International Journal of Advanced Engineering and Management Research

Vol. 7, No. 06; 2022

ISSN: 2456-3676

Multi-resolution Speech Recognition Based on Hyper-rectangular Fuzzy System

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.,

doi: 10.51505/ijaemr.2022.7604 URL: http://dx.doi.org/10.51505/ijaemr.2022.7604

Abstract
This paper presents a multi-resolution feature extraction technique to speech recognition. The proposed multi-resolution feature extraction technique uses wavelet transform and wavelet packet to calculate features of each sub-band in order not to spread noise distortions over the entire feature space. In our previous works, we had developed a method for speech classification. For speech classification, the universe of discourse is divided into many types, and each type is treated as a class. The hyper-rectangular fuzzy system is used to classify frames and integrate the rule-based approach. The variances of each sub-band are utilized to extract both crisp and fuzzy classification rules. In our experiments, the Texas Instruments/Massachusetts Institute of Technology database is used and extracts features of phonemes. The results demonstrate the superior performance to Mel frequency cepstral coefficients. The effectiveness of the proposed system is encouraging.

Keywords: speech recognition, wavelet transform, multi-resolution, feature extraction, and fuzzy system.

1. Introduction

Most of the features of speech recognition systems are Mel frequency cepstral coefficients (MFCC). Davis and Mermelstein showed that a MFCC based speech recognizer outperforms other feature based (Linear Prediction Cepstral Coefficients (LPCC), Linear Prediction Coefficients (LPC), Reflection Coefficients (RC), and Linear filter Cepstral Coefficients (LFCC)) speech recognizers. The windowed Fouier transform (FT) has uniform resolution over the time frequency plane. It is difficult to detect sudden burst in a slowly varying signal by FT. Recently, wavelet transform (WT) has been proposed for feature extraction. In this paper, we propose a multi-resolution feature extraction (MRFE) technique to speech recognition. Noise is one of the most principal problems in speech recognition systems. The performance starts to degrade rapidly when the recognition is transfer to noisy environments. To improve the performance of speech recognition system, it is crucial to extract features of high qualities. The MRFE uses WT and wavelet packet (WP) to calculate features of each sub-band in order not to spread noise distortions over the entire feature space.

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

Vol. 7, No. 06; 2022

ISSN: 2456-3676

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. 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. 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. 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. 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 our previous works, we develop a method for speech classification. For speech classification, the universe of discourse is divided into many types, and each type is treated as a class. The hyper-rectangular fuzzy system (HRFS) is used to classify frames and integrate the rule-based approach. The variances of each sub-band are utilized to extract both crisp and fuzzy classification rules. The behavior of the features can be explained based on fuzzy rules and their performance can be adjusted by tuning the rules.

This paper is organized as follows. In section 2 we discuss the MRFE technique based on WT and WP. Section 3 briefly describes the class of HRFS. The experimental results are given in Section 4. Finally, some concluding remarks are presented in Section 5.

2. Multi-Resolution Feature Extraction Technique Based on Wavelet Transform and Wavelet Packet

2.1 Wavelet Transform and Wavelet Packet

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

$$\phi_{j,k}(t) = 2^{j/2}\phi(2^j t - k) \qquad j,k \in Z \tag{1}$$

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

A signal space of multi-resolution approximation is decomposed by WT in a approximation (lower resolution) space and a detail (higher resolution) space. Figure 1 shows the corresponding tiling description of time-frequency resolution properties of two-scale of WT. WT recursively divides the approximation space, giving a left binary tree structure, and WP decomposes the detail spaces as well as approximation ones. Figure 2 illustrates the corresponding tiling description of time-frequency resolution properties of four-scale of WP.

Vol. 7, No. 06; 2022

ISSN: 2456-3676

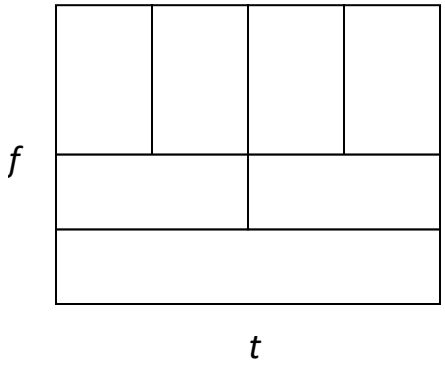

Fig. 1 Two-scale of WT.

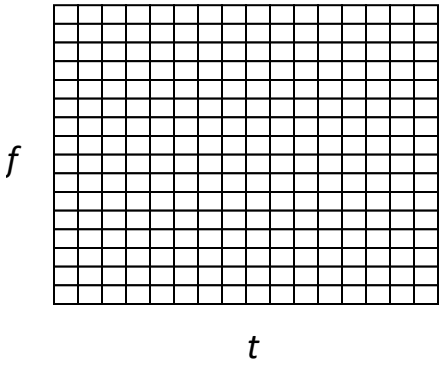

Fig. 2 Four-scale of WP.

2.2 Multi-Resolution Feature Extraction Technique

The speech in the Texas Instruments/Massachusetts Institute of Technology (TIMIT) database is sampled at 16 KHz, giving an 8 KHz bandwidth signal. First, two-scale of WT is calculated. This partitions the frequency axis into three bands. Then, four-scale of WP further decompose the lower band from 0-2 KHz. The partition of the frequency axis into sixteen bands each of 125 Hz. After performing the decomposition of WT and WP, the variance in each frequency band is calculated as features of MRFE. Table 1 illustrates the distribution of frequency bands.

Table 1 The frequency distribution of MRFE.

No. of bands	Lower cut off frequency (Hz)	Higher cut off frequency (Hz)	Bandwidth (Hz)
1	0	125	125
2	125	250	125
3	250	375	125
4	375	500	125
5	500	625	125
6	625	750	125
7	750	875	125
8	875	1000	125
9	1000	1125	125
10	1125	1250	125
11	1250	1375	125
12	1375	1500	125
13	1500	1625	125
14	1625	1750	125
15	1750	1875	125
16	1875	2000	125
17	0	2000	2000
18	2000	4000	2000
19	0	4000	4000
20	4000	8000	4000

Vol. 7, No. 06; 2022

ISSN: 2456-3676

It is acknowledged that the frequency selectivity plays an important role in the human hearing process. A band-limited noise does not spread over the entire feature spaces, since the multi-bands of features are almost independent. A pure sub-bands based approach may lose the information on the correlation between various sub-bands. Therefore, we challenge this view by selecting above frequency distribution.

3. Hyper-Rectangular Fuzzy System (HRFS)

3.1 HRFS Architecture

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 HRFS provides a tool for machine learning. The classification knowledge is easily extracted from the weights in a hyper-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 hyper-rectangle, which defines the existence region of a class and, if necessary, an overlapping hyper-rectangle which overlapped the existence of data in that activation hyper-rectangle. We determine activation hyper-rectangle, which define the input region corresponding to the class, by calculating the maximum and minimum values of input data for each class. Figure 3 illustrates the architecture of a HRFS.

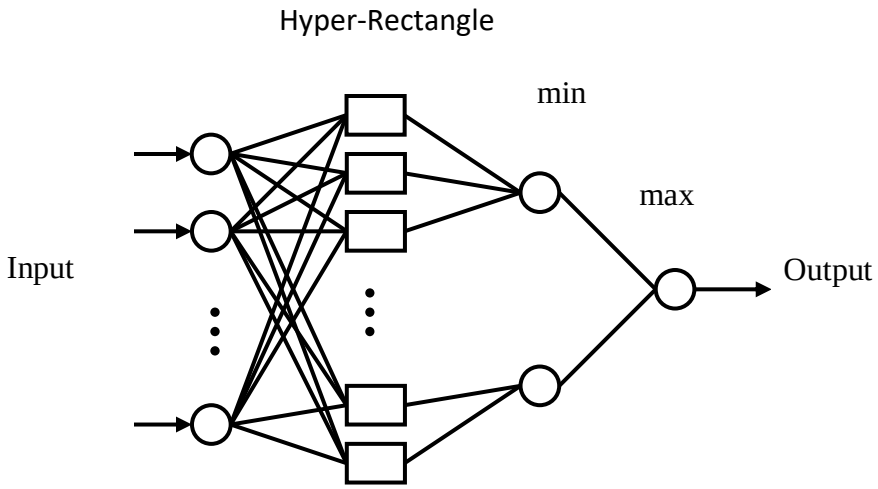

Fig. 3 A HRFS Architecture.

3.2 Fuzzy Rule Extraction of HRFS

Let a set of input data for class i is x_i, where $i = 1, \ldots n$. We define the activation hyper-rectangle A_{ij} as

$$A_{ij} = \left\{ x \,\middle|\, u_{ijk} \le x_k \le U_{ijk}, k = 1,\ldots n \right\} \tag{3}$$

and define the fuzzy rule r_{ij} without overlapping as follows:

Vol. 7, No. 06; 2022

ISSN: 2456-3676

If x is A_{ij}, then x belongs to class i, (4)

the overlapping hyper-rectangle I_{ij} as

$$I_{ij} = \left\{ x \,\middle|\, v_{ijk} \le x_k \le V_{ijk}, k = 1,...n \right\} \quad (5)$$

and define the fuzzy rule r_{ij} with overlapping hyper-rectangle as follows:

If x is A_{ij} and x is not I_{ij}, then x belongs to class i, (6)

If x is not (4) and (6), then calculates the degree of membership of each class by fuzzy rule inference.

3.3 Fuzzy Rule Inference of HRFS

The degree of membership of the fuzzy rule for a given input x is determined by the membership function of the activation hyper-rectangle. While the degree of membership of the fuzzy rule for a given input x is determined by the difference between the membership function of the activation hyper-rectangle and that of the overlapping hyper-rectangle. The membership function for each input variable is a trapezoidal shape. Figure 4 shows one-dimension membership function for the hyper-rectangle, where u_k and U_k denote the minimum and maximum values of the k-th dimension of the hyper-rectangle, respectively.

$$m_X(x) = \min_{k=1,...n} m_X(x,k) \quad (7)$$

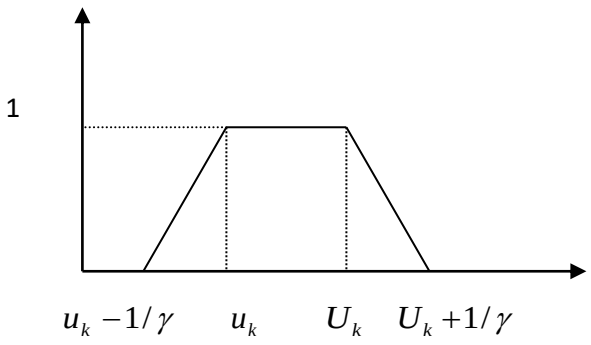

Fig. 4 One-dimension membership function for the hyper-rectangle.

$$m_X(x,k) = \begin{cases} 1 & for\ u_k \le x_k \le U_k \\ 1 - \max(0, \min(1, \gamma(u_k - x_k))) & for\ x_k \le u_k \\ 1 - \max(0, \min(1, \gamma(x_k - U_k))) & for\ x_k \ge U_k \end{cases} \quad (8)$$

Vol. 7, No. 06; 2022

ISSN: 2456-3676

where γ is a sensitive parameter. The minimum value in (7) is taken so that the degree of membership within the hyper-rectangle and on the surface of the hyper-rectangle becomes 1.

The degree of membership of a fuzzy rule respected by (4) is:

$$d_{r_{ij}}(x) = m_{A_{ij}}(x) \quad (9)$$

The degree of membership of a fuzzy rule respected by (6) is:

$$d_{r_{ij}}(x) = \max(0, m_{A_{ij}}(x) - m_{I_{ij}}(x)) \quad (10)$$

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 MRFE based on WP and WP of each frame are calculated.

In our experiments, the speech is TIMIT database with dialect region DR1. DR1 contains a total of 490 utterances, 10 sentences spoken by each of 49 speakers (31 males and 18 females). 24 males and 14 females formed the training set; the test set consisted of the others. Speech signals are sampled at 16KHz with 16 bits resolution. Phonemic features extracted from each frame with 1024 samples. Haar function is used for WT. The variance in each of the sub-bands is calculated. Twenty features of sub-bands are used as the input variables to the HRFS to be trained. The values of the features of the trained HRFS are easily utilized to represent a set of if-then rules. Figure 5 shows the recognition accuracy for different scales of WP. It is shown that scales larger than four-scale give approximation performances, so the four-scale is treated in our experiments.

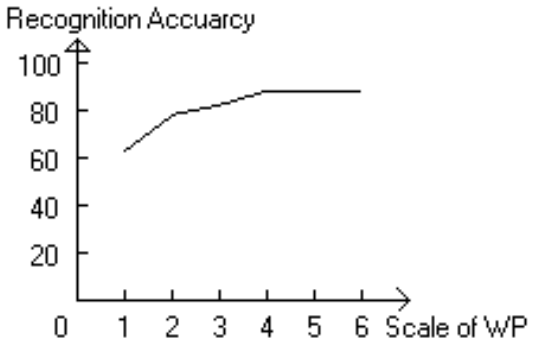

Fig.5 The recognition accuracy for different scales of WP.

The performance of proposed method is compared with the MFCC features. Table 2 shows the results of speech recognition accuracy. It is observed that MFCC is better than MRFE only for the vowels, since MFCC uses Fourier transform that is more efficient to extract the periodic structure from a signal. The overall recognition rate of the MRFE is superior to the MFCC. The

Vol. 7, No. 06; 2022

ISSN: 2456-3676

effectiveness of the proposed system is confirmed by the experimental results. The whole results seem encouraging.

Table 2 The results of speech recognition accuracy.

System	Stops	Affricates	Fricatives	Nasals	Semi-vowels	Vowels	Average
MRFE	76.8%	76.5%	88.6%	86.2%	86.1%	84.8%	83.2%
MFCC	74.6%	73.8%	86.3%	85.6%	85.3%	85.2%	81.8%

5. Concluding Remarks

In this paper, multi-resolution feature extraction technique is presented for speech recognition system. Since the multi-bands of features are almost independent, and the band-limited noise does not spread over the entire feature spaces. Then, we utilize a hyper-rectangular fuzzy system to extract fuzzy rules for classification. The fuzzy rules with variable fuzzy regions were defined by activation hyper-rectangles, which show the existence region of data for a class and overlapping hyper-rectangles, which overlapping the existence of the data for the other classes. These rules were extracted directly from speech features. In the near future, we will try to apply HRFS to adjust features to improve the speech recognition system.

References

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

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

Davis, S. B. and Mermelstein, P. (1980). *Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuous Spoken Sentences,* IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol. 28, No. 4.

Farooq, O. and Datta, S. (2001). *Mel Filter-Like Admissible Wavelet Packet Structure for Speech Recognition* (pp. 196-198), IEEE Signal Processing Letters, Vol. 8, No. 7, July.

Hsieh, C. T., Su, M. C. and Hsu, C. H. (1996). *Continuous Speech Segmentation Based on a Self-Learning Neuro-Fuzzy System* (pp. 1180-1187), IEICE, Trans. Fund., Vol. E79-A, No. 8 August.

Hsieh, C. T. and Hsu, C. H. (2001). *Application of Hyper-Rectangular Fuzzy System for Speech Classification* (pp. 300-303), 2001 Ninth National Conf. on Fuzzy Theory and Its Applications, Nov., Taiwan.

Hsu, C. H. (2019). *Endpoint Detection Based on Wavelet Transform for Speech* (pp.1-5). International Journal of Advanced Engineering and Management Research, Vol. 4, No. 6.

Vol. 7, No. 06; 2022

ISSN: 2456-3676

Luo, T. H. and Hsu, C. H. (2014). *Signal Analysis for "Kagaya Miyamoto Shiki" Music Therapy* (pp. 283-284). In Proceedings of International Scientific Conference on Engineering and Applied Sciences, Singapore.

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

TIMIT (1990). *Acoustic-Phonetic Continuous Speech Corpus*, NIST Speech Disc 1-1.1, Oct.