updating seems don't work well with mean-value based and histogram based algorithms. Voice truncation is serious, and most of the sentences are not entire. The reason is that for each update, the thresholds will be boosted up if the recent 2 seconds interval contains voice, as we have shown in the first experiment. The boosted-up thresholds will lead to the voice truncation in the following detection.

However, fuzzy clustering overcomes this defeat. It can estimate threshold from any intervals, thus keep a consistent performance. In all the algorithms, the fuzzy-clustering based threshold online updating shows the best performance, and final correct-understood sentence rate given by recognizer is improved from 61.5% to 88.5%.

6. CONCLUSIONS

Fuzzy clustering and Bayesian Information Criterion provides a good framework for threshold estimate in VADs. Experiments show that it works well in both stationary and non-stationary environment. It is robust because of theoretical advantage and heuristic-rules-free.

Fuzzy clustering can be applied to any features that have discrimination between voice and background noise, not limited to energy. Energy in time-domain shows poor performance under quite low SNRs. Future work can be focused on integrating fuzzy clustering and BIC with other robust parameter, such as entropy [3], the NP Parameter [5], and etc, for VAD in quite low SNRs.

6. REFERENCES

[1] S. G. Tanyer and H. Ozer, "Voice activity detection in nonstationary noise", *IEEE Trans. Acoust., Voice, Signal Processing*, v8, pp. 478–482, Jul. 2000.

[2] L. F. Lamel, L. R. Rabiner, A. E. Rosenberg, and J. G. Wilson, "An improved endpoint detector for isolated word recognition", *IEEE Trans. Acoust., Voice, Signal Processing*, v29, pp. 777–785, Aug. 1981.

[3] J. L. Shen, J. W. Hung, and L. S. Lee, "Robust entropy based endpoint detection for voice recognition in noisy environments", in *Proc. ICSLP'96*, 1996.

[4] K. H. Woo, T. Y. Yang, K. J. Park, and C. Y. Lee, "Robust voice activity detection algorithm for estimating noise spectrum", *Electronics Letters*, v36, pp. 180-181, Jan. 2000.

[5] P. Joseph and N. Douglas, "NP voice activity detection algorithm", in *Proc. ICASSP'95*, v1, pp.381-384, 1995.

[6] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, Academic Press, 1999.

[7] S. S. Chen, P. S. Gopalakrishnan, "Clustering via the Bayesian information criterion with applications in voice recognition", in *Proc. ICASSP'98*, v1, pp. 645-648, 1998.