Robust Entropy-based Endpoint Detection for Speech Recognition in Noisy Environments

Jia-lin Shen, Jeih-weih Hung, Lin-shan Lee
Institute of Information Science, Academia Sinica
Taipei, Taiwan, Republic of China
jlshen@iis.sinica.edu.tw

ABSTRACT

This paper presents an entropy-based algorithm for accurate and robust endpoint detection for speech recognition under noisy environments. Instead of using the conventional energy-based features, the spectral entropy is developed to identify the speech segments accurately. Experimental results show that this algorithm outperforms the energy-based algorithms in both detection accuracy and recognition performance under noisy environments, with an average error rate reduction of more than 16%.

1. INTRODUCTION

Endpoint detection and verification of speech segments become relatively difficult in noisy environments, but are definitely important for robust speech recognition. The short-time energy or spectral energy has been conventionally used as the major feature parameters to distinguish the speech segments from other waveforms [1-4]. However, these features become less reliable and robust in noisy environments, especially in the presence of non-stationary noise and sound artifacts such as lip smacks, heavy breathing and mouth clicks etc. In this paper, a new algorithm for endpoint detection is proposed based on the entropy[5] in time-frequency domains, referred to as spectral entropy here. In this approach, the probability density function (pdf) of spectrum for each frame of speech signal is first estimated, on which the spectral entropy is defined and measured. It is found that this spectral entropy value is very useful in distinguishing the speech segments in a continuously recorded utterance from the non-speech parts, especially under sophisticated noisy environments. Several approaches are further developed to enhance the discriminability of the approach. The probability density function for speech spectra is also statistically estimated over a large set of speech data to be used as a weighting function for the spectral entropy. Experiments results indicated that the embedded speech segments can be very accurately extracted in utterances consisting of various types of serious background noise and sound artifacts, and the performance of speech recognition in such environments can be significantly improved as compared to the conventional energy-based algorithms.

2. ENTROPY-BASED ENDPOINT DETECTION

In the conventional endpoint detection algorithms, the short-time energy or spectral energy is usually used as the primary feature parameters with the augmentation of zero-crossing rate, pitch and duration information[1-4]. But these features become less reliable in the presence of non-stationary noise and various types of sound artifacts. Some other algorithms utilizing the noise adaptive thresholds have been proposed [1-4], but they become less helpful as well in the presence of sound artifacts and relatively high noise levels. This is why the entropy-based algorithm is proposed here to overcome the above problems. The spectrogram of a continuously recorded utterance is first derived. For each frame, the spectrum is obtained by fast Fourier transform (FFT). This FFT spectrum can be viewed as a vector of coefficients in the orthonormal basis. The probability density function (pdf) for the spectrum can thus be estimated by normalization over all frequency components :

$$p_i = s(f_i) / \sum_{k=1}^{N} s(f_k), \quad i = 1...N \qquad (1)$$

where $s(fi)$ is the spectral energy for the frequency

component *fi*, *pi* is the corresponding probability density, and N is the total number of frequency components in FFT. To enhance the discriminability of this pdf between speech and non-speech signals, several empirical constraints are further developed. First, only the frequency components between 250 Hz and 6000 Hz are considered, i.e.,

$$s(f_i) = 0, \quad \text{if } f_i < 250Hz \text{ or } f_i > 6000Hz. \quad (2)$$

This is because this region covers most of the frequency components of speech signals. Secondly, the upper and lower bounds of the probability densities are applied, i.e.,

$$p_i = 0, \quad \text{if } p_i < \delta_2 \text{ or } p_i > \delta_1. \quad (3)$$

where the lower bound δ_2 is used to cancel that noise with almost constant power spectral density values over all frequencies like white noise, while the upper bound δ_1 is used to eliminate the noise concentrating on some specific frequency bands. After the above normalization and enhancement processes, the corresponding spectral entropy for each frame is defined as [5]:

$$H = -\sum_{k=1}^{N} p_k \log p_k. \quad (4)$$

A set of weighting factors wk can be further applied to adjust the frequency component to the spectral entropy. These weighting factors are statistically estimated from a large collection of speech signals. Accordingly, the spectral entropy to be used for endpoint detection can be modified as the following :

$$H = -\sum_{k=1}^{N} w_k p_k \log p_k. \quad (5)$$

In the process of endpoint detection, the sum of the spectral entropy values over a duration of frames is first evaluated and smoothed by a median filter throughout the utterance. Some thresholds are then used to detect the beginning and ending boundaries of the embedded speech segments in a continuously recorded utterance [1][2]. A short period of background noise is first taken as the reference for some initial boundary detection process, and another set of thresholds derived from the analysis of speech signals are thus used for the refinement of the detected boundaries. Finally, some boundary pairs with the period of the corresponding speech segment less than a pre-defined minimum duration are rejected.

3. EXPERIMENT RESULTS

The speech database used in the experiments here contains the isolated digits in Mandarin Chinese produced by 100 speakers. Here 10 speakers among these 100 speakers were taken the testing speakers while the speech data produced by other speakers are used to train a speaker-independent model. 12 order mel-frequency cepstral coefficients and the corresponding 12 delta cepstral coefficients were used as the feature parameters. First order, left-to-right continuous density hidden Markov models(CHMM) were trained for each digit with 6 states for each model and 3 mixtures per state [6][7]. The recognition accuracy was evaluated by the average of the 10 testing speakers. A variety of types of noise were collected from the NOISEX-92 noise-in-speech database including white noise, pink noise, volvo noise (car noise), F16 noise and machinegun noise [8]. Furthermore, some sound artifacts were included such as breath noise, cough noise and mouse click noise. Fig. 1 (a) and (b) show the waveform and spectrogram of different types of noise and a speech segment, respectively. One can find that each type of noise has its special distribution on the spectrum, all of which are quite different from that of speech signal. Also, the signal-to-noise ratios (SNR) for various noisy environments can be very low such that the short-time energy and spectral energy as shown in Fig. 1 (c) become useless in detecting the boundaries for the speech segments. In addition, Fig. 1 (d) is the corresponding zero-crossing rate, which also becomes useless in the condition here. However, from Fig. 1 (e), it can be noted that the spectral entropy values in eq.(5) for the speech segments are still much higher than those different types of noise regardless of the very low energy levels for the speech signal. Here the summation of the spectral entropy values over a duration of 20 frames is first evaluated and then smoothed by a median filter to be used for endpoint detection, and the minimum duration of a speech segment is set to be 100ms. The distribution of the weighting factors used in eq.(5) is also plotted in Fig. 2, which indicates the concentration of speech energy on low frequency bands. The detected beginning and ending

boundaries of the speech segments using the entropy-based algorithm are also shown in Fig. 1 (e), which indicates very successful endpoint detection. In fact, two boundary pairs are obtained due to the pause within this speech segment as shown in Fig. 1 (a). Fig. 3 shows an example of the endpoint detection results based on entropy-based and energy-based algorithms with white noise at SNR of 5 dB. In the next experiment, the recognition results with the speech boundaries obtained by hand-labeling, energy-based and entropy-based algorithms are compared in the presence of different levels and types of noise. The error rates with respect to different levels of white noise, pink noise and car noise are shown in Table 1 (a). It can be found that the error rates obtained using entropy-based endpoint detection are only slightly higher or even identical to those obtained with hand-labeled endpoints in all the cases, while the error rates using the energy-based algorithm [1][2] are always higher. In fact, it can be found that the entropy-based algorithm is slightly better than the energy-based algorithm at medium SNR(15 dB), but significantly better at low SNR's(≤ 10 dB), with an average error rate reduction of more than 16%. Apparently, the recognition errors caused by inaccurate speech boundaries are increased as the SNR decreases. One can also notice that the degradation in recognition rates varies quite significantly for different types of noise. On the other hand, the average deviations of the detected boundaries using entropy-based and energy-based approaches with respect to the manually labeled results under different types and levels of noise are listed in Table 1 (b). It can be found from this table that the detected boundaries using entropy-based algorithm are much more close to the manual boundaries than those using energy-based algorithms, with deviation for the former being nearly half of the deviations for the latter. Accordingly, with the proposed approach, not only the speech boundaries in noisy environments can be accurately detected for improved speech recognition performance, but the undesired acoustic events can be separated and detected as well.

4. CONCLUSION

An entropy-based algorithm for accurate and robust speech endpoint detection is proposed in this paper. The spectral entropy is developed as the major feature to separate the speech segments from other signals. Experimental results show that not only the embedded speech segments can be successfully extracted from utterances containing a variety of background noise and sound artifacts, but also improved performance on speech recognition can be achieved in the presence of various types of noise.

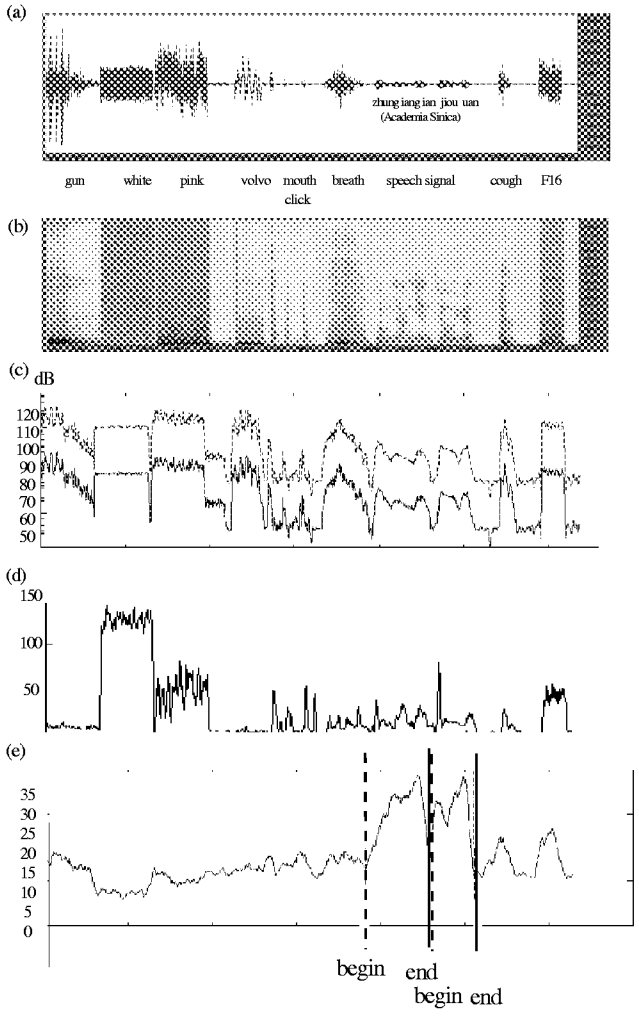

Figure. 1 : Various parameters for various types of noise and speech signal in a continuously recorded utterance : (a). waveform, (b). spectrogram, (c). short-time energy and spectral energy (d) zero-crossing rate and (e). spectral entropy (also showed the detected beginning and ending boundaries).

REFERENCES

1. J.C. Junqua, B. Mak, and B. Reaves, "A Robust Algorithm for Word Boundary Detection in the Presence of Noise", IEEE Trans. on Speech and Audio Processing, Vol. 2, No. 3, pp. 406-412, Apr. 1994.

2. L. Lamel, L. Labiner, A. Rosenberg, and J. Wilpon, "An Improved Endpoint Detector for Isolated Word Recognition", IEEE ASSP Magazine, Vol. 29, pp. 777-785, 1981.

3. M.H. Savoji, "A Robust Algorithm for Accurate Endpointing of Speech", Speech Communication, Vol. 8, pp. 45-60, 1989.

4. H. Ney, "An Optimisation Algorithm for Determining the Endpoints of Isolated Utterances", Proc. ICASSP, pp. 720-723, 1981.

5. S. Kullback, "Information Theory and Statistics", Wiley, New York, 1959.

6. L.R. Rabiner, "A Tutorial on Hidden Markov Models and selected Applications in Speech Recognition", Proc. IEEE, 77(2):257-286, Feb. 1989.

7. B.H. Juang, L.R. Rabiner, "The Segmental K-Means Algorithm for Estimating Parameters of Hidden Markov Models", IEEE Trans. on Audio Speech and Signal Processing, 38(9):1639~1641, Sep. 1990.

8. A.P. Varga, H.J.M Steeneken, M. Tomlinson, D. Jones, "The NOISEX-92 Study on the Effect of Additive Noise on Automatic Speech Recognition", In Technical Report, DRA Speech Research Unit, 1992.

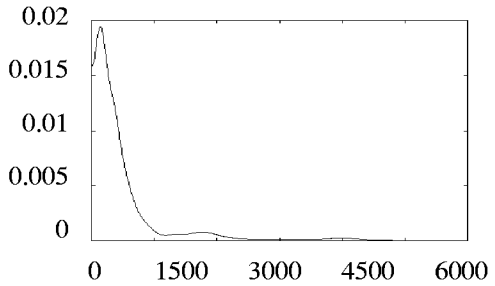

Figure 2 : The weighting factors used in the entropy evaluation.

(a)

Noise	Method	error rate (%)		
		5dB	10dB	15dB
white	energy	51	28	12
	entropy	43	24	11
	hand	36	20	11
pink	energy	40	14	0
	entropy	32	10	0
	hand	30	10	0
volvo (car)	energy	7	0	0
	entropy	2	0	0
	hand	0	0	0
average	energy	32.7	14	4
	entropy	25.7	11.3	3.7
	hand	22	10	3.7

(b)

Noise	Method	Derivations with manual boundaries (ms)		
		5dB	10dB	15dB
white	energy	32.80	23.36	13.95
	entropy	12.90	13.36	6.56
pink	energy	36.44	32.76	42.38
	entropy	16.80	18.84	19.06
volvo (car)	energy	29.00	23.50	19.06
	entropy	17.93	14.80	10.10
average	energy	32.75	26.54	25.13
	entropy	15.83	15.67	11.91

Table 1 : (a) The error rates for endpoints obtained by hand-labeling, entropy-based and energy-based algorithms with different types and levels of noise and (b) the corresponding deviations of speech boundaries with respect to hand-labeled endpoints.

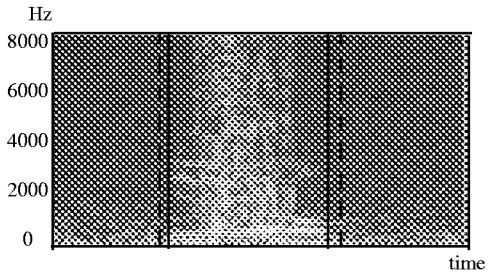

Figure 3 : An example of the word “six” with additive white noise at SNR of 5 dB. Solid and dashed lines denote the detected boundaries for energy-based and entropy-based algorithms, respectively.