Speech Recognition System Design Using Short-time Energy and Zero Crossing Rate

Xianchao Wang
School of Automation Science and Engineering, Xi'an Jiaotong University

Abstract—**When processing speech signal, short-time analysis in time domain is important and efficient. This paper presents the author's design of a speech recognition system using short-time energy and zero crossing rate. In this system, speech signal of the number 0 to 9 in Chinese can be distinguished effectively. Based on short-time analysis, these two characteristics can help detect the end point of the voiced part in the speech signal. Besides, since every speech signal has its unique short-time energy and ZCR, this can be used to correctly identify the numbers. Methods in machine learning like cross-validation and k-nn algorithm are also used to make the system function better. This system is based on programs written in Matlab. Experiments have been carried out to verify theoretical studies and figure out the most suitable parameters for this system. The system designed by the author is particularly for identifying characters in a small-scale character set. In that case, sophisticated algorithms and systems are unnecessary.**

Keywords—Short-term energy; zero crossing rate; end point detection; cross-validation; k-nearest neighbors algorithm.

I. INTRODUCTION

Speech is an indispensable means of communication used by people since ancient times. In the field of automatic speech recognition, or ASR, meaningful reasons like technological curiosity, pursuit of more intelligent human-machine interaction, and desire for convenience brought by advanced technology inspire people from generation to generation to do researches and inventions. Back in 1773, people already succeeded in producing vowel sounds using specific tools [1]. From speech production models to spectral representations, the speech pioneers later have done quantities of researches, attempting to equip the machine with the ability to speak naturally and respond properly. Genuine automatic speech recognition starts in 1952 in Bell Laboratories, where a digit recognizer was born [2]. After years of development, IBM came up with the first commercial system Via Voice in 1990s. In recent years, as automatic speech recognition becomes mature, it has been successfully applied in work place and people's daily life.

According to previous researches, using two characteristics of the speech signals, i.e., short-time energy and zero crossing rate proves effective in classifying voiced and unvoiced parts in the signals [3-6]. However, lacking proper models and algorithms, these researches stop there. In this study, the author focuses on further exploiting the two features, applying them to the design of a system for identifying characters in a small-scale character set. In general, speech signal is non-stationary, but it remains nearly unvaried during short time [7]. Based on this, analyzing speech signals on time domain is feasible. Since the scale of the character set in this system is small, calculating and comparing the above-mentioned signal characteristics is enough. In the character set, different characters have their unique acoustic features. The system makes good use of this point and can distinguish one character from another effectively. Constructed by the programs written in Matlab, the recognition system can recognize numbers 0 to 9, which functions well when dealing with the collected speech signal. Moreover, to find out the best parameters for this system, i.e., the window function type, the segment numbers and the weight in k-nn algorithm, experiments have been carried out and results are compared.

There is no denying that ASR is of great applied value and has a promising future. While it's still far from inventing a real machine that speaks fluently and understands words correctly like what humans do, many important scientific and technological advances have taken place, bringing us closer to the goal.

II. BASICS OF THEORY OF THE SYSTEM

A. *Short-time Energy*

In most cases, the energy of the signal is defined as [8]

$$\mathrm{E} = \sum_{m=-\infty}^{\infty} x^2(m)\,. \tag{1}$$

After separating the speech signal into small segments, its energy in the n-th frame is defined as

$$\mathrm{E}_n = \sum_{m=n}^{n+N-1} x^2(m)\,. \tag{2}$$

It is a function measuring the change in magnitude of the speech signal [9]. However, since E_n uses the square of the signal value while calculating, it is sensitive to signal which is in a logic high state. The difference is prominent when dealing with the small sample values and the big ones.

B. *Zero Crossing Rate*

Zero crossing rate, namely ZCR, measures the number of zero crossings in the speech signal. For continuous time signals, zero crossing refers to the waveform in time domain crossing the time axis, while for discrete time signals, we define zero crossing as neighboring sample values have different signs [10]. As a result, ZCR counts the changing times of the signal sign. To some extent, it embodies the characteristic of signal spectrum, as high-frequency components are likely to have a high ZCR [11].

ZCR of the n-th speech signal, namely Z_n, is given by [12]

$$Z_n = \frac{1}{2}\sum_{m=n}^{n+N-1}\left|\operatorname{sgn}\left[x_n(m)\right]-\operatorname{sgn}\left[x_n(m-1)\right]\right|, \quad (3)$$

where

$$\operatorname{sgn}\left[\mathrm{x(n)}\right] = \begin{cases} -1 & x(n) < 0 \\ 1 & x(n) \geq 0 \end{cases}. \quad (4)$$

C. Frame

In general, speech is a non-stationary signal, the characteristics of which are time variant. Therefore, digital signal processing methods dealing with stationary signals is not applicable. However, different voices result from certain shape of vocal tract formed by the movement of mouth. Compared with the frequency of speech signal, such movement is rather slow. On the other hand, although the speech signal is time variant, it remains nearly unvaried during a short time [7]. These segments are called as frame [13].

D. Window

To separate the overall speech signal into frames, moveable finite-length window which is zero-valued despite the chosen interval, is utilized for weighting [14]. Mathematically, when multiplying the speech signal with a window function, the result we get is also zero-valued outside the interval, and the left overlapped part is often different from the formal signal to make the characteristics easier to be seen, like viewing a single part of speech signal through a window. This process can be described as

$$s_w\left(n\right) = s(n) \times w(n). \quad (5)$$

When using window function for signal processing, the shape and length of the window greatly affects the analysis of the short-term parameters. The short-term parameters got from a proper window can represent the characteristics of the speech signal better. In this research, rectangular, Hann and Hamming windows are tested and the corresponding results compared. The rectangular window can be expressed as

$$w(n) = \begin{cases} 1 & 0 \leq n \leq N-1 \\ 0 & else \end{cases}. \quad (6)$$

The Hann window can be expressed as

$$w(n) = \begin{cases} 0.54 - 0.46\cos\dfrac{2\pi n}{N-1} & 0 \leq n \leq N-1 \\ 0 & else \end{cases}. \quad (7)$$

The Hamming window can be expressed as

$$w(n) = \begin{cases} 0.5(1-\cos\dfrac{2\pi n}{N-1}) & 0 \leq n \leq N-1 \\ 0 & else \end{cases}. \quad (8)$$

III. Construction of Recogntion System

A. Collection of Speech Signal

The function "Audiorecorder" in Matlab is used to collect the speech signals. The sample rate, sample bits and channel number here is 44100Hz, 16 and 1 accordingly. The signals are then saved in the form of "wav" file. The waveform of the collected signal is shown in the sequence of 0 to 9 in Fig. 1.

B. Processing of Speech Signal

1) Preprocessing

The speech signal is inevitably affected by factors like voice volume and background noise, thus necessitating the preprocessing, which includes normalization and dc-removal. Normalization is meant to eliminate the error caused by different volumes, while dc-removal helps focus on the useful part of the signal. The difference of the waveform of the number 0 before and after preprocessing is illustrated in Fig. 2.

2) Windowing

As mentioned in the former part, the processing should be based on short-time analysis. Windowing is a useful tool for segmentation. Here are the waveforms using different window functions in Fig. 3.

3) End point detection

From the previous figures, it is easy to find that only part of the speech signal is voiced, the remaining of which is unvoiced and unnecessary. To acquire the voiced part, short-time energy and ZCR can be used. Use the speech signal of number 0 under Hamming window as an example. According to the results, there appears to be great difference of two characteristics between the voiced and unvoiced part. To be exact, the voiced part has much higher short-term energy and low ZCR, while the unvoiced part of the speech signal has low short-term energy but much higher ZCR. Interestingly, in Fig. 4, zero crossing rate witnesses a sharp decrease near the end of the curve. Compared to the original waveform, it can be seen that this represents background noise, i.e., small fluctuations in the corresponding position of the speech signal. Actually, after checking the result frame by frame, there is a small deviation between the end point got by using the short-time energy or ZCR alone, which is mainly less than three frames. To make the result as accurate as possible, here the author uses the threshold detection. Firstly, set a high threshold to the short-term energy which most of the frames can exceed. Then set a relatively low threshold to figure out the end point. Finally, use another threshold for zero crossing rate to check if this point is the right one and make corrections accordingly. The results are shown in Fig. 5.

4) Segmentation

To recognize the correct number, the direct way is to compare the two characteristics of the signal under test with the data set frame by frame. However, due to the fact that each

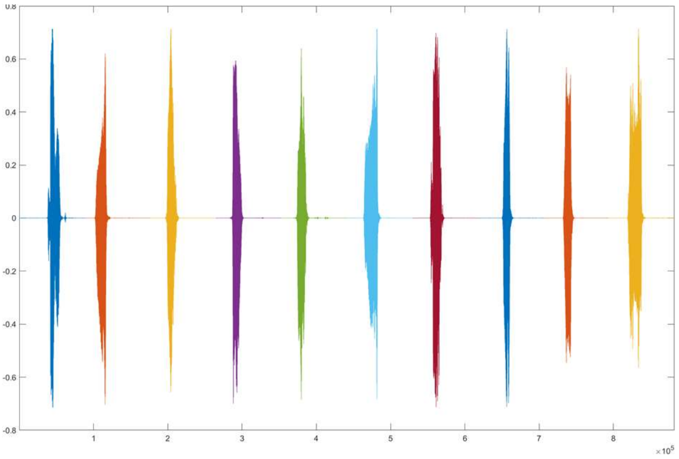

Fig. 1. Waveform of the original speech signal of numbers 0 to 9.

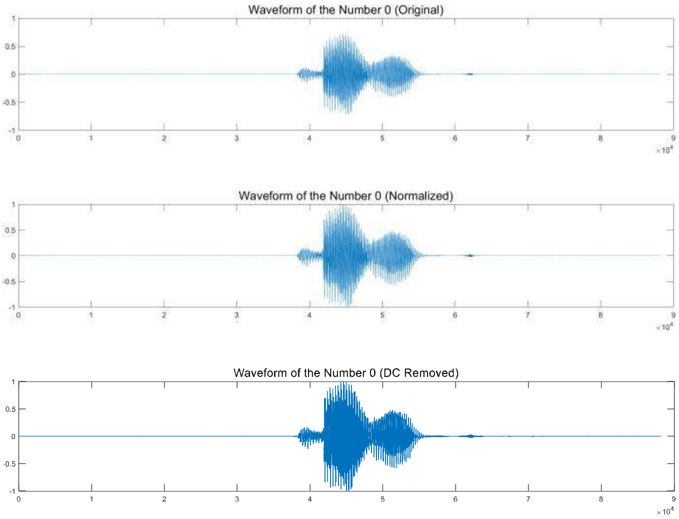

Fig. 2. Waveform of the number 0 before and after preprocessing.

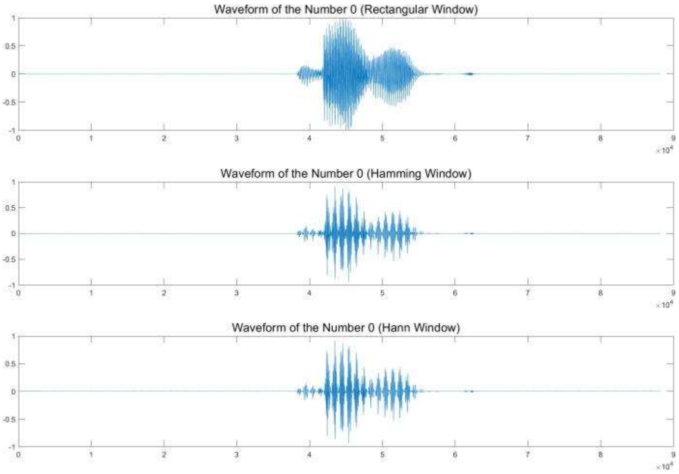

Fig. 3. Waveform of the number 0 under different window functions.

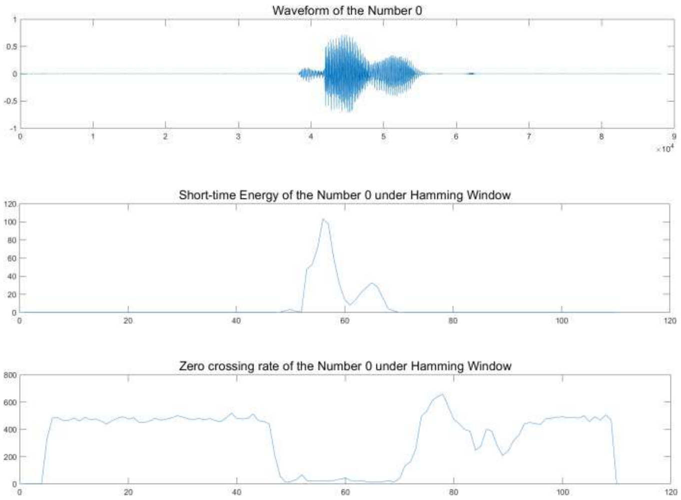

Fig. 4. Short-time energy and ZCR of the number 0.

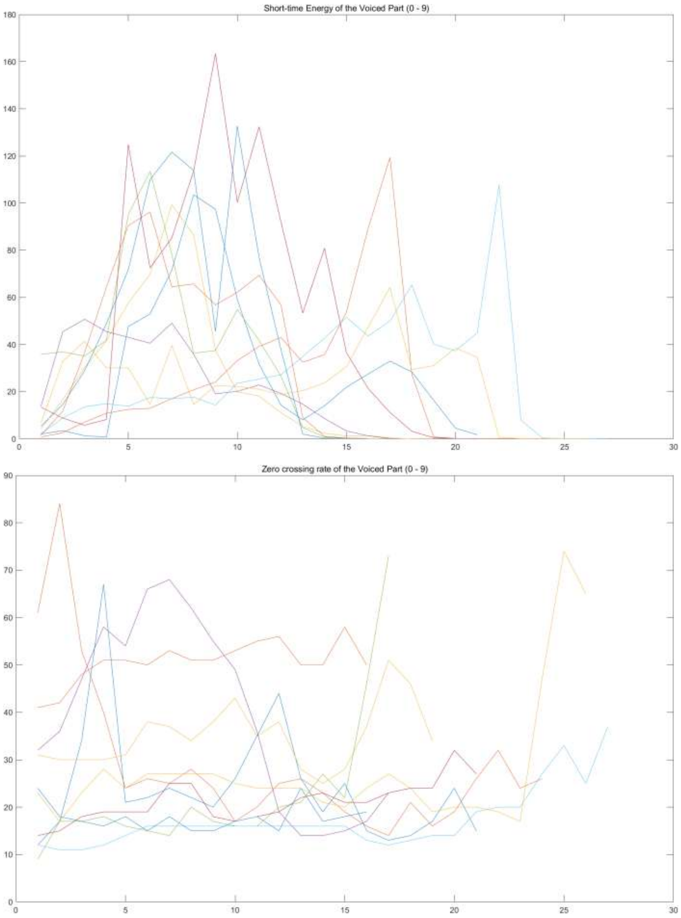

Fig. 5. Short-time energy and ZCR of the voiced part (number 0-9).

speech signal has different lengths of voiced part, it is hard to achieve that. This paper gives a method of solving this problem: for each number, separate their voiced part into several segments, and calculate the mean value of the two characteristics within a segment. The number of segments is fixed, and in this case the signals can be compared. Here gives the charts of the two characteristics when the segment number equals 4 in Fig. 6.

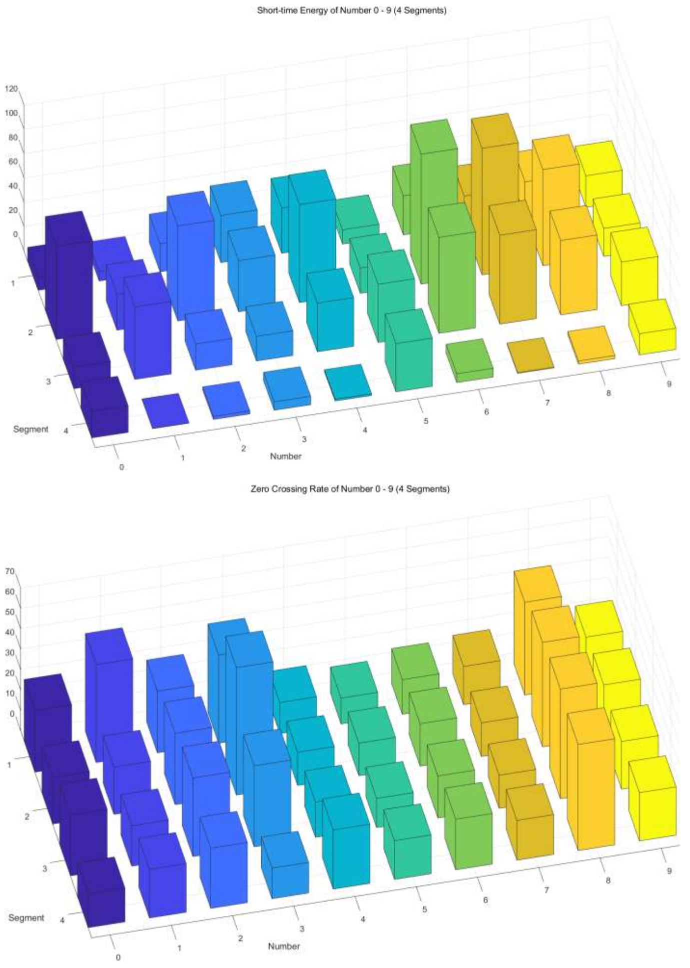

Fig. 6. Short-time energy and ZCR of number 0-9 (voiced part divided into 4 segments).

C. *Test of Speech Signal*

1) *Cross-validation*

Because the number of speech samples is limited, cross-validation is used to guarantee enough test times. In the experiment, the author collects 100 speech samples of his own, namely 10 samples for each number. In cross-validation, the 100 samples are divided into two groups: the test signals and the data set. The test signals are random samples of each number and contains 10 samples in all, while the remaining 90 samples function as the data set. The test time can be any time because in each test, the division is random. The author set the test times to 10000 for each number in the following experiments.

2) *K-nn Algorithm*

Both short-time energy and ZCR has to be calculated and compared. And an important question is the method of quantification. This experiment applies K-nn algorithm for the classifier. The final result in the data set is the number which has the shortest "distance" with the test signal ("distance" means the summation of the absolute deviation of the two characteristics). In this algorithm, the weight of the two characteristics in the summation may influence the final result, which is talked about in the next section.

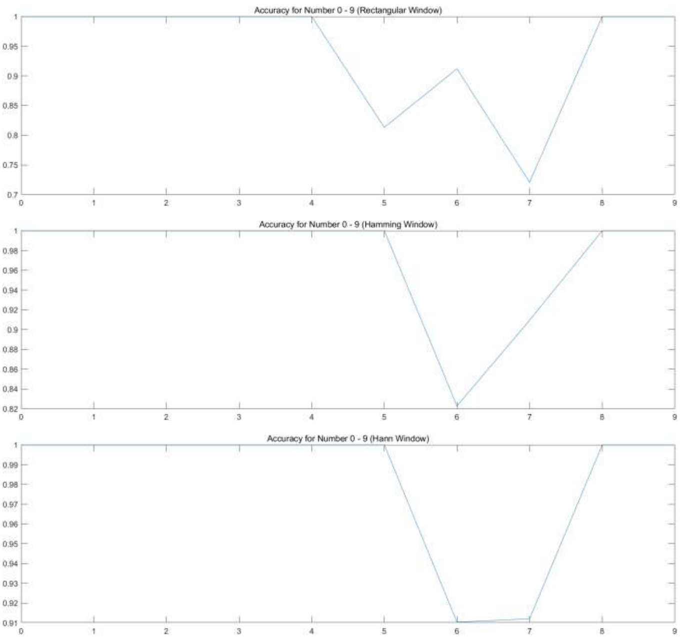

Fig. 7. Overall accuracy under different window functions.

IV. System Performance under Different Parameters

In the recognition system, several parameters may affect the final result. In this paper, the type of window function, the segment numbers and the weight in k-nn algorithm are discussed using control variable to find out best parameters for the system.

A. *Window Function*

Fig. 7 shows the accuracy for all the numbers under different window functions. All the three windows tested perform well, among which the Hann Window is the best choice.

B. *Segment Number*

Fig. 8 shows the overall accuracy under different segment numbers. From the result, we can see that the accuracy is low when the segment number is small, which is because the division is rough and cannot reflect the trend of the signal. Circumstances where segment number is around 10 almost function perfectly, and the accuracy reaches the peak at 14 segments. The accuracy is low again when the segment number is over 15. The author speculates that this might be due to the fact the pronunciations of some numbers in Chinese is hasty, causing the sequence length to be small and even less then the segment number. According to the designed program, this may lead to the recognition result being bound to be incorrect.

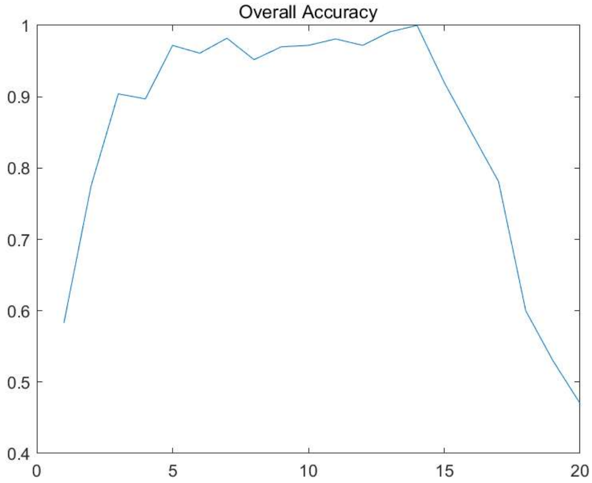

Fig. 8. Overall accuracy under different segment numbers.

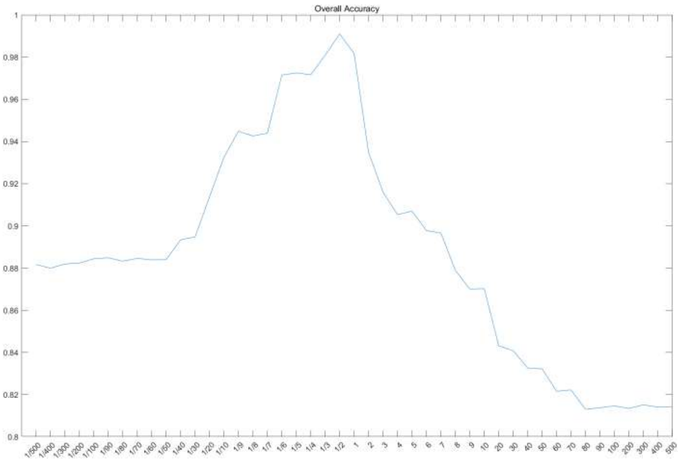

Fig. 9. Overall accuracy under different weight.

C. *Weight in k-nn Algorithm*

Fig. 9 shows the overall accuracy under different weights. The x-axis represents the proportion of the weight of short-time energy to the weight of ZCR. The best proportion for the system is 0.5. However, the reasons behind the distribution of accuracy shown in the graph is unknown, which can be further researched.

V. CONCLUSION

In this paper, speech recognition based on time-domain analysis is studied. Using the characteristics of speech signals, namely, short-term energy and ZCR, a recognition system is designed that is particularly suitable for recognizing small character sets. Admittedly, there remain some errors and unknow reasons in the research. For example, the result only comes from cross-validation. The speech samples are not enough to confirm such system also works well in a real situation where different speakers have different accents, tones, speeds, and so on. Further researches can focus on those drawbacks and improve the system.

REFERENCES

[1] Erdol N., Castelluccia C., Zilouchian A. (1993) Recovery of missing speech packets using the short-time energy and zero-crossing measurements. IEEE Transactions on Speech & Audio Processing, 1(3): 295-303.

[2] Rojathai S., Venkatesulu M. (2016) Tamil Speech Word Recognition System with Aid of ANFIS and Dynamic Time Warping (DTW). Journal of Computational and Theoretical Nanoscience, 20: 116.

[3] Jalil, M., Butt, F. A., & Malik, A. (2013, May). Short-time energy, magnitude, zero crossing rate and autocorrelation measurement for discriminating voiced and unvoiced segments of speech signals. In 2013 The international conference on technological advances in electrical, electronics and computer engineering (TAEECE) (pp. 208-212). IEEE.

[4] Bachu, R. G., Kopparthi, S., Adapa, B., & Barkana, B. D. (2010). Voiced/unvoiced decision for speech signals based on zero-crossing rate and energy. In Advanced techniques in computing sciences and software engineering (pp. 279-282). Springer Netherlands.

[5] Shete, D. S., Patil, S. B., & Patil, S. (2014). Zero crossing rate and Energy of the Speech Signal of Devanagari Script. IOSR-JVSP, 4(1), 1-5.

[6] Zaw, T. H., & War, N. (2017, December). The combination of spectral entropy, zero crossing rate, short time energy and linear prediction error for voice activity detection. In 2017 20th International Conference of Computer and Information Technology (ICCIT) (pp. 1-5). IEEE.

[7] Kpuska V Z., Eljhani M M., Hight B H. (2013) Front-end of Wake-Up-Word Speech Recognition System Design on FPGA. Journal of Telecommunications System & Management, 2(108): 203-208.

[8] Mcdermott E., Nakamura A. (2006) Large-scale continuous speech recognition system design using discriminative training. The Journal of the Acoustical Society of America, 120(5): 3042.

[9] Miyanaga Y. (1999) Autonomous-Agent Speech Recognition System and its VLSI System Design. Ieice Technical Report Speech, 99: 55-60.

[10] Wang S R., Huang S., Yuan F. (2011) Design and Implementation of Speech Recognition System Based on SPCE061A. Advanced Materials Research, 187: 389-393.

[11] Xing Y L., Chen L., Zhang X W. (2003) The Design and Implementation of Speech Recognition Intelligent Telegraphy System. Journal of Military Communications Technology, 20: 103-109.

[12] Pinto D., Arnau J M. (2020) Design and Evaluation of an Ultra Low-power Human-quality Speech Recognition System. ACM Transactions on Architecture and Code Optimization, 17(4): 1-19.

[13] Nilakhe A., Shelke S. (2016) A design for wireless music control system using speech recognition. 2016 Conference on Advances in Signal Processing (CASP). IEEE.

[14] Saraswathi S., Geetha T V. (2010) Design of language models at various phases of Tamil speech recognition system. International Journal of Engineering Science & Technology, 2(5): 230-239.