ACOUSTICAL LETTER

A speech endpoint detector based on space-energy-entropy

Wang Xu[*], Qi Ding and Bing-xi Wang

Information and Engineering University,
No. 306, P.O. Box 1001, Zhengzhou 450002, Henan, P.R. China

(*Received 29 April 2003, Accepted for publication 12 June 2003*)

Keywords: Eigenvalue, Space-energy-entropy, Endpoint detection
PACS number: 43.60.Lq [DOI: 10.1250/ast.25.54]

1. Introduction

Precisely locating endpoints of the speech signal under noisy environments can remove the noisy segment and reduce the recognition error rate in speech recognition. In this paper, the probability density function of eigenvalues of embedded time-delayed space of speech signal is first estimated, on which space-energy-entropy is defined. Experimental results show that this entropy is very useful in distinguishing the speech segment from the non-speech parts. In this paper, a new algorithm for endpoint detection is proposed based on the space-energy-entropy.

2. Principal component analyses

The eigenspace of the noisy speech data can be divide into two subspaces [1]. One is the signal-plus-noise subspace and another is the noise subspace. The energy of the speech signal mainly focuses on the signal-plus-noise subspace. The energy of the random noise can be approximately considered to distribute uniformly in the noise subspace.

Consider one frame of speech signal $x(n)$ corrupted by an additive stationary background noise. We obtained the following M-dimensional vectors by embedding $x(n)$ in the space of the delayed coordinates:

$$\boldsymbol{y}(n) = [x(n), x(n-lag), \cdots, x(n-(M-1)\times lag)]^{\mathrm{T}} \quad (1)$$

where M is the embedding dimension, lag is the delay, $n = (M-1)\times lag, \cdots, N-1$, N is the frame size. We assume that the data set has zero mean and consider the following orthogonal transformation:

$$\boldsymbol{y}(n) = \sum_{m=1}^{M} s_m(n)\boldsymbol{A}_m \quad (2)$$

The goal of PCA is to find the orthogonal eigenvectors $\{\boldsymbol{A}_m, m = 1, \cdots, M\}$ associated with real non-negative eigenvalues $[\lambda_1, \cdots, \lambda_M]$ of the covariance matrix $C = \langle \boldsymbol{y}(n)\boldsymbol{y}(n)^{\mathrm{T}}\rangle$. The eigenvalues are ordered $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$. The principal components $s_m(n)$ can be found by projecting the data vectors onto each eigenvector:

$$s_m(n) = \boldsymbol{y}(n)^{\mathrm{T}}\boldsymbol{A}_m \quad m = 1, \cdots, M \quad (3)$$

The eigenvalues equal to the variances of the principal components and they also represent the energy's magnitude of each dimension of the eigenspace.

Figure 1 shows the waveform of the isolated Mandarin digits "Wu"and"Ba" embedded in additional white Gaussian noise. The SNR of the waveform is 5dB. The following parameters have been chosen: 1) frame size N=120 without overlap. 2) Hamming window. Then one frame of noisy speech signal is time-delayed embedded to get an $M\times(N-M+1)$ matrix $\boldsymbol{X}$:

$$\begin{bmatrix} x(0) & x(1) & \cdots & x(N-M) \\ x(1) & x(2) & \cdots & M \\ M & M & O & M \\ x(M-1) & x(M) & \cdots & x(N-1) \end{bmatrix} \quad (4)$$

Doing PCA on matrix $\boldsymbol{X}$, a group of eiginvalues $[\lambda_1, \cdots, \lambda_M]$ are determined.

Figure 2 shows the magnitude of eigenvalue of the different frames of the noisy speech signal in Fig. 1. We can find that each eigenvalue's magnitude is almost similar in the frame of noise. It implies that the noise's energy distributes uniformly in the each dimension of the time-delayed embedding space. But in the frame of the signal-plus-noise, the eigenvalue's magnitude of the first 10 dimensions is much larger than that of the last 5 dimensions. So the first 10 dimensions can be considered as the signal-plus-noise subspace. The last 5 dimensions are primarily occupied by noise and can be considered as the noise subspace.

3. Endpoint detection based on space-energy-entropy

The probability density function p_i is defined as:

$$p_i = \lambda_i \Bigg/ \sum_{i=1}^{M} \lambda_i \quad i = 1, \ldots, M \quad (5)$$

where λ_i is the eiginvalue of the i dimension of the time-delayed embedding space. p_i is the proportion of contribution of principal component s_i to speech signal and can be also considered as the probability of speech signal's energy concentrating on the i dimension. The corresponding space-energy-entropy H_k for the k frame is defined as:

$$H_k = -\sum_{i=1}^{M} p_i \log p_i \quad (6)$$

Figure 3 shows the space-energy-entropy of the waveform in Fig. 1. Because the probability distribution of the speech's space-energy is different from that of the noise's space-energy, voiced space-energy-entropy is quite different from non-voiced one. Based on this character, the endpoints can be properly pointed out by comparing space-energy-entropy with

[*]e-mail: xw7777@163.com

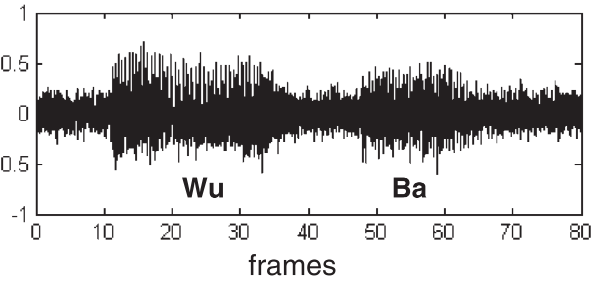


Fig. 1 Noisy signal (additive white Gaussian noise SNR=5 dB).

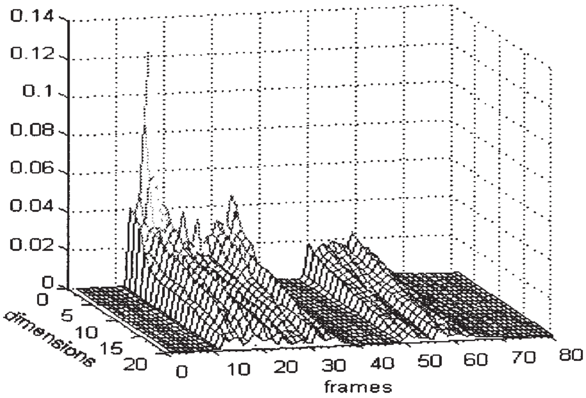


Fig. 2 Eigenvalues of time-delay embedding space.

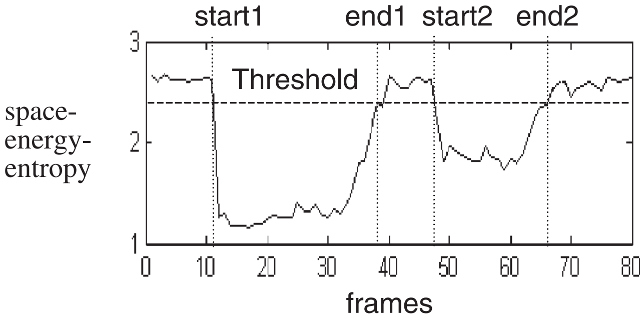


Fig. 3 Space-energy-entropy (additive white Gaussian noise also showed the detected beginning and ending boundaries).

the properly decision threshold.

Because the Eq. (5) computes the probability of speech energy concentrating on one dimension, it only considers energy-distribution within one frame but neglects the difference of energy among frames. It is obvious that voiced and non-voiced segment can not be discriminated by space-energy-entropy when the background noise i.e. a segment of musical has the same energy-distribution in time-delayed embedding space as the clean speech signal. The locations of clean speech signal in Fig. 1 are remained and the background noise is replaced by a segment of saxophone musical. Figure 4 shows the space-energy-entropy of this waveform. Then the speech segment can not be distinguished from the non-speech parts by the entropy. Considering the sum of speech plus noise is always greater than energy of noise, we can revise Eq. (6) by two ways. One is that via the equation $\sum_{i=1}^{M} \lambda_i = E_k$ Eq. (6) can be rewritten as the following:

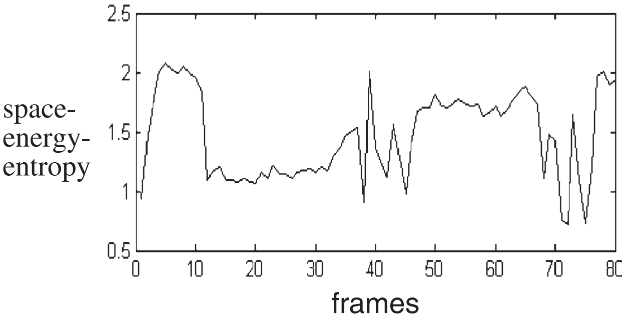


Fig. 4 Space-energy-entropy (additive saxophone muscial).

$$\begin{aligned} H_k &= -\sum_{i=1}^{M} \lambda_i/E_k \cdot \log(\lambda_i/E_k) \\ &= -E_k{}^{-1}\left(\sum_{i=1}^{M} \lambda_i \log\lambda_i - \sum_{i=1}^{M} \lambda_i \log E_k\right) \\ &= -E_k{}^{-1}\left(\sum_{i=1}^{M} \lambda_i \log\lambda_i - E_k \log E_k\right) \end{aligned} \tag{7}$$

$$\Leftrightarrow E_k \cdot H_k - E_k \log E_k = -\sum_{i=1}^{M} \lambda_i \log \lambda_i$$

$$\text{define :} \quad H_k{}' = -\sum_{i=1}^{M} \lambda_i \log \lambda_i = E_k \cdot H_k - E_k \log E_k \tag{8}$$

$H_k{}'$ can be viewed as space-energy-entropy H_k weighted with energy.

The second method is that Eq. (6) can be modified as:

$$H_k = -E_k{}' \times \sum_{i=1}^{M} p_i \log p_i \tag{9}$$

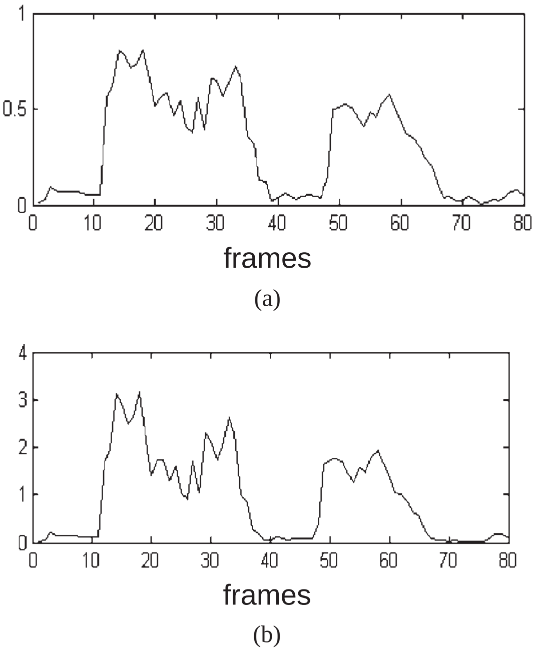
where $E_k{}' = \sum_{n=1}^{N} x^2(n)$ denotes the energy of the k frame noisy speech signal.

Figures 5(a) and 5(b) show the space-energy-entropy respectively computed by Eqs. (8) and (9).Their contours of waveform are almost the same. For decreasing the complexity of computation, the first method is chosen in this paper.

4. Experiment results and discussion

4.1. Performance evaluation criteria and speech database

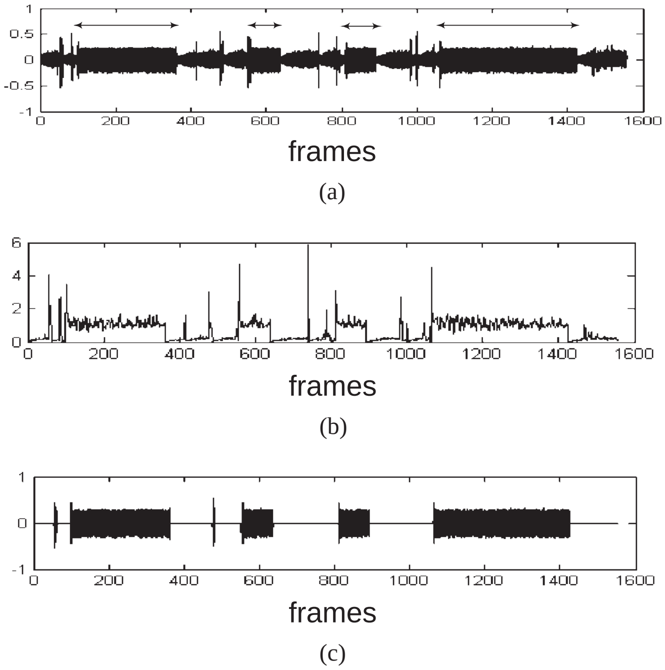
We use Weighted Accuracy [2] a evaluation parameter which gives greater weight to the effect of word clipping than to widening. The clean speech data used in the experiments are the set of isolated Mandarin digits from 19 speakers. The testing database has been created by adding different types of background noises from the Noisex Database to the clean speech data, at SNRs ranging from 5 dB to 15 dB. The time duration of each group of test data is 15 s. The sampling frequency is 8 kHz with 16 bits resolution. The frame size N=120 samples with non-overlapped Hamming windowing, the embedding dimension is 15.



(a)

(b)

Fig. 5 (a) Space-energy-entropy by Method 1 (additive axophone muscial). (b) Space-energy-entropy by Method (additive axophone muscial).



(a)

(b)

(c)

Fig. 6 (a) The actual pilot's voice in ultra shortwave channel. (b) Space-energy-entropy. (c) Detected pilot's voice.

Table 1 Average Weighted Accuracy of the test data by two methods.

Background noise		WGN		AIR		HWY	
Method		Space-energy-entropy	Spectral-entropy	Space-energy-entropy	Spectral-entropy	Space-energy-entropy	Spectral-entropy
Average clipping frames	5 dB	35.8	34.4	19.9	39.3	19.9	52.7
	10 dB	11.9	15.0	15.0	35.7	18.1	45.7
	15 dB	4.8	7.6	14.4	30.3	17.0	43.8
Average widening frames	5 dB	0.3	0.0	19.2	23.2	19.2	0.0
	10 dB	1.1	0.0	19.6	0.0	19.6	0.0
	15 dB	1.1	0.0	19.6	0.0	19.6	0.0
Average Weighted Accuracy	5 dB	50.3	48.2	39.7	69.0	39.7	75.7
	10 dB	17.3	21.0	31.8	49.9	33.2	63.9
	15 dB	7.3	10.6	30.9	42.5	31.6	61.2

The entropy must compare with some decision thresholds to discriminate voice and non-voice. In this paper, two thresholds are used. One is the voice-threshold and another is non-voice threshold. More detail can refer to the algorithm 2 in paper [3]. In addition, A fast algorithm based on discrete cosine transform is used to approximate Karhunen-Loeve transform for the eigen decomposition of a $N \times N$ covariance matrices,which reduces computation cost from $O(N^3)$ to N^2 [4].

4.2. Compared with spectral-entropy

Spectral-entropy is calculated by the probability density function for spectrum [5]. The energy and entropy are integrated for endpoint detection in a noisy in-car environment [6]. There is correspondence between space-energy-entropy and spectral-entropy. The former estimates the energy-distribution on space by orthogonal transform while the latter estimates the energy-distribution on spectral by Fourier transform. Table 1 shows the results of endpoint detection on testing database with space-energy-entropy and spectral-entropy [6].

It can be noted from Table 1 that the Average Weighted Accuracy of space-energy-entropy is as much as that of spectral-entropy under white Gaussian noise. When the background noise is aircraft cockpit noise or automobile highway noise, the results of space-energy-entropy are better than those of spectral-entropy and the average clipping frames of the space-energy-entropy are smaller than those of spectral-entropy., which is suitable fro automatic speech recognition.

4.3. Endpoint detection of actual noisy speech signal

Figure 6(a) is a segment of actual pilot's voice in ultra shortwave channel. The double arrowheads in Fig. 6(a) point to the pilot's voice and the others are the noise of the ultra shortwave channel. Figure 6(b) shows the space-energy-entropy of the waveform of Fig. 6(a). Figure 6(b) shows the detected voice. It is obvious that using the space-energy-entropy can discriminate the voice and non-voice effectively.

5. Conclusion

A space-energy-entropy-based algorithm for accurate and robust speech endpoint detection is proposed in this paper. Experimental results show that the embedded speech segments can be successfully extracted from utterance containing a variety of background noise.

References

[1] Y. Ephrahim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Process.*, **3**, 251–266 (1995).

[2] F. Beritelli, "A robust endpoint detector based on differential parameters and fuzzy pattern recognition," *Proc. ICSP '98*, pp. 747–751 (1998).

[3] S. Van Gerven and F. Xie, "A comparative studyof speech detection methods," *Eur. Conf. Speech Communication and Technology* (1997).

[4] J. Huang and Y. Zhao, "A DCT-based fast signal subspace technique for robust speech recognition," *IEEE Trans. Speech Audio Process.*, **8**, 747–751 (2000).

[5] J.-L. Shen, J.-W. Hung and L.-S. Lee, "Robust entropy-based endpoint detection for speech recognition in noisy environments," *ICSP '98*, pp. 232–235 (1998).

[6] L.-S. Huang and C.-H. Yang, "A novel approach to robust speech endpoint detection in car environments," *ICASSP 2000*, vol. 3, pp. 1751–1754 (2000).