Mandarin Chinese Tone Recognition Based on Hyper-rectangular Fuzzy System

Chih-Hsu Hsu
Centre for General Education, Deh Yu College of Nursing and Health,
No.336, Fu Hsin Rd., Keelung 203, Taiwan, R.O.C.,

doi.org/10.51505/ijaemr.2025.1418 URL: http://dx.doi.org/10.51505/ijaemr.2025.1418

Received: Oct 11, 2025 Accepted: Oct 21, 2025 Online Published: Oct 28, 2025

Abstract
This paper presents a multi-resolution feature extraction technique to Mandarin Chinese tone 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. Mandarin Chinese tone recognition involves identifying and distinguishing between four tones. Accurate tone recognition is vital for language understanding systems because tones carry lexical meaning. In our experiments, the Chinese corpus is used and extracts features of tones. The effectiveness of the proposed system is encouraging.

Keywords: tone recognition, wavelet transform, feature extraction, and fuzzy system.

1. Introduction

Mandarin Chinese four tones refer to the four tones, namely level, rising, departing and entering. Mandarin Chinese tone recognition involves identifying and distinguishing between the language's four main lexical tones (high, rising, low, and falling) and the neutral tone. It is a crucial skill for language learners because different tones change the meaning of a syllable, and accurate recognition is also a key area of research for automatic recognition systems, which use techniques like Hidden Markov Models (HMMs) and neural networks.

In the field of tone recognition research, speech feature extraction is a crucial processing step and is crucial to the overall performance of the system. Among various speech features, cepstral coefficients are the most widely used and representative speech parameters. Linear Predictive Cepstral Coefficients (LPCCs) are used in speaker identification primarily due to their computational simplicity and ability to effectively represent vowel characteristics. Another speech feature that is even more widely used in voiceprint identification is Mel-Scale Frequency Cepstral Coefficients (MFCCs). This feature's greatest advantage is that it accounts for the characteristics of human hearing. Since the ear's response to frequency is logarithmic rather than linear, the calculation of cepstral coefficients places a greater emphasis on low frequencies. This characteristic makes the resulting coefficients more resistant to noise interference.

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 tone recognition. Noise is one of the most principal problems in tone 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 quality. 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.

The four tones are a concept in Chinese phonology, referring to the four tones of Middle Chinese and their evolution. The four tones have evolved differently in various Chinese dialects and other languages that have borrowed Chinese vocabulary. 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 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 complicated separate boundaries exist. To overcome this problem, two types of hyper-boxes: (1) activation hyper-boxes and (2) inhibition hyper-boxes were proposed.

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 pitch contours of Mandarin Chinese's four tones and 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

In the field of signal analysis, Fourier analysis is a well-known technique, decomposing a signal into a combination of sinusoidal waves of different frequencies. From a mathematical perspective, Fourier analysis involves converting a signal from the time domain to the frequency domain to observe its characteristics, a process known as the Fourier transform (FT). For many signals, Fourier analysis is very useful because the signal's frequency content contains important information. However, Fourier analysis has a serious drawback: when a signal is converted from the time domain to the frequency domain, its temporal information is lost. To address this shortcoming, the Short-Time Fourier Transform (STFT) is used to analyze only a small portion of the signal at a time, essentially creating a window on the signal. The drawback of the STFT is that its time-frequency window is fixed in size. Wavelet analysis addresses this issue by allowing the time-frequency window used to observe the signal to have an adjustable size, which is the adaptability of wavelet analysis.

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 scales of WP.

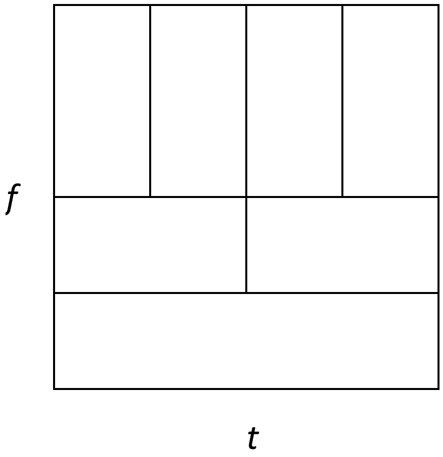

Figure 1: Two-scale of WT

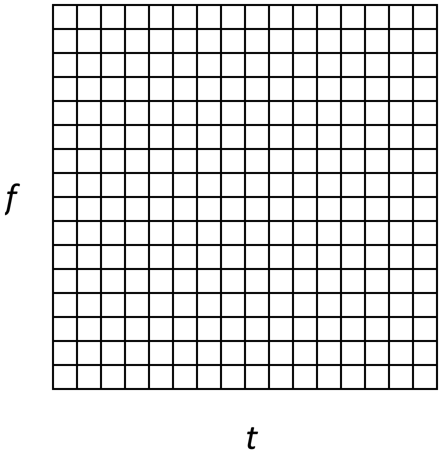

Figure 2: Four-scale of WP

2.2 Multi-Resolution Feature Extraction Technique

The Chinese corpus 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 multi-resolution feature extraction (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

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 space, since the multi-bands of features are almost independent. A pure sub-band-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)

The hyper-rectangle fuzzy system has an identifier for complex decision boundaries and uses a non-uniform cutting method. This general cutting method can reflect the interactions between input variables, modifying weights through error propagation, and approximating hyper-rectangle fuzzy sets using a multidimensional rectangle.

3.1 HRFS Architecture

We use the so-called supervised decision-directed learning (SDDL) algorithm to train a hyper-rectangle fuzzy system. Based on the clustering characteristics of the training data, we extract information implicit in the data. This extracted information is represented as if-then rules for the hyper-rectangle fuzzy system. As long as there is no overlap in the training data, the classification success is guaranteed. Therefore, it is easy to extract classification rules using if-then rules. Figure 3 shows the basic architecture of the HRFS.

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.

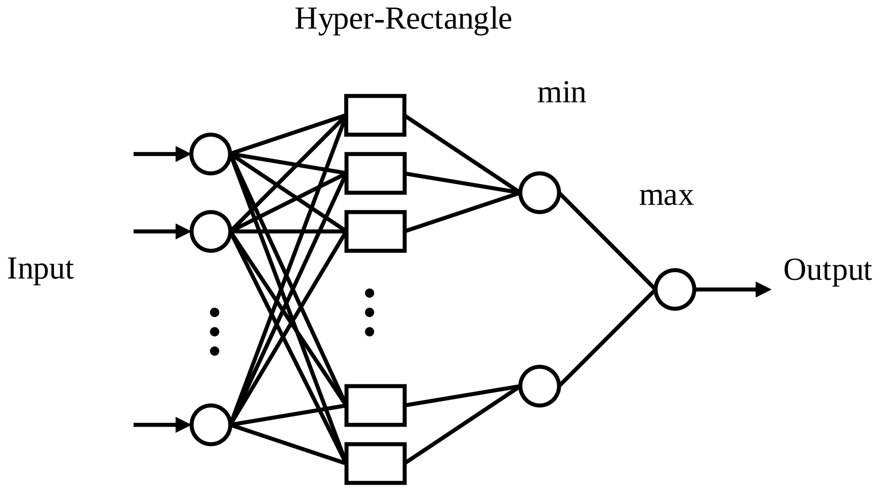

Figure 3: A HRFS Architecture

3.2 Fuzzy Rule Extraction of HRFS

Let a set of input data for class *i* is x_i, where *i* = 1*, … 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:

$$\text{If } x \text{ is } A_{ij}, \text{ then } x \text{ belongs to class } i, \tag{4}$$

the overlapping hyper-rectangle I_{ij} as

$$I_{ij} = \left\{ x \,\middle|\, v_{ijk} \le x_k \le V_{ijk}, k = 1,\dots n \right\} \tag{5}$$

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

$$\text{If } x \text{ is } A_{ij} \text{ and } x \text{ is not } I_{ij}, \text{ then } x \text{ belongs to class } i, \tag{6}$$

If x is not (4) and (6), then calculate 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.

$$m_X(x) = \min_{k=1,\dots n} m_X(x,k) \tag{7}$$

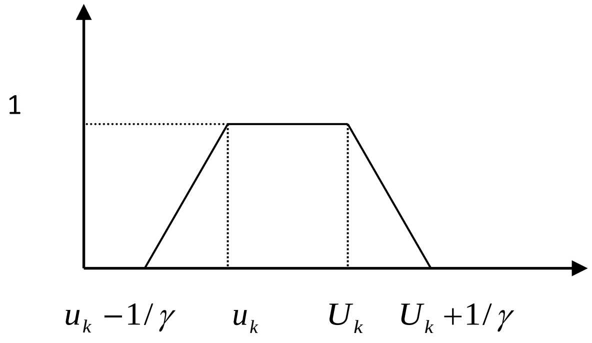

Figure 4: One-dimension membership function for the hyper-rectangle

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

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:

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

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

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

4. Performance Evaluation

The Chinese corpus 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 decomposes the lower band from 0-2 KHz.

4.1 The pitch contours of Mandarin Chinese four tones

Mandarin Chinese four tones refers to the four tones, namely level, rising, departing and entering. Figure 5 shows the pitch contours of Mandarin Chinese four tones of Mandarin Chinese.

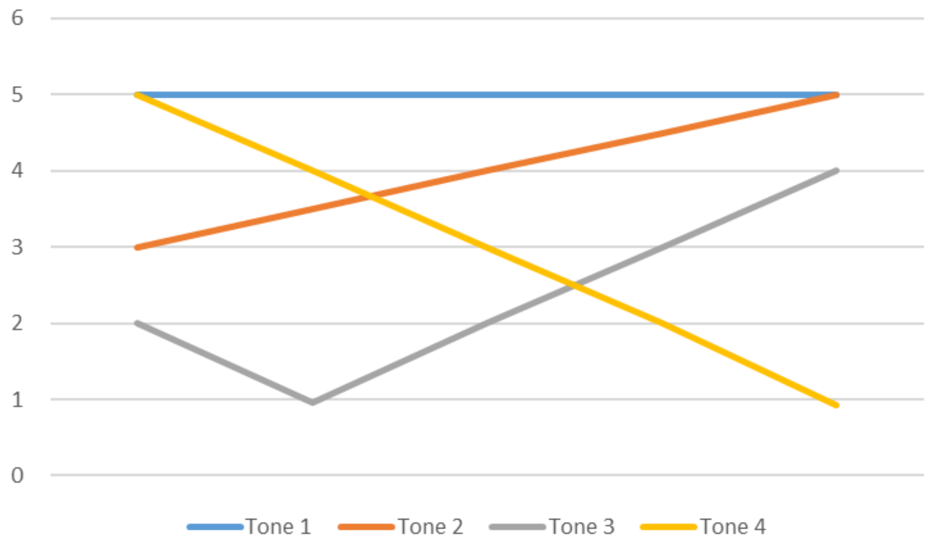

Figure 5: The pitch contours of Mandarin Chinese four tones

From the four-tone fundamental frequency modes, the pitch length of the four tones can be roughly estimated, of which the third tones are the longest, the first and second tones are the second, and the fourth tones are the shortest. If it is consonant combined with vowel, the sound length is longer. In the structure of the Chinese syllabes, apart from the problems of pitch and length of tones, there are also energy problems. Among the four tones, the fourth tones have the highest energy, the first and second tones are the second, and the third tones have the lowest energy. Each word itself also has an energy problem; the volume at the beginning and the end is low, and the volume in the middle is higher.

A phonetic notation for the pitch component of Mandarin Chinese tones devised is commonly used in the Mandarin Chinese literature on tone. It consists of a 5-degree scale dividing an ideal speaker's voice range, from 1-low to 5-high. The isolation contours of the four Mandarin Chinese tones are the following:
Tone 1 is [55] high level
Tone 2 is [35] mid to high rising
Tone 3 is [214] mid-low to low falling, then rising to mid-high
Tone 4 is [51] high to low falling

4.2 The experimental results

In our experiments, the corpus sampled at 16 KHz with 16-bit resolution. Tones 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 6 shows the recognition accuracy for Mandarin Chinese tones. The effectiveness of the proposed system is encouraging.

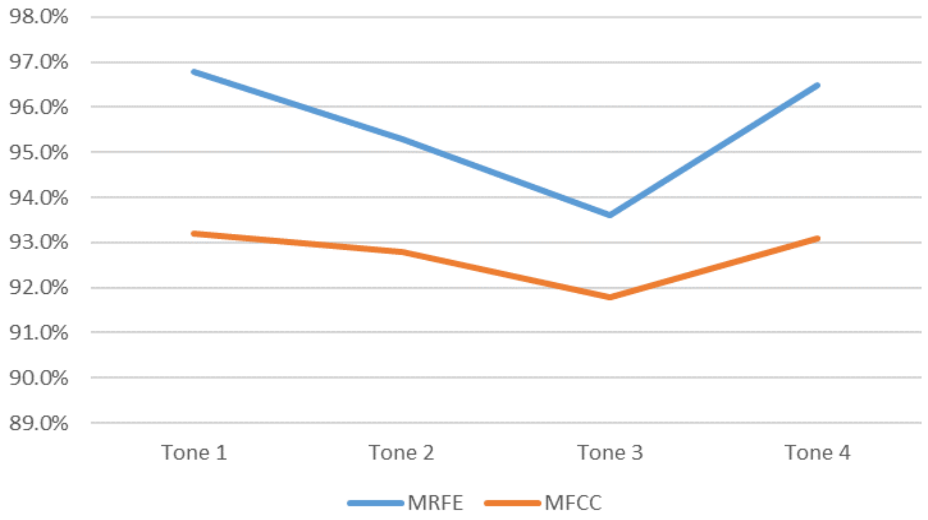

Figure 6: The recognition accuracy for Mandarin Chinese tones

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 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	Tone 1	Tone 2	Tone 3	Tone 4	Average
MRFE	96.8%	95.3%	93.6%	96.5%	95.6%
MFCC	93.2%	92.8%	91.8%	93.1%	92.7%

5. Concluding Remarks

In this paper, multi-resolution feature extraction technique is presented for Mandarin Chinese tone recognition system. For speech features, the wavelet transform parameters of the speech signal are extracted. A hyper-rectangle fuzzy system (HRFS) is used as the discriminator. This HRFS features a discriminator with a complex decision boundary. It modifies weights through error propagation. Training the HRFS with a supervised decision-directed learning (SDDL) algorithm facilitates the extraction of classification rules using an if-then approach. The Chinese corpus is used and extracts features of tones. The effectiveness of the proposed system is encouraging.

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.

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. (2022). *Multi-Resolution Speech Recognition Based on Hyper-Rectangular Fuzzy System* (pp. 36-43). International Journal of Advanced Engineering and Management Research, Vol. 7, No. 6.

Hsu, H. L. (2020). *A Preliminary Study of the Tonal Features of Central Taiwan Mandarin* (pp. 115-157). Taiwan Journal of Linguistics (in Chinese). Vol. 18, No 1.

Hu, X. and Perry, J. (2018). *The syntax and phonology of non-compositional compounds in Yixing Chinese* (pp. 701-742). Natural Language and Linguistic Theory Vol. 36.

Rabiner, L. R. (1989). *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition* (pp. 257-286). Proceedings of the IEEE, Vol. 77, No. 2.

Sagart, L. (1999). *The origin of Chinese tones.* (pp. 91-104). Proceedings of the Symposium /Cross-Linguistic Studies of Tonal Phenomena/Tonogenesis, Typology and Related Topics. Tokyo, Japan.

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

Wang, William S.-Y. and Sun, Chaofen (2015). *The Oxford Handbook of Chinese Linguistics* (pp. 80-90). Oxford University Press.