International Journal of Advanced Engineering and Management Research

Vol. 4, No. 06; 2019

ISSN: 2456-3676

ENDPOINT DETECTION BASED ON WAVELET TRANSFORM FOR SPEECH

Chih-Hsu Hsu
Department of Applied Healthcare Informatics, Ching Kuo Institute of Management and Health,
No.336, Fu Hsin Rd., Keelung 203, Taiwan, R.O.C.,

ABSTRACT

This paper proposed a fuzzy system to discriminate speech signals from background. In previous works, we had developed a method for speech classification. For speech classification, the universe of discourse is divided into two types, and each type is treated as a class. These are background and speech signals. The rectangular fuzzy system (RFS) is used to classify frames and integrate the rule-based approach. The variance of first detail and the third approximation can extract fuzzy classification rules. Experimental results demonstrate the superior performance to the conventional ones. The effectiveness of the proposed system is confirmed by the experimental results.

Keywords: fuzzy system, endpoint detection, wavelet transform.

1. INTRODUCTION

In this paper, speech detection is presented with a rectangular fuzzy system (RFS). Speech detection is very important in speech processing. Major cause of errors in speech recognition systems is inaccurate detection from the utterance. For conventional technique, the energy and zero crossing rates are performed [1]. They play important roles in the higher signal-to-noise ratio (SNR) environments, but their performance becomes poor in the lower SNR conditions. In our previous works, we develop a method for speech classification [2]. The first wavelet and the third scaling functions give good indications to discriminate speech signals from the background. In the most of speech detection, it is necessary to determine the threshold to discriminate between speech and background signals. We proposed a novel detection method, i.e. fuzzy rules rather than the threshold, to distinguish speech signals from background. The behavior of the features can be explained based on fuzzy rules and their performance can be adjusted by tuning the rules.

Recently, several approaches focus on generating fuzzy if-then rules directly from numerical data. In most of fuzzy systems, construction of fuzzy rules from numerical data for classification problems consists of two phases: (1) fuzzy partition of a pattern space and (2) identification of a fuzzy rule for each fuzzy subspace. The major restriction of this approach is that the number of divisions of each input variable must be pre-selected. In addition, the degree of partitions will affect the classification power and the number of generated fuzzy rules. One approach to remedy the mentioned disadvantages is to use the concept of distributed representation of fuzzy rules which is implemented by super-composing many fuzzy rules corresponding to different fuzzy partitions of a pattern space [3]. However, this approach will still result in a lot of unnecessary fuzzy rules. The genetic algorithm has been proposed for choosing an appropriate set of fuzzy rules [4]. Fuzzy rules with variable fuzzy regions are extracted for classification problems. These approaches do not need to define the number of divisions of each input variable in advance [5,6]. Each class is represented by a set of hyper-boxes, in which overlaps among hyper-boxes for the same class are allowed, but no overlaps are allowed between different classes [5]. However this approach may not easily handle patterns where complicate separate boundaries exist. To overcome this problem, two types of hyper-boxes: (1) activation hyper-boxes and (2) inhibition hyper-boxes were proposed in [6]. A RFS is proposed to extract both crisp and fuzzy rules from numerical data.

After we have finished the RFS, the speech signal is correctly segmented into syllabic units. If we assume that every frame is classified correctly into one of the two types: silence (0) and speech (1), then the frames of the speech signals can be represented by the labeling sequences as

$$00...011...111\ldots100...0. \qquad (1)$$

This paper is organized as follows. In section 2 we discuss the speech features based on wavelet transform. Section 3 briefly describes the class of RFS. The experimental results are given in Section 4. Finally, some concluding remarks are presented in Section 5.

2. Speech Features Based on Wavelet Transform

The multi-resolution formulation of wavelet transform is obviously designed to represent signals where a single event is decomposed into finer and finer detail, but it turns out also to be valuable in representing signals where a time-frequency or time-scale description is desired even if no concept of resolution is needed. In many applications, one studies the decomposition of a signal in terms of basis function. For example, stationary signals are decomposed into the Fourier basis using Fourier transform. For non-stationary signals (i.e. signals whose frequency characteristics are time-varying like music, speech, image, etc.) the Fourier basis is ill-suited because of the poor time-localization. The classical solution to this problem is to use the short-time (or windowed) Fourier transform. However, the short-time Fourier transform has several problems, the most severe being the fixed time-frequency resolution of the basis functions. Wavelet techniques give a new class of bases that have desired time-frequency resolution properties. The “optimal” decomposition depends on the signal studied.

The definition of the scaling function $\phi_{j,k}(t)$ and wavelet function $\psi_{j,k}(t)$ is given by [7].

$$\phi_{j,k}(t) = 2^{j/2}\phi(2^j t - k) \qquad j,k \in Z \qquad (2)$$

$$\psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k) \qquad j,k \in Z \qquad (3)$$

This two-variable set of basis function is used in a way similar to the short time Fourier transforms. A signal space of multi-resolution approximation is decomposed by wavelet

Vol. 4, No. 06; 2019

ISSN: 2456-3676

transform in an approximation (lower resolution) space and a detail (higher resolution) space. In order to generate a basis system that would allow higher resolution decomposition at higher frequencies, we will iterate the wavelet transform recursively to divide the approximation space, giving a left binary tree structure. The wavelet packet was allowed a finer and adjustable to particular signals or signal classes. The wavelet packet decomposes the detail spaces as well as approximation ones.

In the higher SNR environment, the energy and ZCR of speech signals are higher than background in a stream of utterance. However, the performance is not satisfactory under the lower SNR environment. Here, the first detail and the third approximation play very important roles under the noisy environment. The windowed Fourier transform has uniform resolution over the time frequency plane. It is difficult to detect sudden burst in a slowly varying signal by Fourier transform. Wavelet transform overcomes the problem of fixed resolution, using adaptive window sizes, which allocate more time to the lower frequency and less time for the higher frequency [8].

3. Rectangular Fuzzy System (RFS)

The construction of a rule-based expert system involves the process of acquiring production rules. Production rules are often represented as " IF condition THEN act. The class of RFS provides a tool for machine learning. The classification knowledge is easily extracted from the weights in a rectangle. First, we divided the range of an output variable into many intervals and using the input data belonging to each interval. Each rule is composed of an activation rectangle, which defines the existence region of a class and, if necessary, an overlapping rectangle which overlapped the existence of data in that activation rectangle. We determine activation rectangle, which define the input region corresponding to the class, by calculating the maximum and minimum values of input data for each class. Figure 1 illustrates the architecture of a RFS.

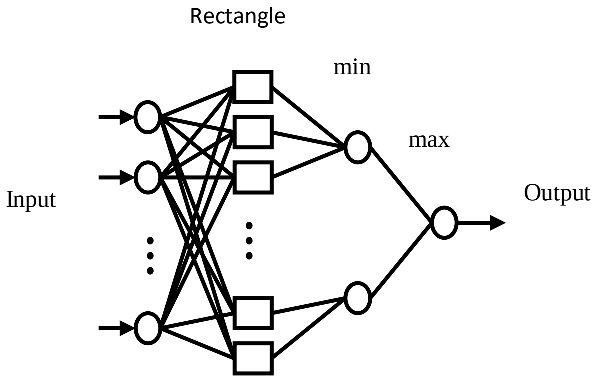

Fig. 1 A RFS Architecture.

4. PERFORMANCE EVALUATION

For a given utterance, it is first sampled and converted to digital form through the A/D converter. After the framing process that divides sequence of signals into sequence of frames, the discrete wavelet transformed features of each scaling and wavelet functions are calculated.

In our experiments, two male and two female utter the database. Speech signals are sampled at 22.05KHz with 8 bits resolution. Speech features extracted from each frame with 1024 samples. Haar function is used for wavelet mother function. The first detail and the third approximation are used as the input variables to the RFS to be trained. The values of the features of the trained RFS are easily utilized to represent a set of if-then rules.

The performance of proposed method is compared with the conventional speech detection algorithm that used energy and ZCR, whose threshold is adjusted to get a best performance. Table 1 shows the results of speech detection accuracy. The effectiveness of the proposed system is confirmed by the experimental results. The whole results seem encouraging.

Table 1 The results of speech detection accuracy.

System	Male 1	Male 2	Female 1	Female 2	Average
Conventional	76.3%	79.6%	77.6%	81.2%	78.7%
RFS	96.1%	95.2%	96.8%	97.9%	96.5%

5. PERFORMANCE EVALUATION

In this paper, a rectangular fuzzy system for speech detection is presented. We utilize RFS to classify the speech data. The fuzzy rules with variable fuzzy regions were defined by activation rectangles, which show the existence region of data for a class and overlapping rectangles, which overlapping the existence of the data for the other classes. These rules were extracted directly from speech features, the first detail and the third approximation. A novel detection method is determined by fuzzy rules rather than the threshold, to distinguish speech signals from background. The class of RFS is utilized to detect frames of speech signals into two classes: background and speech signals. In the near future, we will try to apply RFS to adjust features to speech recognition system.

References

L. R. Rabiner and R. W. Schafer, “Digital Processing of Speech Signals,” Prentice-Hall, 1978.

W.C. Chen, C.T. Hsieh, and C.H. Hsu, “Two-stage vector quantization based multi-band models for speaker identification,” In Proceedings of International Conference on Convergence Information Technology, Gyeongju, Korea, pp.2336-2341, Nov. 21-23, 2007.

H. Ishibuchi, K. Nozaki and H. Tanaka, “Distributed Representation of Fuzzy Rules and its Application to Pattern Classification,” Fuzzy Sets and System 52, 1992, pp. 21-31.

H. Ishibuchi, K. Nozak and N. Yamamoto, “Selecting Fuzzy Rules by Genetic Algorithm for Classification Problems,” 2nd IEEE Int. Con. on Fuzzy Systems, 1993, pp. 1119-1124.

P. K. Simpson, “Fuzzy Min-Max Neural Networks-Part1: Classification,” IEEE Trans. on Neural Networks, Vol. 3, Sept. 1992, pp. 776-786.

S. Abe and M. S. Lan, “Fuzzy Rules Extraction Directly from Numerical Data for Function Approximation,“ IEEE Trans. on System, Man, and Cybernetics, Vol. 25, No. 1, Jan. 1995, pp. 119-129.

C. S. Burrus, R. A. Gopinath, and H. Guo, “Introduction to Wavelets and Wavelet Transforms,” Prentice-Hall, 1998.

T.H. Luo and C.H. Hsu, “Signal Analysis for “Kagaya Miyamoto Shiki” Music Therapy,” In Proceedings of International Scientific Conference on Engineering and Applied Sciences, Singapore, pp. 283-284, Aug. 15-17, 2014.