Savitzky-Golay Filtering and Improved Energy Entropy for Speech Endpoint Detection under Low SNR

Zhenye Gan¹,²*, Miaomiao Hou¹, Hexiang Hou¹, Hongwu Yang³

¹College of Physics and Electronic Engineering, Northwest Normal University, Lanzhou, China
²Engineering Research Center of Gansu Province for Intelligent Information Technology and Application
³School of Educational Technology, Northwest Normal University, Lanzhou, China
*ganzy@nwnu.edu.cn

Abstract—Under the condition of low signal-to-noise ratio (SNR) or non-stationary noise, the performance of endpoint detection is poor. Therefore, this paper proposes a speech endpoint detection algorithm for low SNR. In this paper, Savitzky-Golay filtering, improved sub-band energy entropy and constant Q-transform (CQT) are used to extract features, and single parameter double threshold method is used to realize endpoint detection. In this paper, clean speech and noise-92 speech fragments are used to evaluate the algorithm. Experimental results show that the algorithm in this paper can distinguish speech endpoints well. For Gaussian white noise, Factory noise and Volvo noise, the detection accuracy of the improved algorithm is improved by 5.5%, 5.3% and 4.6% respectively. Therefore, the algorithm in this paper can separate the mute and voice in complex environment.

1. INTRODUCTION

The purpose of speech endpoint detection is to realize the separation of speech and silence, which plays a very important role in speech recognition, speech enhancement, speaker recognition and so on [1]. The existing endpoint detection methods are mainly divided into two kinds: one is based on threshold. This method first extracts the feature value of speech, and then compares the pre-set threshold value with the feature value for endpoint discrimination. This method is simple and real-time. It can also meet the requirements of real-time system. In [2], a speech/silence classification algorithm based on alternative energy is proposed. The algorithm can track the non-stationary signal and dynamically calculate the instantaneous value of the threshold using the adaptive scale parameters. In [3], an improved method based on integrated empirical mode decomposition (EMD) algorithm and Teager kurtosis is proposed. Teager energy operator is used to track the modulation energy of each IMF, which is decomposed by integrated EMD. The root power function and order statistical filter are used for feature extraction. This method can automatically estimate the threshold value by tracking the minimum value of the extracted eigenvalue. This algorithm can achieve similar results in the case of high signal-to-noise ratio (SNR). In the case of low SNR, compared with the original algorithm, it can keep lower error detection rate and higher detection rate.

In the case of low SNR, due to the interference of large noise, the signal waveform fluctuates greatly. There are many unstable burrs, and the serious spectrum distortion of the speech signal itself will bring...
difficulties to the threshold estimation. In the face of very low SNR, the filter is introduced into endpoint detection. The sub-band energy entropy ratio is used, then the median filter is used to filter the features, and finally the endpoint detection of speech is realized in [4]. This method can achieve speech separation accurately under the low SNR of Gaussian white noise, but it is not universal for different noise environments. In [5], after noise reduction by Multitaper spectrum, a new speech endpoint detection method based on EMD is proposed. Compared with the traditional energy entropy ratio algorithm, this algorithm can adapt to speech separation in very low SNR.

In [6], a speech activity detection algorithm based on Mel frequency cepstral coefficient (MFCC) distance and order statistical filter (OSF) is proposed. Although it can improve the detection effect, the detection accuracy still needs to be improved. MFCC features of silence, speech and noise are extracted, and Euclidean distance discrimination method is replaced by cosine similarity theorem to solve the problem that Euclidean distance is sensitive to numbers in [7]. Compared with the original algorithm, the overall recognition rate is increased by about 10%. However, the recognition rate needs to be improved when the SNR is very low. In [8], the method of combining energy and frequency band variance is proposed, which has a good detection effect in the environment of low SNR Gaussian white noise or colored noise.

The second method uses pattern matching decision algorithm. This algorithm needs a lot of data training, which is more complex to implement, but its recognition rate is higher. In [9], a simple frame by frame method is used for endpoint detection. In this method, the features of each frame are sent to deep neural network (DNN) model in turn, and speech discrimination is made after the probability of speech is obtained. To reduce the error probability, viterbi algorithm is used to regularize the hopping probability of the whole speech sequence. In [10], a DNN using amplitude and phase information is proposed. Compared with the DNN algorithm of amplitude, the recognition rate of the algorithm is greatly improved. The endpoint detection algorithm based on neural network has the problems of slow training speed and high complexity.

Combined with the above, this paper improves the first method, and proposes an endpoint detection algorithm based on Savitzky-Golay filtering and improved sub-band energy entropy. Experimental results show that the algorithm not only has strong anti-noise ability in low SNR, but also can accurately detect speech endpoint in complex environment, which has strong anti-interference performance.

2. CORRELATION THEORY

2.1 Savitzky-Golay Filtering

Savitzky–Golay filtering mainly realizes signal smoothing by convolution of time-domain signals and using polynomial least squares in each convolution window [11,12]. The principle of the filter is to find a suitable polynomial solution to fit the number of local signal sampling points, and smooth the local signal to achieve filtering. The polynomial is shown in Equation (1), inputting \(2W+1\) signal samples (the number of samples must be singular), letting \(v = 0\) be the center point, and \(V\) as the power sum of polynomials \(V \leq 2W + 1\).

\[
p(v) = \sum_{k=0}^{V} a_k v^k
\]

Then the Equation (2) is used to describe the mean square approximation error, and the polynomial that is approximate to the input sample is obtained. That is to say, the minimum mean value of the input signal centered on \(v = 0\). The square approximation error is used to smooth the local signal.

\[
e_v = \sum_{v=-W}^{W} \left( p(v) - x(v) \right)^2 = \sum_{v=-W}^{W} \left( \sum_{k=0}^{V} a_k v^k - x(v) \right)^2
\]
2.2 Improved Sub-Band Energy Entropy

Because the detection effect of the sub-band spectral entropy algorithm is poor in the low SNR, then it combines the sub-band spectral entropy with the sub-band energy, improves the sub-band spectral entropy, and designs the sub-band energy entropy ratio algorithm. Dividing the sub-band energy by the sub-band spectral entropy can highlight the value of the voice range, and the value of the noise range becomes smaller, which widens the numerical gap between the voice range and the noise range, and makes it easier to detect the voice endpoint.

The speech signal $x_i(m)$ is obtained by framing and windowing the noisy speech, and then the frequency signal $X_i(k)$ is obtained by discrete fourier transform (DFT). Where DFT is:

$$X_i(k) = \sum_{m=0}^{N-1} x_i(m) e^{-j2\pi km/N} \quad k \leq N \qquad (3)$$

Each frame of speech signal is divided into several sub-bands, and then the spectral entropy of each sub-band is calculated, so that the influence of noise on the spectral line can be reduced. Let each sub-band contain 4 spectral lines, and there are $M$ sub-bands in total. The sub-band energy of the $m$ sub-band in frame $i$ is:

$$E_i(m) = \sum_{k=(m-1)4+1}^{(m-1)4+4} |X_i(k)|^2 \quad 1 \leq m \leq M \qquad (4)$$

Similarly, the probability of sub-band energy $pa_i(m)$ and sub-band spectral entropy $Ha_i(m)$ can be obtained.

$$pa_i(m) = \frac{E_i(m)}{\sum_{k=1}^{M} E_i(k)}, 1 \leq m \leq M \qquad (5)$$

$$Ha_i = -\sum_{m=1}^{M} pa_i(m) \log pa_i(m) \qquad (6)$$

Combining sub-band energy with sub-band spectral entropy to get the endpoint detection of SHE, the formula is shown in (7).

$$SEH_i = \sqrt{1 + \left|\frac{SE_i}{Ha_i}\right|} \qquad (7)$$

where, $SEH_i$ represents the sub-band energy entropy ratio of the speech signal in the $i$ frame, $SE_i$ and $Ha_i$ represent the sub-band energy and sub-band spectral entropy of the speech signal in each frame respectively. The energy entropy ratio of sub-band has a good accuracy in low SNR, but it is not effective in dealing with lower SNR and other noise environment. In order to solve this problem, the energy entropy ratio of sub-band is improved.

The $M$ sub-band of the frequency domain signal is divided into $R$ sub-band blocks. The probability of the energy of the $r$ sub-band block $pb_i(r)$ and the spectral entropy of the sub-band block $Hb_i(r)$ in the $i$ frame are as follows:

$$pb_i(r) = \frac{\sum_{m=M(r-1)+1}^{M(r-1)+R} E_i(m)}{\sum_{m=1}^{M} \sum_{r'=1}^{R} E_i(m,r')}, 1 \leq r \leq R \qquad (8)$$

$$Hb_i(r) = -\sum_{k=1}^{R} pb_i(r) \log pb_i(r) \qquad (9)$$

The final energy entropy matrix of sub-band block is as follows:

$$HE_i = [SBE_i(1)/Hb_i(1), SBE_i(2)/Hb_i(2), \ldots, SBE_i(R)/Hb_i(R)] \qquad (10)$$

where, $SBE_i$ is the energy of sub-band. Using this method to block the sub-band can get more detailed eigenvalues, which can effectively reduce the noise interference on the speech and is conducive to endpoint detection.
2.3 CQT

Because the resolution of DFT is the same in high frequency and low frequency, it can’t easily reflect the speech signal, and constant Q-transform (CQT) guarantees a constant Q factor in the whole spectrum, so it has higher frequency resolution for low frequency and higher time resolution for high frequency. In order to make the sub-band block better distinguish noise and voice from entropy feature, constant coefficient Q-transform is added to open the mute and voice feature value, and simplify the selection of threshold \[13,14\].

\[
Q = \frac{f_k}{f_k - f_{k+1}} = \left(1 - 2^{-\frac{1}{b}}\right)^{-1}
\]

(11)

\[
f_k = \left(2^{\frac{1}{b}}\right)^i \cdot f_{\text{min}}
\]

(12)

where, \(f_k\) is the frequency of the \(k\)-th frequency component, \(f_{\text{min}}\) is the minimum frequency of CQT, and \(b\) is \(1/b\) octave. \(b\) is a constant, usually default to 24. Research shows that the octave of 1/24 is similar to the human auditory system, but actually for the best \(b\), it is different for different applications. This \(b\) is set to 24.

3. THE PROPOSED ALGORITHM

![Diagram](image)

Figure 1. System block diagram of the proposed endpoint detection algorithm

In this paper, the feature of frequency domain is adopted, and the time-frequency transformation of speech signal is needed. Before signal processing, Savitzky-Golay filtering can effectively reduce the interference of noise to speech signal, and distinguish the voice segment and silent segment better. The sampling frequency of the test speech signal is 16kHz. The pre-processing needs to be divided into frames and windows. The frame length is 200 sampling points, and the frame shift is 80 sampling points. The window function uses Hamming window. Through fast fourier transform (FFT), the parameter value of each frame frequency domain is obtained. The sub generation is divided evenly, and then the energy entropy ratio of each sub block is calculated. The final features are obtained by CQT, and the start and end of speech are determined by selecting appropriate threshold value, so as to detect speech frame and non-speech frame. The block diagram of the algorithm in this paper is shown in Figure 2. The specific steps are as follows:

**Step 1:** The speech signal \(x^{(m)}\) is filtered by Savitzky-Golay to get the filtered speech signal \(x^{(n)}\).

**Step 2:** The filtered speech signal \(x^{(n)}\) is windowed by frame. The speech signal \(x^{(n)}\) is divided into \(N\) frames to get the pre-processed speech signal \(x^{(i)(m)}\), where the subscript \(i\) represents the \(i\)-th frame after the framing.

**Step 3:** FFT is applied to the speech signal \(x^{(i)(m)}\), and the time domain signal is converted to the frequency domain signal to obtain the frequency domain \(A^{(m)}\). In this paper, the signal of each frame is divided into 25 sub-bands, and then evenly divided into 5 sub-bands. The energy and spectral entropy of sub-band block are obtained, and the improved energy entropy matrix \(HE\) is obtained.
Step 4: The improved sub-band energy entropy matrix is transformed by CQT, and each frame of speech signal is transformed to get two-dimensional feature matrix $C_{k,r}$. The matrix of each frame of speech signal is calculated to get the mean value, and finally the feature matrix $D$ is obtained.

$$D = \text{mean}[C_{k,R}(1), C_{k,R}(2), \cdots, C_{k,R}(N)]$$

(13)

Step 5: Calculate the maximum value of the sound section, $Det = \max(D) - \text{eth}$. Set threshold $T_1$ and $T_2$ according to $Det$, among $T_1 = a \times Det + \text{eth}, T_2 = b \times Det + \text{eth}; (a < b)$.

Step 6: When the signal value is higher than $T_2$, the voice segment is entered; when the signal value is lower than $T_1$, the silence segment is entered; until the detection of the whole voice segment is completed.

4. EXPERIMENT AND ANALYSIS

This paper uses MATLAB to simulate the experiment. 200 recorded clear speech samples are selected as the samples to be tested, and different kinds of noise and different SNR are added to them. Test speech is "zhe shi hou, liu ren an she ji hao de lu xian, zai hui yang zhen long zhen yi shan shang hui he." Fig. 2 (a).

4.1 Anti-Noise Performance Analysis

First, the SNR is set to 5dB to detect the endpoint detection with Gaussian white noise. The results of the algorithm are compared with those of the short-term energy combining sub-band variance algorithm [8] (Method I), the spectral subtraction sub-band energy entropy ratio algorithm [4] (Method II) and the improved cosine value algorithm [7] based on MFCC. The comparison results are shown in Fig. 2.

Figure 2. Comparison of detection effects of different algorithms under 5dB Gaussian white noise.

Figure 3. Comparison of detection effects of different algorithms under 0dB Gaussian white noise.
Figure 4. Comparison of detection effects of different algorithms under -5dB Gaussian white noise.

The results of Fig. 2(c) and Fig. 2(d) are spectral subtraction, and the results of Fig. 2(g) are Savitzky-Golay filtering in this paper. The signal-to-noise ratio of Fig. 2(c) and Fig. 2(d) is 13.62dB, and that of Fig. 2(g) is 11.68dB. Although Savitzky-Golay filter improves the SNR relatively less than spectral subtraction, the difference is only 1.94dB. Compared with spectral subtraction, Savitzky-Golay filtering has less influence on the spectrum and does not affect the subsequent feature extraction. Compared with Fig. 2(e), Fig. 2(f) and Fig. 2(i), the algorithm in this paper is easier to select the threshold value, and there is a large difference between the mute and the characteristic value of the voice. Compared with the energy combination band variance and the sub-band energy entropy ratio, the algorithm has a better detection accuracy and can guarantee the integrity of the detected voice. Fig. 2(h) and Fig. 2(j) are endpoint detection algorithms based on MFCC cosine value. This algorithm does not carry out filtering processing, although the detection accuracy is poor, it can still detect speech completely. Compared with the other three algorithms, the algorithm in this paper has better detection accuracy.

In order to further verify the performance of the algorithm, the SNR is further reduced. Fig. 3 and Fig. 4 are the experimental results of different algorithms at 0 and -5dB respectively. Compared with Savitzky-Golay filtering, spectral subtraction can still improve the SNR better, but it has a great influence on the amplitude of speech and the detection accuracy. Savitzky-Golay filtering is more suitable for low SNR. It can be seen from the figure that the algorithm in this paper has a good recognition effect under the lower SNR. When SNR is 0dB, the detection effect of Fig. 3(d) and Fig. 3(f) in "shi hou" is not good, and the segmentation is serious. Compared with this paper and method I, the overall detection effect is biased. As can be seen in Fig. 4, in the case of -5dB, the silence is detected as a voice signal when the energy combined with the frequency band variance appears under the spectrum subtraction, and the energy entropy ratio under the spectrum subtraction misses the voice segment, while the algorithm in this paper still has a good recognition effect. The change of cosine effect of MFCC in Fig. 3 and Fig. 4 is small, and the detection effect is different from the algorithm in this paper. The algorithm based on cosine distance can make good use of the distance between noise and silence to achieve endpoint detection, and can also deal with endpoint detection in low SNR.

4.2 Detection of Recognition Accuracy

In order to detect the accuracy of the algorithm under low SNR, the speech signals with SNR of 5dB, 0dB, -5dB and -10dB are selected for experiments, and the noise is taken from the noise standard noise database. White noise, Factory, Pink and Volvo are selected. Figure 5 shows the effect of factory noise detection under the SNR 0dB of different algorithms. It can be seen from the figure that after the factory noise is added, the noise reduction speech after the spectral subtraction contains noise, and the noise is wrongly identified as speech, which appears in the filtered waveform. Method I and method II using spectral subtraction have serious miscalculation. The algorithm based on MFCC cosine also has a lot of miscalculation. It is difficult to distinguish the feature value of speech and silence, and it can’t find a suitable threshold. Although the improved SNR is lower than the spectral subtraction, the speech...
waveform after noise reduction is similar to the original speech, which realizes complete and accurate speech endpoint detection.

In order to evaluate the performance of the algorithm more intuitively, detection accuracy is introduced, which is defined as follows:

\[
\text{Accuracy} = \frac{N - N_R}{N}
\]  

(14)

Where, \(N_R\) is the number of error frames (including the number of noise frames misjudged as voice frames and the number of noise frames misjudged as voice frames), \(N\) is the total number of frames. Adding noise to 200 segments of clean speech, and then calculate the average detection accuracy. Table 1 shows the comparison results of endpoint detection of different algorithms in different environments. Fig. 6 is the corresponding histogram.

![Figure 5. Comparison of detection effects of different algorithms under 0dB Factory noise.](image)

Table 1. Comparison of endpoint detection results of different algorithms in different environments

| SNR  | This algorithm | Method I | Method II | Ref [7] |
|------|----------------|----------|-----------|---------|
|      | Gaussian white noise | Factory noise | Pink noise | Volvo noise |
| 5dB  | 95.56%         | 89.02%   | 82.56%   | 97.02%  |
| 0dB  | 94.60%         | 86.31%   | 76.34%   | 91.18%  |
| -5dB | 91.76%         | 80.35%   | 68.17%   | 88.39%  |
| -10dB| 87.67%         | 71.63%   | 63.82%   | 84.32%  |
|      | 96.12%         | 83.93%   | 88.27%   | 95.40%  |
| 0dB  | 94.32%         | 79.37%   | 84.23%   | 90.12%  |
| -5dB | 89.51%         | 74.28%   | 76.95%   | 86.31%  |
| -10dB| 75.29%         | 68.34%   | 70.30%   | 80.34%  |
| 5dB  | 94.34%         | 82.67%   | 89.58%   | 93.69%  |
| 0dB  | 89.33%         | 78.98%   | 86.39%   | 89.31%  |
| -5dB | 84.41%         | 76.21%   | 79.63%   | 83.13%  |
| -10dB| 78.54%         | 69.95%   | 73.51%   | 77.13%  |
| 5dB  | 82.56%         | 65.23%   | 70.31%   | 86.21%  |
| 0dB  | 78.48%         | 51.21%   | 63.61%   | 80.65%  |
| -5dB | 71.24%         | 47.17%   | 57.33%   | 76.83%  |
| -10dB| 66.53%         | 38.17%   | 43.12%   | 69.17%  |

As can be seen from Table 1 and Fig. 6, the detection accuracy of this algorithm is better than that of the other three algorithms for Gaussian white noise, factory noise and Volvo noise. Especially in the case of low SNR, the algorithm in this paper also has a good recognition effect. For pink noise, compared with method I and method II, the detection accuracy of this algorithm is lower. Because Savitzky-Golay filtering can't filter pink noise well, it can't separate speech and mute very well. The algorithm based on MFCC cosine has low
5. CONCLUSIONS
In a word, the improved sub-band energy entropy endpoint detection algorithm has better effect than the original algorithm in the low SNR. This improvement can distinguish silence and speech to a large extent, and ensure the integrity of voice detection. At the same time, this paper also studies the effect of speech detection in different noise environment. For Gaussian white noise, factory noise, pink noise and Volvo noise, the improved recognition rate increases about 5.5%, 5.3%, 9% and 4.6% respectively, which has good anti-interference. In this paper, the structure of the algorithm is complex and the efficiency is low, so it is difficult to detect the large-scale speech endpoint, which needs further study.

ACKNOWLEDGMENT
This work is supported by the research fund from the National Natural Science Foundation of China (Grant No.11664036). Additionally, part of this work is performed in High School Science and Technology Innovation Team Project of Gansu (Grant No.2017C-03).
REFERENCES

[1] Zhang T, Shao Y, Wu Y. An overview of speech endpoint detection algorithms[J]. Applied Acoustics, 2020, 160: 107133.

[2] Sakhnov K, Verteletskaya E, Simak B. Approach for energy-based voice detector with adaptive scaling factor. J. Inorg. Chem. 2019; 46(4).

[3] Feng C, Zhao C. Voice activity detection based on ensemble empirical mode decomposition and teager kurtosis. In: Presented at 2015 IEEE Int. Conf. on Signal Processing. IEEE; 2015. p. 455–60.

[4] Zhang Yi, Wang Kejia, Xi Bing. Speech endpoint detection algorithm based on subband energy entropy ratio [J]. Computer Science, 44(5).

[5] Zhu Yuqing, Zhang Xiaobing, Huang Zhenkun. Multitaper spectrum and EMD improved new energy entropy ratio endpoint detection[J]. Audio Engineering, 2019, 43 (9): 17—21.

[6] Chen Z, Weilan W, Liu J. Voice activity detection algorithm based on Mel cepstrum distance order statistics filter. J Univ Chin Acad Sci 2014;31(4):524–9.

[7] Cao, Danyang, Xue Gao, and Lei Gao. "An improved endpoint detection algorithm based on MFCC Cosine Value." Wireless Personal Communications 95.3 (2017): 2073-2090.

[8] Chen Haoze, Zhang Zhijie, A speech endpoint detection method based on energy and frequency band variance[J]. Science Technology and Engineering, 2019, 19(26): 249-254.

[9] Li Wenjie, Zhang Qingqing, Zhang Pengyuan. Voice activity detection based on deep neural networks with Viterbi[J]. Journal of Chongqing University of Posts and Telecommunications (Natural Science Edition), vol 030, no 002, pp:210-215, 2018.

[10] L Wang, K Phapatanaburi, Z Go. Phase aware deep neural network for noise robust voice activity detection. Presented at 2017 IEEE Int. Conf. on Multimedia and Expo. IEEE Computer Society, 1087-1092.

[11] Rahman M A, Rashid M A, Ahmad M. Selecting the optimal conditions of Savitzky–Golay filter for fNIRS signal[J]. Biocybernetics and Biomedical Engineering, 2019, 39(3): 624-637.

[12] Sadeghi M, Behnia F, Amiri R. Window Selection of the Savitzky-Golay Filters for Signal Recovery from Noisy Measurements[J]. IEEE Transactions on Instrumentation and Measurement, 2020.

[13] Shi Z, Lin H, Liu L. Is CQT more suitable for monaural speech separation than STFT? an empirical study[J]. arXiv preprint arXiv:1902.00631, 2019.

[14] O’Hanlon K, Sandler M B. Comparing CQT and Reassignment Based Chroma Features for Template-based Automatic Chord Recognition[C]//ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019: 860-864.