Proceedings of the 2005 IEEE
Engineering in Medicine and Biology 27th Annual Conference
Shanghai, China, September 1-4, 2005

Speech Endpoint Detection Based on Speech Time-Frequency Enhancement and Spectral Entropy

Fan Yingle, Li Yi, and Wu Chuanyan, *Member, IEEE*

***Abstract*—In the process of speech recognition, it is especially crucial to precisely locate endpoints of the input utterance to be free of non-speech regions. This paper proposes a novel approach that finds robust features for endpoint detection in a noisy environment. In this proposed method, we integrate both time-frequency enhancement and the spectral entropy feature. Firstly, the noisy speech is enhanced using spectral subtraction method, in frequency domain to remove the additive noises. Then in time domain, a weight function built by short-time energy and zero-crossing rate is used to remove the noise produced by the spectral subtraction. Finally spectra entropy -based method is used to detect the endpoints. By monitoring the transition of the extracted feature, more precise endpoints could be found. The proposed algorithm is shown to be well suited for the detection of speech endpoint and is very robust for different types of noise, especially for low SNR. Furthermore, the algorithm has a low complexity and is suitable for real-time DSP system.**

I. Introduction

THE detection of the endpoints of an utterance is required in many speech applications. Accurate endpoint detection is crucial for good speech recognition accuracy.

The most popular existing detection method is the simple energy detector which performs adequately for clean speech. Problems arise in noisy environments for low energy phonemes (some fricatives and plosives, for example) at the endpoints. A major source of error in isolated word speech recognition systems is the inaccurate detection of the beginning and ending boundaries of test and training patterns. The performance of existing endpoint detection severely degrades in noisy environments [1].

The types of errors for energy-based detectors introduced by poor SNR include [2]: (1) Missing the leading or trailing low-energy sounds such as fricatives; (2) Classifying clicks, pops and other background noise as part of speech due to their high energy content;(3)Falsely classifying background noise as speech while missing the actual speech. This is particularly true when the background noise consists of speech from other speakers, such as in babble noise.

Spectral entropy a metric of uncertainty for random variables, then the spectral entropy of speeches is different from that of noise signals because of the inherent characteristics of speech spectrums [3].So the spectral entropy is adopted as the basic feature for endpoint detection in this paper.

Conventional endpoint detection methods pay much attention to the features, and neglect the enhancing speech before exacting the feature. In result, the process can't obtain accurate detection, especially for low SNR. In this paper, time-frequency enhancement and spectral entropy endpoint detection algorithm is proposed. In frequency domain, spectral subtraction is used to remove the additive noises. In time domain, a weighting function is used to remove the residual noise brought by the spectral subtraction. After time-frequency enhancement, spectral entropy endpoint detection is used to locate the endpoints.

II. Time-Frequency Enhancement

For low SNR, the spectral entropies of the noisy speech and the noises are almost the same [4].It is important to enhance the speech before endpoint detection. There are many methods for speech enhancement. Spectral subtraction [5] is the better one. But it isn't perfect, while the spectral subtraction can remove the additive noises; it also brings the residual noises such as music noise. In order to remove the residual noises, a kind of filters in time-domain is introduced in this paper.

A. Speech enhancement in frequency domain

The theory of spectral subtraction is shown as following:

$$|\hat{S}_m(f)|^2 = \begin{cases} |s_m(f)|^2 - \gamma |D_m(f)|^2 \\ |s_m(f)^2| \geq |D_m(f)|^2 \\ 0.015 \times |s_m(f)|^2 \\ |s_m(f)|^2 \leq |D_m(f)|^2 \end{cases} \quad (1)$$

In formula (1) γ is equal to 1 or 2. $|\hat{S}_m(f)|$ is the energy spectrum after enhancement. $|D_m(f)|$ is the estimation of the energy spectrum of the background noises. It can be obtained by the following methods. Firstly, apply one of the

Manuscript received May 2, 2005. This work was supported by National Nature Science Foundation of China (60302027). the Zhejiang Province Nature Science Foundation of China(602127)

Fan Yingle is with Department of Instrument Science and Technology, Hangzhou Dianzi University, Hangzhou 310018, China (e-mail: Fan@hziee.edu.cn).

Li Yi was with Department of Instrument Science and Technology, Hangzhou Dianzi University, Hangzhou 310018, China (phone: 86-571-88986156; e-mail: liyi7699 @163.com); Li Yi was also an Doctor Graduate Student in Department of Biomedical Engineering, Zhejiang University

Wu Chuanyan is with Department of Instrument Science and Technology, Hangzhou Dianzi University, Hangzhou 310018, China (e-mail: Wuchuanyan@163.com).

conventional endpoint detection methods to separate the background noises from the noisy speech. Then compute the average energy spectrum of the background noises.

B. Speech enhancement in time domain

In order to remove the residual noises [6] brought by the spectral subtraction, a weight function is proposed. The weight function can be constructed by the original speech short-time energy and zero-crossing rate. The algorithm is presented as followed.

Step1: Compute the short-time energy and zero-crossing rate for each frame is defined as formula (2)and (3) [7].

$$E(m)=\sum_{n=0}^{N-1}s_w^2(n) \tag{2}$$

$$Z(m)=\frac{1}{2}\left\{\sum_{n=0}^{N-1}\left|\operatorname{sgn}\left[s_w(n)\right]-\operatorname{sgn}\left[s_w(n-1)\right]\right|\right\} \tag{3}$$

Where sgn(x) is the sign function.

$$\operatorname{sgn}(x)=\begin{cases}1 & \text{if} \quad x>0\\ 0 & \text{if} \quad x=0\\ -1 & \text{if} \quad x<0\end{cases} \tag{4}$$

Step 2: Since short-time energy and zero-crossing rate can detect the speech endpoints, a weighting function can be defined as formula (5). Then compute the function values for each frame.

$$f(m)=\log(E(m)/Z(m)) \tag{5}$$

Where $f(m)$ is the weighting function of the mth frame, $E(m)$ the short-time energy and $Z(m)$ the zero-crossing rate of the mth frame.

Step 3: Assume that x(n) is the final speech signal removed noises, then

$$x(n)=\begin{cases}k_1\hat{S}(n) & f(m)<\lambda_1\\ k_2\hat{S}(n) & \lambda_1\le f(m)<\lambda_2\\ k_3\hat{S}(n) & f(m)>\lambda_2\end{cases} \tag{6}$$

Where k_1,k_2k_3 is the weight coefficients, and the value is 0.45, 1.1, 0.8 by experience. λ_1,λ_2 is the thresholds, The initial value can be set randomly, then they can be adjusted by formula (7).

$$\begin{aligned}\lambda_1&=0.3*\max(f(m))+0.7*\min(f(m))\\ \lambda_2&=0.8*\max(f(m))+0.2*\min(f(m))\end{aligned} \tag{7}$$

In conclusion, computing the weight function to adjust the thresholds λ_1,λ_2 dynamically and then to smoothen the residual noises brought by spectral subtraction.

III. Spectral Entropy Base Endpoint Detection

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:

$$p_i=s(f_i)/\sum_{k=1}^{N}s(f_k),i=1\cdots N \tag{8}$$

Where $s(f_i)$ is the spectral energy for the frequency component f_i, and p_i 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 250Hz and 6000 Hz are considered, i.e.,

$$s(f_i)=0,\ \text{if}\ f_i<250Hz\ \text{or}\ f_i>6000Hz \tag{9}$$

The most frequency components of speech signals is covered. Secondly, the probability densities are defined as followed, i.e.,

$$p_i=0,\ \text{if}\ p_i<\delta_2\ \text{or}\ p_i>\delta_1 \tag{10}$$

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

$$H=-\sum_{k=1}^{N}p_k\log p_k \tag{11}$$

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

$$H=-\sum_{k=1}^{N}w_kp_k\log p_k \tag{12}$$

In the process of endpoint detection, the sum of the spectral entropy values over duration of frames is first evaluated and smoothened 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 continuous utterance. 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 predefined minimum duration are rejected.

IV. Experiment Results

The proposed algorithm was evaluated on clean speech database collected with 16 KHz sampling rate, 16 bits A/D in a normal environment (without background noise) The speech signal is processed by hamming window with 256 length ,and FFT transform with 512 length. The clean speech database contains 50 people, each person speaking from'0' to '9' once. The noise database is NOISEX92. To simulate speech production in noisy conditions, Each clean speech add noise to obtain noisy speech. Several levels of signal-to-noise ratio (SNR) have been considered, ranging from clean speech (with no additive noise) to 0 dB SNR. We chose the representative stable noise white and F-16 cockpit noise, unstable noise, factory noise and office noise.

Here are two possible ways to evaluate the correctness of an endpoint detection algorithm: one is to compare the detected results to hand labeled ones, and the other is to pass the detected words through a speech recognizer and compare the recognition rates. Here, we choose the first option for the most straightforward comparison. Comparing the algorithm performance with that of the TFE endpoint detector is shown in Table I and Table II.

TABLE I
RECOGNITION ACCURACY OBTAINED FOR ENTROPY ALGORITHMS AND TFE IN STABLE NOISE CASES.

	White noise				F-16 cockpit noise			
SNR	Entropy		TFE		Entropy		TFE	
	S	E	S	E	S	E	S	E
15DB	95.8	94.3	99.5	98.4	93.4	96.2	96.7	99.0
10DB	95.0	88.3	98.7	96.6	91.1	86.9	95.0	93.6
5 DB	91.2	73.9	97.3	83.0	79.5	63.9	95.0	86.6
0 DB	70.9	63.5	83.6	72.1	60.5	42.8	80.7	60.9

TABLE II
RECOGNITION ACCURACY OBTAINED FOR ENTROPY ALGORITHMS AND TFE IN UNSTABLE NOISE CASES.

	Office noise				Factory noise			
SNR	Entropy		TFE		Entropy		TFE	
	S	E	S	E	S	E	S	E
15DB	71.2	66.5	89.8	82.6	90.1	89.4	97.1	90.4
10DB	69.8	64.0	83.8	73.2	79.2	68.0	92.1	80.1
5 DB	60.4	40.2	69.7	52.3	69.2	56.3	76.4	78.2
0 DB	42.7	35.6	62.9	49.2	58.1	43.2	70.2	60.5

V. Conclusion

Endpoint detection and verification of speech segments become relatively difficult in noisy environment, 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. 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 in time-frequency domains, referred to as spectral entropy here.

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 at low SNR. Furthermore, the algorithm has a low complexity and is suitable for real-time DSP system.

References

[1] Lingyun Gu and Stephen A. Zahorian, "A New Robust Algorithm for Isolated Endpoint Detection," In Proc. lEEE ICASSP - 02, Vol. 4, pp. 4161 -pp. 4164, 2002

[2] Sahar E. Bou-Ghazale and Khaled Assaleh, "A Robust Endpoint Detection of Speech for Noisy Environments with Application to Automatic Speech Recognition." In Proc. IEEE ICASSP - 02, Vol. 4, pp. 3808 - pp. 3811, 2002

[3] JIA Chuan, XU Bo. "An Improved Entropy-Based Endpoint Detection Algorithm ," ISCSLP, pp.96, 2002.

[4] HU Hang, Speech Signal Processing, Harbin Polytechnic University Publishing House, Harbin, 2000.

[5] XU Yi-fang, ZHANG Jin-jie, YAO Kai-sheng etc, "Speech enhancement applied to speech recognition in noisy environments," Tsinghua Univ (Sci & Tech), Beijing. Pp.41-44, 2001.

[6] YAO Kai-sheng, Bertram E. shi, etc, "Residual noise compensation for robust speech recognition nonstationary noise," Proc. ICASSP, (2):pp.1125-1128, 2000.

[7] XU Da-wei, WU Bian and ZHAO Jian-wei, etc, "A Study on Noisy Speech Recognition (Linear Predictive Coding Prediction Error)," Comuputer Engineering and Applications, Vol.39 No.1, pp.115-117, 2003.

[8] CHEN Shang-qin, LUO Cheng-lie and YANG Xue, " Modern Speech Recognition " Electronic publishing house of University of Science and Technology,Chengdu.pp.7-19,1991.