AN IMPROVED ENTROPY-BASED ENDPOINT DETECTION ALGORITHM

Chuan JIA, Bo XU

National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences, Beijing
{cjia, xubo@nlpr.ia.ac.cn}

ABSTRACT

It is found that the detection using basic spectral entropy becomes difficult and inaccurate when speech signals are contaminated by high noise. This paper presents an improved entropy-based algorithm. The way to compute spectral probability density function of entropy is altered by the introduction of a positive constant. The modification improves the discriminability between speech and noise and the robustness of entropy so that it becomes easier to set thresholds. Experiment results reveal the validity of the improved entropy and prove that the improved entropy outperforms basic entropy. Moreover, the improvement of accurate rate (5db SNR) reaches 12.9% for the detection of start and end points averagely comparing with a pure energy-based algorithm.

1. INTRODUCTION

Endpoint detection, which aims at distinguishing the speech and non-speech segments from digital speech signal, is considered as a crucial part of the speech signal processing, such as automatic speech recognition. A good endpoint detector can improve the accuracy and speed of a speech recognition system. With the increasing deployment of speech recognition and voice-based systems across a wide range of voice-based services, it is desirable to develop a robust endpoint detector.

In the last several decades, a number of endpoint detection methods have been developed. We can categorize approximately these methods into two classes. One is based on thresholds [1-3]. Generally, this kind of method first extracts the acoustic features for each frame of signals and then compares these values of features with preset thresholds to classify each frame. The other is pattern-matching method [4,5] that needs estimate the model parameters of speech and noise signal. The detection process is similar to a recognition process. Compared with pattern-matching method, thresholds-based method does not need keep much training data and train models and is simpler and faster.

Endpoint detection by thresholds-based method is a typical classification problem. In order to achieve satisfied classification results, it is the most important to select appropriate features. Many experiments have proved that short-term energy and zero-crossing rate fail under low SNR conditions. It is desirable to find other robust features superior to short-term energy and zero-crossing rate. J. L. Shen [6] first used the entropy that is broadly used in the field of coding theory on endpoint detection. Entropy is a metric of uncertainty for random variables, thus it is definite that the entropy of speeches is different from that of noise signals because of the inherent characteristics of speech spectrums.

The algorithm [6] is based on weighted spectral entropy and experiment results proved that the algorithm outperforms energy-based algorithms in both detection accuracy and recognition performance under noisy environments. However, the weights cannot be obtained easily and accurately. L.S. Huang [7] combined basic spectral entropy and energy to solve the detection in babble and background music environments.

However, it is found that the basic spectral entropy of speech varies to different degrees when the spectrum of speech is contaminated by different noise signals especially high noise signals. The varieties make it difficult to determine the thresholds. Moreover, the basic spectral entropy of various noises disturbs the detection process. It is expected that there exists a way by which it is possible that (1) the entropy of various noise signals approaches to one another under the same SNR condition, (2) the curve of noise entropy is flat, and (3) the entropy of speech signals differs from that of noise signals obviously. Moreover, it is advantageous to include energy information besides distribution information. In a word, the improved method based on entropy should be simple, robust, reliable and accurate. This paper proposes an algorithm based on improved spectral entropy almost obtaining the target. The experiment results prove that the improved spectral entropy is superior to basic spectral entropy and the proposed algorithm outperforms energy-based algorithm.

The paper is organized as follows: Section 2 describes the definition of basic spectral entropy and its properties. Section 3 describes the improved spectral entropy and the proposed algorithm. The experiments are shown in section 4 and the conclusion is given in section 5.

2. BASIC SPECTRAL ENTROPY

2.1 Basic spectral entropy

According to the paper [6], the spectrum is first obtained for each frame by fast Fourier transform (FFT). Then the probability density function (pdf) for the spectrum can be calculated by normalizing every frequency component over all frequency components of one frame:

$$p_i = Y(f_i) / \sum_{k=0}^{N-1} Y(f_k) \ , \qquad i = 0 \ldots N\text{-}1 \quad (1)$$

where N is the total number of frequency components in FFT, $Y(f_i)$ is the spectral energy of the frequency component f_i, p_i is the corresponding probability density. Generally, we use a heuristic constraint to improve the discriminability of the pdf

between speech and non-speech signals. Since most of the energy of speech is in the region between 250Hz and 3750Hz, we use the constraint as follows:

$$Y(f_i) = 0, \quad \text{if } f_i<250\text{Hz or } f_i>3750\text{Hz} \qquad (2)$$

After applying the above constraint, the negative spectral entropy H_i of frame i can be calculated:

$$H = \sum_{k=0}^{N-1} p_k \log p_k \qquad (3)$$

2.2 The properties of spectral entropy

The validity of the basic spectral entropy as a feature used on endpoint detection is indicated by the following properties. On the other hand, it shows that the basic spectral entropy needs to be improved.

- The entropy of speech signals is different from that of most noise signals because of the intrinsic characteristics of speech spectrums and the different probability density distributions.
- Equation (1) is a normalizing process, and then spectral entropy is not influenced by the total energy in theory if the spectral distribution keeps unchanged. In practice, the distribution is changed by the actual pronunciation so that the entropy becomes different. However, the change of entropy is small compared with that of energy. For example, Fig. 1(a) shows a time-domain signal including two speech segments where the energy of the second segment is much low. It is difficult to detect accurately the end point of the second segment by energy-based algorithms. From Fig. 1(b), the negative entropy of the two segments is different at the second character "hai", but it is easier to detect the end of the second segment by the entropy compared with the energy. Moreover, even for the plosive and nasal consonants such as /f/, /c/, /t/, /s/, /sh/, there are considerable entropy values.

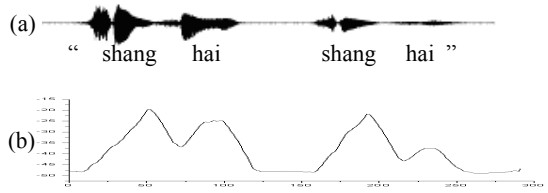

Fig. 1: (a). Waveform, (b) Negative entropy.

- The spectral entropy is robust to noise to some extent. For example, in Fig. 2, each line represents the entropy under different SNR conditions with white noise. With the drop of the SNR, the shape of negative entropy is almost kept. Nevertheless, the negative entropy decreases so that endpoint detection becomes more difficult if the SNR decreases.

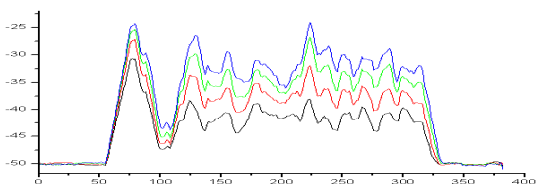

Fig. 2: Negative entropy under different SNR conditions with white noise.

3. THE PROPOSED ALGORITHM

Section 2 proves that spectral entropy can be used as a useful feature for endpoint detection and it is superior to energy. On the other hand, it implicitly shows that the entropy of speech is altered to become confused with the entropy of noise when the spectrum is contaminated by noises especially under the serious SNR conditions. At the same time, the spectral entropy contour of noise makes the detection more difficult. Thus the thresholds are hard to set and the endpoints of utterance are difficult to be identified.

In order to improve the robustness of spectral entropy against various noises, we consider to alter the original spectral probability density function of signals to make the entropy meet the three requirements described in section 1. In this section, we modify the computational form of entropy and propose a new algorithm to enhance the robustness and make the thresholds easy to be tuned so as to make algorithm more practical and accurate.

3.1 The improved feature

Rewrite Equation (1) into the following form by introducing a positive constant K:

$$p_i' = (Y(f_i)+K) / \sum_{k=0}^{N-1}(Y(f_k)+K) \quad i=0\ldots N\text{-}1, K>0 \qquad (4)$$

New "negative spectral entropy" is then obtained by taking Equation (4) into Equation (5).

$$H' = \sum_{k=0}^{N-1} p_k' \log p_k' \qquad (5)$$

After some simple derivations, the difference between old pdf p_i and new pdf pi' is as follows:

$$\Delta p_i = p_i' - p_i = \frac{1}{\frac{\sum_{k=0}^{N-1} Y(f_k)}{K} + N} \cdot (1 - N \cdot p_i) \qquad (6)$$

According to Equation (6), we can analysis the influence of the introduction of K to spectral entropy in two aspects.

Firstly, within one frame, the total energy $\sum_{k=0}^{N-1} Y(f_k)$ is definite and the difference Δp_i is determined by K and p_i.

- If $p_i \approx 1/N$, then $\Delta p_i \approx 0$. It means that the new probability density is close to the old probability density.
- If $p_i > 1/N$, then $\Delta p_i < 0$. The new probability density p_i' of the corresponding frequency component whose energy is greater than the average energy of spectrum components is lower than the old probability density p_i. Moreover, $|\Delta p_i|$ increases along with the increase of K and p_i.
- If $p_i < 1/N$, then $\Delta p_i > 0$. The new probability density p_i' of the corresponding frequency component whose energy is less than the average energy of spectrum components is higher than the old probability density p_i. Moreover, $|\Delta p_i|$ increases along with the increase of K and the decrease of p_i.

Form above analysis, we can deduce that the introduction of K into Equation (1) leads that the higher the old probability density is, the more it decreases, on the contrary, the lower the old probability density is, the more it increases. As a result, the probabilities in one frame tend to be equal. As we know, entropy increases along with the increase of uncertainty. Thus, the entropy of each frame increases (negative entropy decreases).

Secondly, for different frames, the total energies are different and Δp_i is determined by three terms: $\sum_{k=0}^{N-1} Y(f_k)$, K and p_i. If K and p_i are the same for the noisy speech and the noise signal, $|\Delta p_i|$ of the corresponding noisy speech spectrum component is smaller than that of noise signal because the energy of speech plus noise is commonly greater than that of noise signal. Thus, the increase of entropy (the decrease of negative entropy) of noise signals is possibly much more than noisy speech. Furthermore, the special spectrum character of speeches assures that the influence for speech signals is much different from that for noise signals.

In conclusion, negative entropy of both noisy speech and noise signal decreases. However, the decrease of negative entropy of noise is much more obvious than speech signal and entropy of various noises becomes close to one another, which makes the thresholds easy to be preset. Hence, the discriminability between speech signals and noise signals under noise environments is improved greatly.

3.2 The algorithm

In this section, an algorithm using the improved negative spectral entropy as a robust feature is presented below.

Step 1:Compute average frame energy *E_noise* of the first N_1 frames assumed as the background noise.

Step 2: Set K of Equation (4) according to the *E_noise*:
If $E\_noise < Th\_E_1$, set $K = K_0$.
If $Th\_E_1 \le E\_noise < Th\_E_2$, set $K = \alpha \cdot K_0$.
If $Th\_E_2 \le E\_noise < Th\_E_3$, set $K = \beta \cdot K_0$.
If $E\_noise \ge Th\_E_3$, set $K = \gamma \cdot K_0$.
where Th_E_1,Th_E_2 and Th_E_3 are preset thresholds, K_0 is an experience value.

Step 3:Compute the average negative entropy *Mean_NE* of the 20 frames before the current frame t.
If the negative entropy of the current frame $NE_t > Mean\_NE+V_1$ and $NE_{t+i} > Mean\_NE+V_2$ (i=1,2…N_2), continue to find the nearest peak and set it as the current frame. Otherwise repeat step 4.

Step 4:If negative entropy of the peak and the consecutive frame meets $NE_t > Th_1$ and $NE_{t+1} > Th_1$, go back to find a valley until its negative entropy $NE_{\text{valley}} < Th_2$ and regard the valley as the start point. Otherwise goto Step 3.

Step 5:After finding the start point, set $K = K_0$.

Step 6:Compute negative entropy until $NE_t < Th_3$. From this point go forward to find the nearest valley and its negative entropy is NE_{valley}.
Compute $Number_1$ = number of frames whose negative entropy $NE_{valley+i}$ meets $NE_{valley}-V_3 < NE_{valley+I} < NE_{valley}+V_3$ (i =1,2…N_3), $Number_2$ = number of frames whose negative entropy meets $NE_{valley} - V_4 < NE_{valley+i} < NE_{valley} + V_4$ (i=1,2…N_4).
If $Number_1 > Th_4$ and $Number_2 > Th_5$, regard the valley as the end point. Otherwise, repeat Step 6.

Step 7: Repeat from step 3 to step 6 until the end of the file.

3.3 Implementation issues

We set different K for different SNR in searching start points and set the same K in searching end points to attain the best results. Th_E_1, Th_E_2 and Th_E_3 are related to noise energy and frame length, α, β and γ used to make the entropy of noises under different SNR conditions close is related to noise energy and SNR.

These thresholds and parameters are easy to be determined by observation on the curves of the proposed features.

Using the constraints of Minimum Utterance Length and Minimum Pause Length, delete segments or combine segments into one segment.

4. EXPERIMENT RESULTS

The speech database used in the experiments here is 863 Chinese Mandarin Corpus. 5 data sets in the Database are used in our experiments. Every data set includes about 520 ~ 650 utterances and every utterance lasts 4 ~ 8 seconds. The noise signals used in the simulation include 4 kinds of noise (white, pink, F16 and Factory noises) of NOISEX-92 Database and office noise we collected. The office noise includes sounds of air-conditioners, knocking on keyboards and footfall. Especially, the sounds of knocking on keyboards are transitory but high-energy. The clean speech signals and various noise signals are mixed at 3 different signal-to-noise ratios (5db, 10db, 15db) to simulate the real noise environments. In our experiments, FFT is 1024 points and K_0 is the order of 10^8.

4.1 Feature comparison

Fig. 3 shows the basic and the proposed negative spectral entropy at 15db SNR with factory noise. In Fig. 3, the upper figure is waveform, the middle is the basic entropy contour, and the lowest is the proposed entropy contour. One can notice that the discriminability of the proposed feature is obviously better than the basic feature.

Fig. 4 includes four improved feature curves representing negative entropy under factory, office, white and pink noise backgrounds respectively and SNR=10db. From Fig. 4, it is found that the introduction of K almost makes the negative entropy of various noises approaches to -70 and the curves of noise become fairly flat under the same SNR condition. It is obvious that the thresholds are easy to be tuned consistently for different noise signals.

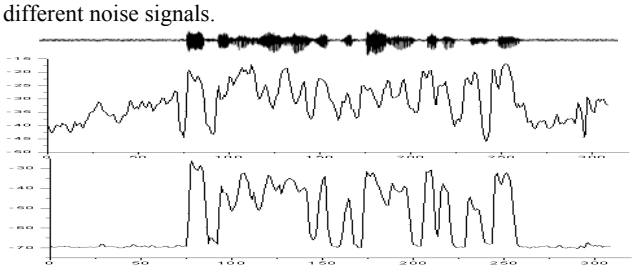

Fig. 3: Waveforms, the basic and improved negative spectral entropy (SNR=15db).

4.2 Endpoint detection experiments

In this experiment, the range of accurate start point is from 2000 points before the hand-labeling start point to 400 points after the hand-labeling start point. Similarly, the range of accurate end point is from 400 points before the hand-labeling end point to 2000 points after the hand-labeling end point. Our recognition experiments tell us that the ranges are rational and don't nearly influence the recognition accurate rate.

Table 1 doesn't include the result of method based on the basic entropy because according to our experience, the parameters are difficult to be tuned for various noises.

The detection accurate rates of start points and end points are showed in Table 1. It can be found that (1) the proposed algorithm is better than pure energy-based algorithm at medium SNR (15db) and significantly better at low SNR's (≤10db). (2) the performance of the proposed algorithm is distinctly superior to energy-based algorithm under non-stationary noise conditions. For example, energy-based algorithm fails under office noise including impulsive and high-energy sounds of keyboards-knocking. (3) the accurate rate of start point detection is better than that of end point detection. Because the energy of the end of utterance is weak, energy-based algorithm deteriorates at the detection of end points. However, the proposed algorithm is considerably stable.

5. CONCLUSION

Improved spectral entropy and a novel algorithm are proposed in this paper. The introduction of *K* into the process of calculating the probability density function of spectrum enhances the discriminability between speech signals and noise signals and improves the robustness of spectral entropy. Moreover, it becomes easier to determine thresholds than before. Experiments results prove that the improved feature can be successfully used in the real noisy environments and the performance is superior to energy-based endpoint detection. The improvement of accurate rate (5db) reaches 12.9% for the detection of start points and end points averagely comparing with a pure energy-based algorithm.

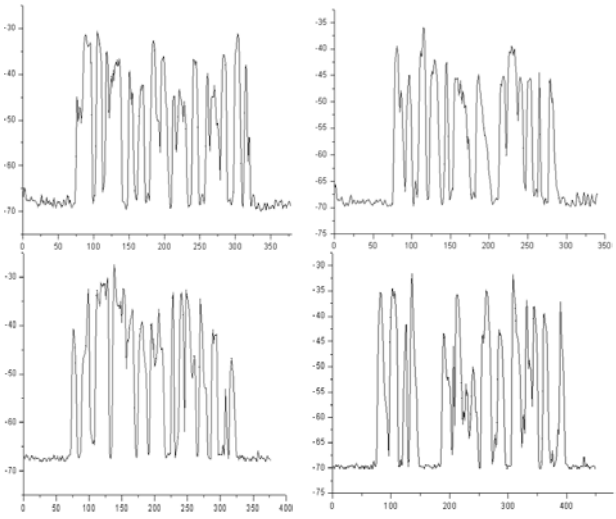

Fig. 4: The improved negative spectral entropy under 4 different noise environments (SNR=10db).

SNR / NOISE		15DB (%)		10DB (%)		5DB (%)	
		start	end	start	end	start	end
White	Energy	95.8	94.3	95.0	88.3	91.2	73.9
	Entropy	99.5	98.4	98.7	96.6	97.3	83.0
Pink	Energy	97.4	94.2	92.6	89.2	88.4	69.5
	Entropy	97.8	98.6	96.9	98.2	93.8	91.5
F16	Energy	93.4	96.2	91.1	86.9	79.5	63.9
	Entropy	96.7	99.0	95.0	93.6	95.0	84.6
Factory	Energy	90.1	89.4	79.2	68.0	69.2	56.3
	Entropy	97.1	90.4	92.1	80.1	76.4	78.2
Office	Energy	71.2	66.5	69.8	64.0	60.4	40.2
	Entropy	89.8	82.6	83.8	73.2	69.7	52.3
Average	Energy	89.6	88.1	85.5	79.3	77.7	60.8
	Entropy	96.2	93.8	93.3	88.3	86.4	77.9
Improvement		6.6	5.7	7.8	9.0	8.7	17.1

Table 1: Endpoint detection accurate rate

6. ACKNOWLEDGEMENT

The research work described in this paper was supported by the National Key Fundamental Research Program of China (the 973 Program) under the grant G19980300504 and the National Natural Science Foundation of China under the grant 69835003.

7. REFERENCES

[1] Woo-Ho Shin, Byoung-Soo Lee, Yun-Keun Lee, Jong-Seok Lee, "Speech/non-speech classification using multiple features for robust endpoint detection", *International Conference on Acoustics, Speech, and Signal Processing*, 2000.

[2] Stefaan Van Gerven, Fei Xie, "A Comparative study of speech detection methods", *European Conference on Speech, Communication and Techonlogy*,1997.

[3] Ramalingam Hariharan, Juha Häkkinen, Kari Laurila, "Robust end-of-utterance detection for real-time speech recognition applications", *International Conference on Acoustics, Speech, and Signal Processing*, 2001.

[4] A. Acero, C. Crespo, C. De la Torre, J. Torrecilla, "Robust HMM-based endpoint detector", *International Conference on Acoustics, Speech, and Signal Processing*, 1994.

[5] E. Kosmides, E. Dermatas, G. Kokkinakis, "Stochastic endpoint detection in noisy speech", *SPECOM Workshop*, 109-114, 1997.

[6] Jialin Shen, Jeihweih Hung, Linshan Lee, "Robust entropy-based endpoint detection for speech recognition in noisy environments", *International Conference on Spoken Language Processing*, Sydney, 1998.

[7] Liang-sheng Huang, Chung-ho Yang, "A novel approach to robust speech endpoint detection in car environments", *International Conference on Acoustics, Speech and Signal Processing*, 2000.