International Journal of Information Technology and Knowledge Management
July-December 2010, Volume 3, No. 2, pp. 695-698

COMPARATIVE STUDY OF SPEECH RECOGNITION SYSTEM USING VARIOUS FEATURE EXTRACTION TECHNIQUES

Kapil Sharma[1], H.P.Sinha[1] & R.K. Aggarwal[2]

It is very important to detect the speech endpoints accurately in speech recognition. This paper presents a comparative analysis of various feature extraction techniques of endpoint detection in speech recognition of isolated words in noisy environments. The endpoint detection problem is nontrivial for no stationary backgrounds where artifacts (i.e., no speech events) may be introduced by the speaker, the recording environment, and the transmission system. An optimum set of characteristics is identified by combining parameters from both time domain and frequency domain, in a robust approach for identification when the speech signal is corrupted by additive noise and channel distortion. The cases of colored noises such as babble noise, factory noise at different SNR values in conjunction with distortions due to recording medium were tested. Experimental results identify the optimal algorithm which significantly achieves the highest performance in the recognition task.

Keywords: OCRZ, DAS, Robust Speech Recognition, Endpoint Detection, VFR, Noise Analysis.

1. Introduction

Endpoint detection is very important problem in many speech-processing systems. The systems that process a word as a unit have to locate its beginning and end. The problem of detecting (locating) the endpoints would seem to be easily solvable for a human, but it has been found to be a very complex and challenging task, in many cases, for a machine. In some situations it is not so difficult to determine the position of the endpoints – e.g. in case the signal-to-noise ratio level is high enough The mismatch of environmental conditions between testing data and actual data obtained from real time recording has a severe effect on the performance of speech recognition systems. The performance degradation is due to background disturbances in the form of additive noise and channel distortion effects. Many algorithms have been proposed to extract features which are robust to this degradation [1]. Addition of silence portions due to human hesitation to utter the words is also responsible for distortions to the signal. The solution to this problem has been implemented in the form of end point detection of the boundaries of actual speech part in the signal [2-5].

This paper attempts to compare the responses of two different algorithms for end point detection and then draw conclusions as to which set of parameters are most effective for the task at hand. The first algorithm (ZCR-EN) is based on a combination of two time domain features called zero crossing rate (ZCR) and short time energy (EN) of the signal. The second algorithm (VFR – Variable Frequency Rate) iteratively uses three parameters, ZCR, EN and Euclidean distance in spectral domain to determine the boundaries.

In the past, algorithms have also been developed to combat the noise degradation problem [6-9]. In this paper we analyze the performances of a group of techniques that are inherently robust to noise. These include autocorrelation based features – RAS (Relative Autocorrelation Sequence) and CHRAS (Channel Relative Autocorrelation Sequence) and differential autocorrelation based technique (DAS) [10].

The experiments were conducted by adding noises in the form of additive coloured noise and multiplicative channel distortion [10]. It is a time domain parameter that is calculated over a framed signal. ZCR[13] is defined as the number of times that a signal changes signs in a particular frame. Speech generally has a higher zero crossing rate, since it is composed of alternating voiced and unvoiced sounds in the syllable rate. Similarly, silence portion has a much lower ZCR. This feature can be exploited to determine the approximate point at which silence changes to speech.

The rest of the paper is organized as follows. The analysis of the disturbances used for the experiments is presented in Section II. The mathematical fundamentals of the two end point detection algorithms are described in Section III. In Section IV, a series of experiments on the recognition process is conducted to analyze the recognition rate of the algorithms. Ultimately, a conclusion is given in Section V.

[1] Department of Electronics and Communication Engineering, M.M Engineering College, Maharishi Markandeshwar University, Mullana (Ambala)

[2] Asst. Prof., Department of Computer Science and Engineering, Institute of National Importance (NIT), Kurukshetra

Email: [2]rai1_kapil@rediffmail.com

2. Analysis of the Disturbances

Additive noise is additive to the speech signal in both, the power spectrum domain and the autocorrelation domain. Channel disturbance on the other hand is multiplicative in nature in the autocorrelation domain. Thus the technique to remove additive noise is to subtract in either the power spectrum domain or the autocorrelation domain. The technique to remove channel distortion however, involves a more complicated set of steps. The multiplicative nature of the distortion can be converted into an additive nature by shifting the speech from the autocorrelation domain to the logarithmic domain. Here, a filtering technique is applied in which mean subtraction is used to remove the channel effect.

The noises used in the experiments can be broadly classified as channel distortion and additive noise The additive noise is coloured in nature. The coloured noises that have been considered are babble noise, factory noise and F-16 noise. The channel distortion is in the form of a random sequence of numbers emulating a Gaussian Channel. All these noises have been extracted from the NATO RSG – 10 databases.

3. Analysis of Endpoint Detection Algorithms

The speech signal x is taken and a preemphasis filter is used to perform the task of enhancing the dominant parts of the signal. The emphasized speech signal is the divided into frames and the framed signal is then used as an input to the end point detection algorithm. This section now presents the mathematical analysis of each of these algorithms.

A. Endpoint Detection using ZCR-EN

This algorithm uses two time domain parameters to decide the boundary between silence voice components. ZCR is zero crossing rate which is defined as the number of times that a signal changes signs in a particular frame and can be calculated using (1).

$$\frac{1}{2\cdot N}\cdot\sum_{m}|\operatorname{sgn}(x(m))-\operatorname{sgn}(x(m-1))|$$

$$1\le n\le N:1\le m\le M \qquad (1)$$

where M is the number of samples per frame and N is the total number of frames in the signal. Short time energy is defined as the sum of the squares of the magnitudes of the samples taken per frame. It can be calculated using (2).

$$\mathrm{E_n} = \frac{1}{N}\cdot\sum_{m}x(m)^2$$

$$1\le n\le N:1\le m\le M \qquad (2)$$

The algorithm to find the endpoints using these to parameters has been presented below:

First, threshold values on the basis of which a certain frame is accepted or rejected are calculated. The zero crossing rate threshold OCRZ is determined using (3).

$$\mathrm{OCRZ} = \frac{1}{N}\Sigma_{i=1}^{N}x_i\times 2\sqrt{\frac{1}{n-1}\Sigma_{i=1}^{N}(x_i-\overline{x})^2} \qquad (3)$$

The upper (TUL) and lower (TLL) thresholds of energy are calculated using (4).

$$\mathrm{TLL} = \min(0.03\times(\mathrm{IMX}-\mathrm{IMN})+\mathrm{IMN},\ 4\times\mathrm{IMN})$$

$$\mathrm{TUL} = 5\times\mathrm{ITL} \qquad (4)$$

where, IMN and IMX are the minimum and maximum energy levels found in the signal.

Searching is started from the beginning of the framed signal until the energy crosses TUL. Then the search is backed off towards the signal beginning until the first point at which the energy falls below TLL is reached. This is marked as the provisional beginning point - N1. N2 (the end point) is evaluated in a similar way. Again, the signal is searched from the beginning and ZCR is examined. If this measure exceeds the OCRZ threshold 3 or more times, N1 is moved to the first point at which the threshold is exceeded. N1 is defined as the formal beginning point. Formal endpoint N2 using OCRZ is evaluated in a similar manner.

B. Endpoint Detection using VFR

Variable frame rate (VFR) is a technique used for discarding frames that are too much alike. The method emphasizes the transient regions, which are more relevant for speech recognition.

The algorithm consists of three steps. At first, the speech signal corresponding to a single word is preprocessed and the background noise is estimated which is used to decide the threshold values for the following steps. In the second step, the starting-point and the ending-point of the voiced sound are located to be used as reference endpoints based on time domain features of short time energy and zero-crossing rate. And finally, the accurate endpoints of the utterance are located according to the frequency parameter called mel-frequency cepstrum of the sequence of speech signals between the reference endpoints.

$$E_b = \begin{cases}\dfrac{E_{k-1}+E_K}{2} & if\ 0.5\le\dfrac{E_{k-1}}{E_k}\le 2\\ \min(E_{k-1},E_k); & \text{otherwise}\end{cases} \qquad (5)$$

For determining background noise ZCR, (6) is used.

$$Z_N = \begin{cases}\dfrac{Z_f+Z_b}{2} & \text{if } 0.5\le\dfrac{Z_f}{Z_b}\le 2\\ \text{rejected}; & \text{otherwise}\end{cases} \qquad (6)$$

$$Z_f = \begin{cases} \dfrac{Z_1 + Z_2}{2}; & \text{if } 0.5 \le \dfrac{Z_1}{Z_2} \le 2 \\ \min(Z_1, Z_2); & \text{otherwise} \end{cases}$$

$$Z_b = \begin{cases} \dfrac{Z_{k-1} + Z_k}{2}; & \text{if } 0.5 \le \dfrac{Z_{k-1}}{Z_k} \le 2 \\ \min(Z_{k-1}, Z_k); & \text{otherwise} \end{cases} \qquad (7)$$

Energy threshold T_E and ZCR threshold T_Z are calculated using the background energy and ZCR levels of E_N and Z_N. The energy function is searched and the first frame whose energy is above T_E, is assumed to be the starting point as in (8).

$$P_{F2} = k\{E_k > T_E; \qquad k = 1,2,.....K\} \qquad (8)$$

where E_k is defined by (2). The energy function is then searched backwards from right to left, the ending-point of the voiced sound is obtained by:

$$P_{B2} = k\{E_k > T_E; \qquad k = 1,2,.....K\} \qquad (9)$$

The zero-crossing parameter is then used to relax the endpoints. The zero-crossing function is searched from point P_{F3} backwards to obtain

$$P_{F2} = k\{Z_k > T_Z; \qquad k = P_{F2}, P_{F2-1},.....1\} \qquad (10)$$

where Z_K is defined in (1). The zero-crossing function is searched from point P_{B3} forwards, to obtain

$$P_{B2} = k\{Z_k > T_Z; \qquad k = P_{B2}, P_{B2+1},.....K\} \qquad (11)$$

D(i, j), the Euclidean distance between the current frame i, and the last retained frame j is evaluated using mel frequency cepstrum coefficients of the signal. Euclidean threshold T_D is experimentally derived to be -5.8. The Euclidean function is searched forward from P_{F2}

$$P_{F1} = k\{D(k, k + 1) > T_D \;\&\&\; D(k, k + 2) > T_D \;\&\&\; D\,(k, k + 3) > T_D\} \qquad (12)$$

The Euclidean function is the searched from P_{B2} backwards

$$P_{B1} = k\{D(k, k - 1) > T_D \;\&\&\; D\,(k, k - 2) > T_D \;\&\&\; D(k, k - 3) > T_D\} \qquad (13)$$

The points finally obtained are the actual endpoints.

4. Experiments

The speech data was collected from different speakers. Since the effects of background noise and channel distortion are minimized, the speech in this database is referred to as the clean speech.

The speech signal was sampled at a 16 kHz sampling rate and weighed by a Hamming window equal to 256 samples, shifted every 128 samples. In computing the MFCC, a 20 channel filter bank with mel scale frequency is applied.

A. Recognition Rates for Endpoint Detection Algorithms

The system was trained using a clean enrollment of the speech signals. A test database of signals for same set of words as in the clean database is formed by recording in real time. The signals were recorded for 5 seconds. The testing speech is polluted by additive noise at different noise decibel levels. The performance of the four endpoint detection algorithms has been plotted in Figure 1(a-c). Three varieties of additive noises in the form of babble noise, factory noise and F-16 noise have been used for the test. Table 1(a-c) shows the actual accuracy rates. The additive noises have been taken at different noise level of 20dB, 15dB, 10dB, 5dB and 0dB SNR.

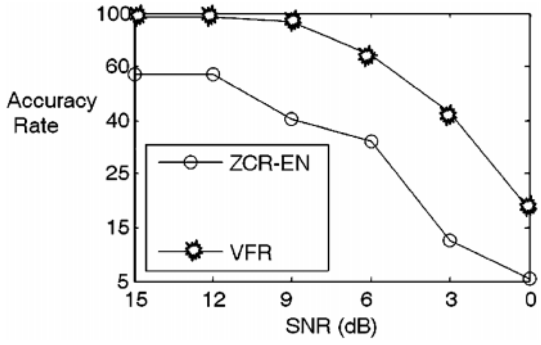

Figure 1(a): Babble Noise

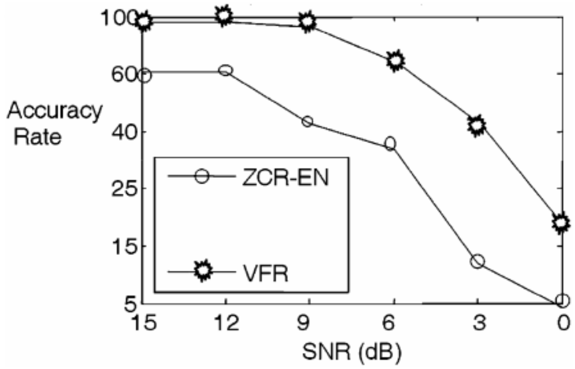

Figure 1(b): Factory Noise

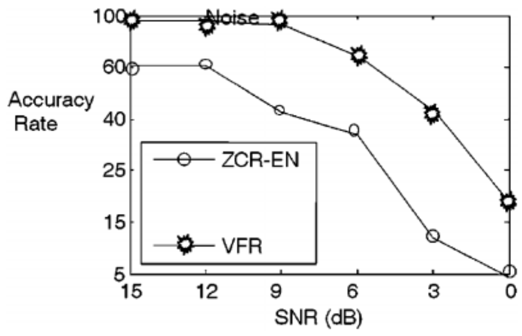

Figure 1(c): F-16 Noise

All the figures have been plotted with noise levels (in SNR) along the x – axis and the accuracy rates of the algorithms in terms of percentages along the y – axis.

Table 1(a)
Endpoint Detection (Babble Noise)

	Clean	20dB	15dB	10dB	5dB	0dB
ZCR_EN	77.3	77.3	60.6	52.3	15.0	1.6
VFR	98.6	98.6	97.0	84.3	62.6	26.6

Table 1(b)
Endpoint Detection (Factory Noise)

	Clean	20dB	15dB	10dB	5dB	0dB
ZCR_EN	77.3	77.3	64.6	43.6	17.3	1.6
VFR	98.6	98.6	97.0	82.6	63.6	29.3

Table 1(c)
Endpoint Detection (F – 16 Noise)

	Clean	20dB	15dB	10dB	5dB	0Db
ZCR_EN	77.3	77.3	63.6	47.0	16.6	1.3
VFR	98.6	98.6	97.0	82.0	69.6	14.0

5. CONCLUSION

This paper proposes a comparative analysis about robust noisy speech recognition method based on the various algorithms at different noise levels for different types of noises, we have come to draw the following conclusions regarding the performance of the algorithms chosen by us to conduct this experiment. . Inaccurate endpoint detection can cause misclassification rather than other possible mistakes. Accuracy of end point detection is much higher for algorithms which integrate both time domain and frequency domain features. Zero crossing rate detection is efficient only at low noise levels. Increase in background disturbances reduces its efficiency.VFR algorithm yields the highest accuracy rate. This is due to the fact that it uses time domain features to evaluate a preliminary set of end points and then makes the boundaries more rigorous by incorporating frequency domain concept of Euclidean distance.

REFERENCES

[1] Y. Gong, "Speech Recognition in Noisy Environments: xxxxA Survey," Speech Commun., 16, pp. 261 – 291, xxxx Apr. 1995.

[2] W. H. Shin, B. S. Lee, Y. K. Lee and J. S. Lee, "Speech / Non-Speech Classification Using Multiple Features for Robust Endpoint Detection", Information Technology Lab., LG Corporate Institute of Technology.

[3] Y. Zhang, X. Zhu, Y. Hao and Y. Luo, "A Robust and Fast Endpoint Detection Algorithm for Isolated Word Recognition", IEEE International Conference on Intelligent Processing Systems, October 1997.

[4] C. Guanghua, L. Junhai and Y. Jun, "An Improved Method of Endpoints Detection Based on Energy Frequency Value", Micro Electronic Research and Development Centre, Shanghai.

[5] Yingle, L. Yi and W. Chuanyan, "Speech Endpoint Detection Based on Time-Frequency Enhancement and Spectral Entropy", Proc. 27th Annual IEEE Conference, Shanghai, September 2005.

[6] M. Afify, Y. Gong, and J. P. Haton, "A General Joint Additive and Convolutive Bias Compensation Approach Applied to Noisy Lombard Speech Recognition," IEEE Trans. Speech Audio Process., 6, No. 6, pp. – 524 – 538, Nov. 1998.

[7] Y. Zhao, "Channel Identification and Signal Spectrum Estimation for Robust Automatic Speech Recognition," IEEE Signal Process. Lett., 5, No. 12, pp. 305 – 308, Dec. 1998.

[8] M.J.F. Gales and S.J. Young, "Robustm Speech Recognition in Additive and Convolutional Noise using Parallel Model Combination," Comput. Speech Lang., vol. pp. 289 – 307, 1995.

[9] A. Sankar and C.H. Lee, "A Maximum – likelihood Approach to Stochastic Matching for Robust Speech Recognition," IEEE Trans. Speech Audio Process., 4, pp. 190 – 202, May 1996.

[10] K.H. Yuo, T.H. Hwang, and H.C. Wang, "Combination of Autocorrelation – based Features and Projection Measure Technique for Speaker Identification," IEEE Trans. Speech Audio Process., 13, No. 4, pp. 565 – 574, Jul. 2005.

[11] G. Farahani, Mohammad Ahadi, and Mohammad Mehdi Homayounpour, "Autocorrelation based Methods for Noise Robust Speech Recognition", Amirkabir University of Technology, Iran.