Available online at www.sciencedirect.com
ScienceDirect
Expert Systems with Applications 36 (2009) 3136–3143
Expert Systems with Applications
www.elsevier.com/locate/eswa

Speaker identification using discrete wavelet packet transform technique with irregular decomposition

Jian-Da Wu *, Bing-Fu Lin

Graduate Institute of Vehicle Engineering, National Changhua University of Education, 1 Jin-De Road, Changhua City, Changhua 500, Taiwan

Abstract

This paper presents the study of speaker identification for security systems based on the energy of speaker utterances. The proposed system consisted of a combination of signal pre-process, feature extraction using wavelet packet transform (WPT) and speaker identification using artificial neural network. In the signal pre-process, the amplitude of utterances, for a same sentence, were normalized for preventing an error estimation caused by speakers' change in volume. In the feature extraction, three conventional methods were considered in the experiments and compared with the irregular decomposition method in the proposed system. In order to verify the effect of the proposed system for identification, a general regressive neural network (GRNN) was used and compared in the experimental investigation. The experimental results demonstrated the effectiveness of the proposed speaker identification system and were compared with the discrete wavelet transform (DWT), conventional WPT and WPT in Mel scale.

Keywords: Speaker identification; Discrete wavelet transform; Wavelet packet transform; General regressive neural network

1. Introduction

The applications of voice signal processing such as sound recognition or speaker identification have been increased rapidly in recent years. Because of its non-contact characteristic, speaker identification system can be utilized in suspect identification. The implementation of speaker identification can be divided into two stages: The first is feature extraction and the second is speaker classification based on the extracted features (Sarikaya, Pellom, & Hansen, 1998). At the stage of feature extraction, the extracted features should be capable of separating the speakers from each other in its space. In traditional techniques, speech features are usually obtained via Fourier transforms and short time Fourier transforms. However, these techniques are not suitable for speaker identification because they accept stationary signal within a given time frame and may therefore lack the ability to analyze the non-stationary signals or signals in transient state (Avci & Akpolat, 2006). Generally speaking, a speaker identification system can be implemented by observing the voiced/unvoiced components or through analyzing the energy distribution of utterances. Some digital signal processing methods, such as linear predictive coding technique (Adami & Barone, 2001; Tajima, Port, & Dalby, 1997), Mel frequency cepstral coefficients (MFCCs) (Mashao & Skosan, 2006; Sroka & Braida, 2005; Kanedera, Arai, Hermansky, & Pavel, 1999), discrete wavelet transform (DWT) (Fonseca, Guido, Scalassara, Maciel, & Pereira, 2007) and wavelet packet transform (WPT) (Lung, 2006; Zhang & Jiao, 2004) are widely used.

In 1990s, Mel frequency cepstral technique became the most widely used technique for recognition tasks due to its ability to represent the speech spectrum in a compact form (Sarikaya & Hansen, 2000). In fact, MFCCs are based on the model of humans' auditory perception and have been proven to be very effective in automatic speech recognition system and modeling the subjective frequency contents of audio signals. Recent research showed that the identification rate with MFCCs can be as high as 99.5% for the noise free TIMIT database (Sarikaya et al.,

* Corresponding author.
E-mail address: jdwu@cc.ncue.edu.tw (J.-D. Wu).

0957-4174/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2008.01.038

1998). However, the identification accuracy reduced to 60% for the same data set that was transmitted over telephone channels. It is pity though MFCCs achieve high identification rates in noise free environment; the results are not as good for same data which is corrupted by background and convolution channel noise (Sarikaya et al., 1998). Furthermore, the dramatic degradation in identification rates can be partly attributed to MFCCs which are not immune to noise. Another drawback is the assumption of a frame of speech that may contain information about two adjacent phonemes. The low frequency spectrum may be dominated by voiced phoneme information, and the high frequency spectrum may be dominated by the unvoiced part if one of these two phonemes is voiced and other is unvoiced (Gowdy & Tufekci, 2000).

In recent years, multi-resolution analysis based on wavelet theory was applied in many recognition tasks. Wavelet theory was proposed in 1984; Goupillaud et al. introduced a new transformation for the frequency analysis of the discretized signals. The transform is known as wavelet transform (Goupillaud, Grossman, & Morlet, 1984; Louis, Maass, & Rieder, 1997). In the aspect of speaker identification, many studies had developed the Mel filter-like structure to integrate the concept of Mel scale and multi-resolution capabilities (Farooq & Datta, 2001; Karam, Phillips, & Robertson, 2000; Torres & Rufiner, 2000). The ideal features for representing a speaker's identity should be substituted by some representative parameters to avoid complex computing. The advantage of WP parameters presented in Mel scale is that the model of extracted features will approach humans' auditory system; moreover, the number of parameters will be decreased.

In this study, a WPT based irregular decomposition approach is presented to improve the performance of speaker identification system. The irregularly-decomposed procedure is based on the energy of speakers' utterances and the motivation of driving this work is to develop a different thought for linguistic recognition. The following section will introduce the principles of wavelet and neural network, and the experimental results show an optimized decomposition using the proposed approach.

2. Feature extraction of speaker voice

2.1. Pre-process of the speech signals

Before the stage of feature extraction, the speech signals are pre-processed. The purpose of signal pre-process is to perform the normalization on speech signals to prevent the error estimation caused by speakers' volume changes. In other words, normalization makes the signals comparable regardless of differences in magnitude. In the present study, the signals are normalized by using the following equation (Lou & Loparo, 2004):

$$S_{P_i} = \frac{S_i - \mu}{\sigma}, \tag{1}$$

where S_i is the ith element of the signal S, μ and σ are the mean and standard deviation of the vector S, respectively; S_{P_i} is the ith element of the signal series S_P after normalization. Fig. 1a shows the diagram of speaking energies before normalization; the energy was dispersed in a wide range. Fig. 1b shows a distinction compared with Fig. 1a, in which the energy was concentrated in a smaller region. In addition, it will be helpful for the classifier to find similarities between the speech signals.

2.2. Sub-band based wavelet parameters

The MFCCs are the most common parametrization schemes employed in the speech recognition tasks (Davis & Mermelstein, 1980). However, MFCCs suffered from the background noise as mentioned in the introduction. A detailed discussion on the MFCCs is beyond the scope of this study, and therefore this paper only focuses on the wavelet transform intrinsically. In the following, the wavelet transform is defined as the inner product of a signal $x(t)$ with the mother wavelet $\psi(t)$:

$$\psi_{a,b}(t) = \psi\left(\frac{t-b}{a}\right) \tag{2}$$

$$W_\psi x(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} x(t)\psi * \left(\frac{t-b}{a}\right) \mathrm{d}t, \tag{3}$$

where a and b are the scale and shift parameters, respectively. Users can dilate or translate the mother wavelet by modulating a and b.

In the discrete version, the wavelet decomposes the signal with variable frames to perform multi-resolution analysis (MRA) in a dyadic form, which is known as discrete wavelet transform (DWT). In DWT, the scale and translation parameters of the discrete wavelet family are given by

$$a = a_0^j \tag{4}$$

$$\text{and } b = kb_0a_0^j \tag{5}$$

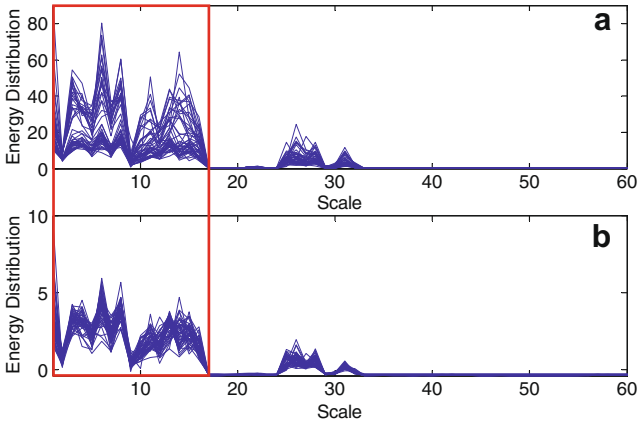

Fig. 1. Energy accumulation of sub-band parameters: (a) before normalization and (b) with normalization.

where j and k are integers. The function family with discretized parameters becomes

$$\psi_{j,k}(t) = a_0^{-j/2}\psi(a^{-j}t - kb_0) \tag{6}$$

In the equation, $\psi_{j,k}(t)$ is called DWT basis. Although it is called DWT, the time variable of the transform is still continuous. The DWT coefficients of a continuous time function are similarly defined as

$$d_{j,k} = \langle f_w(t), \psi_{j,k}(t)\rangle = \frac{1}{a_0^{j/2}}\int f_w(t)\psi(a_0^{-j}t - kb_0)\mathrm{d}t \tag{7}$$

When the transform is complete, the wavelet representation of a function $f_w(t)$ is expressed as

$$f_w(t) = \sum_j \sum_k \langle f_w(t), \psi_{j,k}(t)\rangle \psi_{j,k}(t) \tag{8}$$

Signals transformed by DWT can be regarded as putting data into a series of high-pass and low-pass filters. The high-frequency content of speech signals through a high-pass filter is preserved as "details". In the same way, the low-frequency content is preserved as "approximations", and only approximations can be decomposed iteratively. For speech signals, low-frequency content is the most important part, which provides the signal identity. The high-frequency content imparts flavor or nuance (Alkhaldi, Fakhr, & Hamdy, 2002). If the high-frequency components of the speech signal are removed, the voice will sound different but the speech can still be understood.

The wavelet packets transform (WPT) performs the recursive decomposition of the speech signal obtained by the recursive binary tree. Basically, the WPT is very similar to DWT but WPT decomposes both details and approximations instead of only performing the decomposition process on approximations. The principle of wavelet packet (WP) is that, given a signal, a pair of low pass and high pass filters is used to yield two sequences to capture different frequency sub-band features of the original signal. The two wavelet orthogonal bases generated form a previous node are defined as

$$\psi_{j+1}^{2p}(k) = \sum_{n=-\infty}^{\infty} h[n]\psi_j^p(k - 2^j n), \tag{9}$$

$$\psi_{j+1}^{2p}(k) = \sum_{n=-\infty}^{\infty} g[n]\psi_j^p(k - 2^j n), \tag{10}$$

where $h[n]$ and $g[n]$ denote the low-pass and high-pass filters, respectively. In Eqs. (9) and (10), $\psi[n]$ is the wavelet function. Parameters j and p are the number of decomposition levels and nodes of the previous node, respectively. In this study, both DWT and WPT are applied at the stage of feature extraction, but these data are not suitable for classifier due to a great amount of data length. Thus, we have to seek for a better representation for the speech features. An energy index of these sub-band signals was used and will be discussed in the next section.

2.3. Energy index of the sub-band signals

For a better representation of the sub-band signals, the energy of speech is often computed. Previous studies showed that the use of an energy index as features in recognition tasks is effective. Avci, Hanbay and Varol (2006) proposed a method to calculate the entropy value of the wavelet norm in digital modulation recognition. In the biomedical aspect, Behroozmand and Almasganj (2007) introduced a combination of genetic algorithm and wavelet packet transform used in the pathological assessment, and the energy features are computed from a group of wavelet packet coefficients. In 2003 Kotnik, Kacic and Horvat (2003) proposed a robust speech recognition system in a noisy environment using wavelet-based energy as a threshold for de-noise estimation. As seen in above studies, the energy index of the specific sub-band signal can be employed as features for recognition tasks. In this study, the energy of the signal will be partitioned into different resolutions followed by carrying out an examination on these indexes. Mathematically,

$$P_j = \frac{1}{N}\sum_k |w_{j,k}|^2 = \frac{\|w_j\|^2}{N_j}, \tag{11}$$

where $\|w_j\|$ is the norm of the expansion of the expansion coefficient w_j. Fig. 2a shows utterances in the time domain

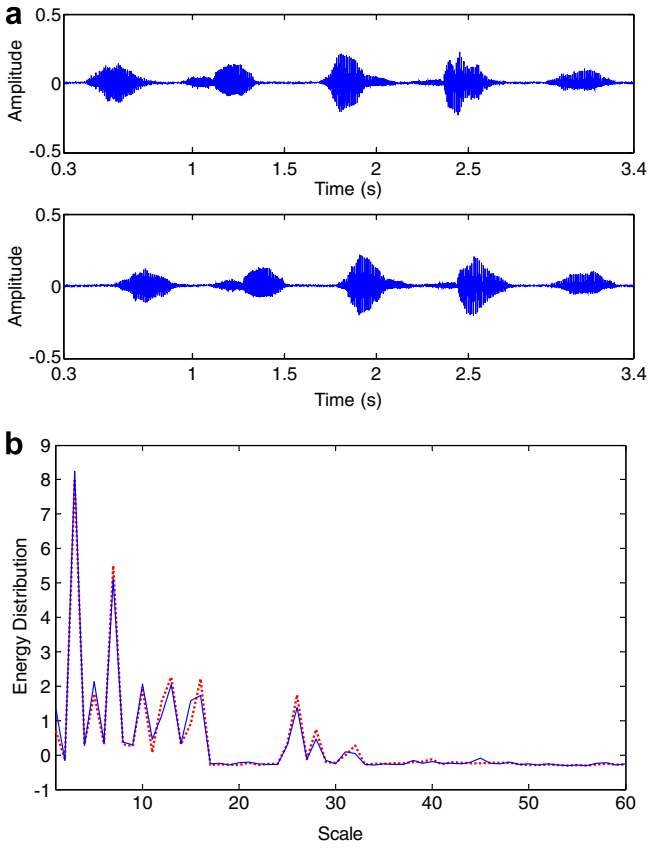

Fig. 2. (a) Utterances of a same speaker in the time domain with different time lag and (b) energy distributions of the given signals are mostly the same for the same speaker.

from a same speaker recorded with a different time lag. Fig. 2b shows the results of energy indexes computed from Fig. 2a, in which the energy distributions of the given signals are the same.

3. Speaker classifier using neural network

In the design of speaker identification scheme, a recognition method of the features based on DWT and WPT was chosen to evaluate the effectiveness of the selected feature sets for the identification system. In this study, the general regressive neural network (GRNN) is used. The GRNN was first proposed by Specht (1991). Fig. 3 shows the block diagram of the GRNN architecture. It is a one-passing learning algorithm, which can be used for estimation continuous variables such as some transient content in speech signal. It does not require an iterative training procedure to converge to the desired solution as in the back-propagation (BP) neural network.

By definition, the regression of a dependent variable y on an independent x estimates the most probable value for y, given x and a training set. The GRNN is a method for estimating the joint probability function of x and y in order to produce the estimated value of y, given only a training set. Assume that $f(x, y)$ represents the known joint continuous probability density function of a vector random variable, x, and a scalar random variable. Let X be a particular measured value of the random variable x. The conditional mean of y given X is given by

$$E[y|X] = \frac{\int_{-\infty}^{\infty} yf(X,y)\mathrm{d}y}{\int_{-\infty}^{\infty} f(X,y)\mathrm{d}y} \tag{12}$$

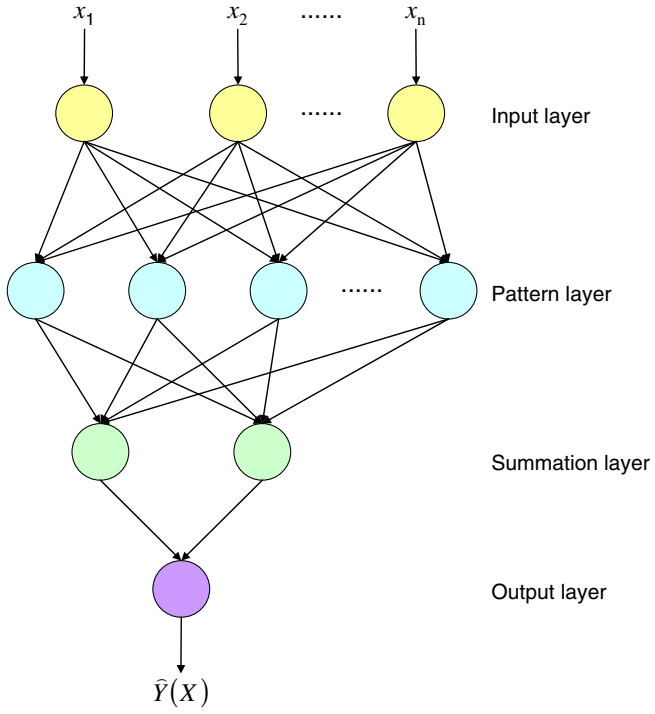

Fig. 3. Block diagram of GRNN architecture.

When the density $f(x, y)$ is not known, it must usually be estimated from the sample of observations of x and y. The probability estimator $f(X, Y)$ is based upon sample values X^i and Y^i of the random variables x and y, where n is the number of sample observations and p is the dimension of the vector variable x:

$$\hat{f}(X,Y) = \frac{1}{(2\pi)^{(p+1)/2}\sigma^{(}p+1)}\frac{1}{n}\times\sum_{i=1}^{n} \times \exp\left[-\frac{(X-X^i)^{\mathrm{T}}(X-X^i)}{2\sigma^2}\right]\exp\left[-\frac{(Y-Y^i)^2}{2\sigma^2}\right] \tag{13}$$

A physical interpretation of the probability estimate $\hat{f}(X, Y)$ is that it assigns the sample probability of width σ for each sample X^i and Y^i, and the probability estimate is the sum of those sample probabilities. Defining the scalar function

$$D_i^2 = (X-X^i)^{\mathrm{T}}(X-X^i) \tag{14}$$

and performing the indicated integration yields:

$$\hat{Y}(X) = \frac{\sum_{i=1}^{n} Y^i \exp\left[-\frac{D_i^2}{2\sigma^2}\right]}{\sum_{i=1} \exp\left[-\frac{D_i^2}{2\sigma^2}\right]} \tag{15}$$

When the smoothing parameter σ is made large, the estimated density is forced to be smooth and the limit becomes a multivariate Gaussian with covariance $\sigma^2 I$. On the other hand, a smaller value of σ allows the estimated density to assume non-Gaussian shapes, but with the hazard that wild points may have significant effect on the estimate.

4. Experimental investigation and results

4.1. Experimental arrangement

In order to evaluate the proposed method, a sound recording experiment was carried to verify the recognition performance by performing different forms of decomposition. First, utterances of a speaker was recorded by a microphone with a data acquisition system and then normalized by the proposed algorithm. The features were extracted using DWT and WPT. After feature extraction, the wavelet based energy indexes of the features were obtained. At the last stage, energy indexes were fed into the GRNN for identification. The experimental setup of the speaker identification system is shown in Fig. 4. The sampling rate was set at 16 kHz and the measured maximum frequency was 8 kHz. The recording apparatus consists of a microphone (PCB 130D20) and a data acquisition system (NI-6024E). The speech database was made up of 50 speakers including 25 female and 25 male speakers. Each speaker repeated an assigned sentence for 50 times, and there were five sentences in all. Experimental investigations were divided into four parts. The first

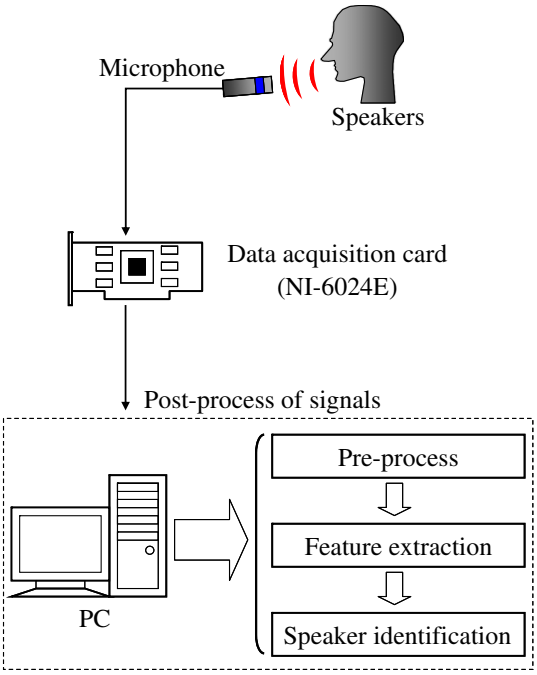

Fig. 4. Experimental setup of the speaker identification system.

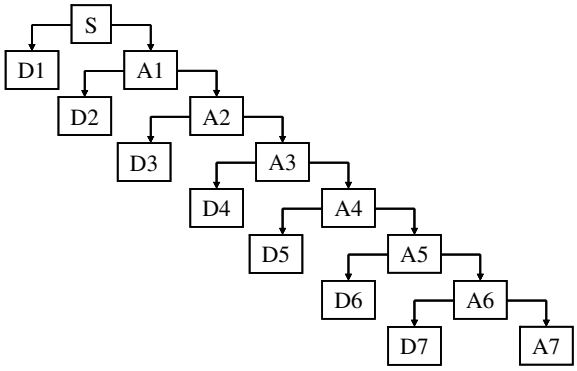

Fig. 5. Tree diagram of DWT.

part described in Section 4.2, is to evaluate feature sets from DWT. In Sections 4.3 and 4.4, feature sets were extracted by WPT, both in conventional way and Mel scale. In Section 4.5, an irregular decomposition based on the speech energy distribution is proposed. The final section states these changes working on the recognition performance and makes a comparable result among these methods.

4.2. Sub-band based parameters of DWT

DWT decomposes the data in a dyadic form, and the recursive decomposition only act on the low frequency content, named "approximation" (abbreviated as A). The high frequency content is preserved as "detail" (abbreviated as D). In this section, feature sets were acquired from 8-level decomposition. Fig. 5 is a typical wavelet tree, in which the decomposition is stuck by the above rules. Given the features for speaker identification, $D1 \sim D7$ and $A7$ are chosen as the inputs.

4.3. Multi-level decomposition by WPT

WPT is the extension version of DWT, but the distinction is that WPT performs recursive decomposition both on A and D. In this section, speech data will be decomposed into 5, 6 and 7 levels, respectively, by using WPT, i.e., there are 32, 64 and 128 sub-band feature sets for identification. The number of feature sets is doubled with further decomposition. The reason for comparing 64 WP feature sets with the methods in Sections 4.4 and 4.5 is that these methods have similar number on feature sets. The usage of 32 and 128 WP feature sets is for investigating the relationship between the recognition rate and the number of feature sets. Fig. 6 shows the complete decomposition of a different level WPT.

4.4. Wavelet packer feature sets spaced in Mel-scale

In this section, feature sets of WPT were spaced in a similar Mel scale form. Fig. 7 is a demonstration of bandpass filters in Mel scale. Previous studies demonstrated the recognition rates with fewer feature sets, but in this paper, the number of feature sets was extended to 60 to make a comparable result with the proposed method. In the experi-

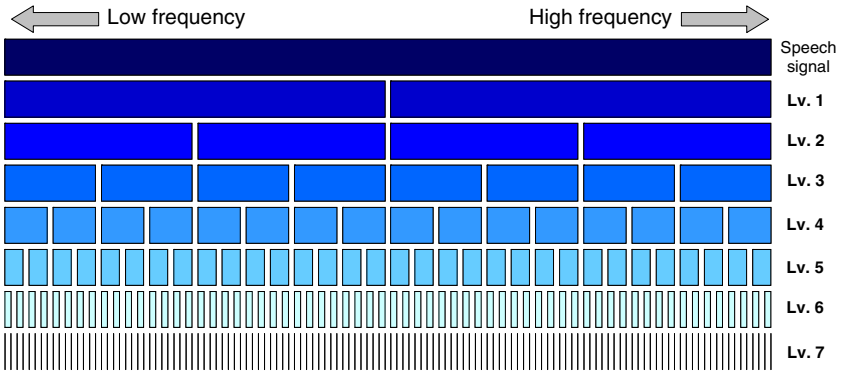

Fig. 6. Conventional WPT in different resolution.

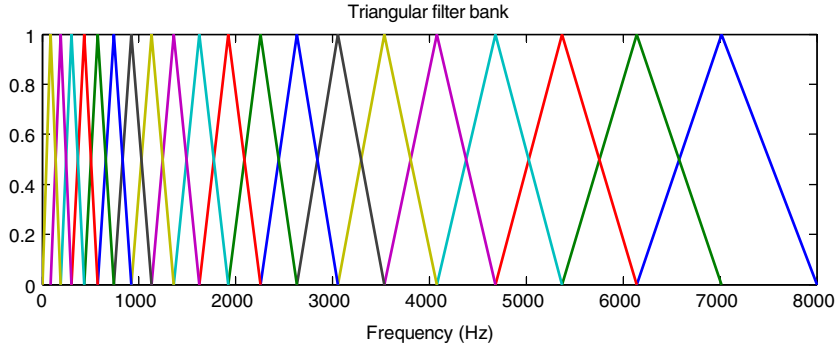

Fig. 7. Bandpass filters in Mel scale.

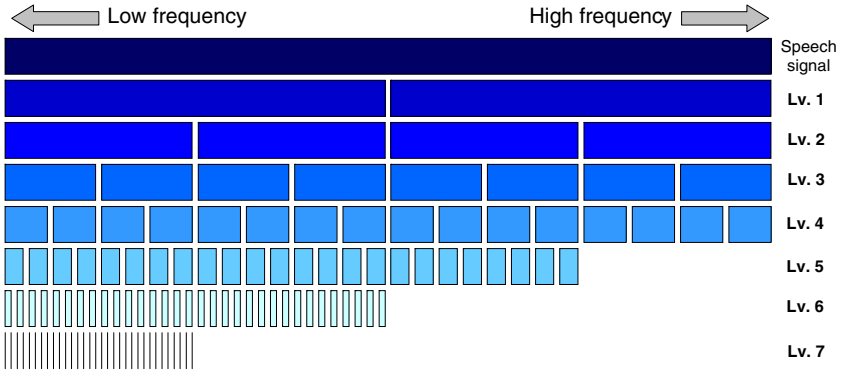

Fig. 8. WP feature sets spaced in Mel scale.

ment, a 7-level decomposition was applied on an interval of 1 Hz to 2 kHz; on the interval of 2 kHz to 4 kHz, a 6-level decomposition was applied, and so forth till the 8 kHz mark is reached. Fig. 8 shows the features sets of WPT spaced in Mel scale, which has similar form to previous studies (Farooq & Datta, 2001; Karam et al., 2000; Sarikaya et al., 1998; Torres & Rufiner, 2000), but the resolution is more specific.

4.5. Irregular decomposition based on energy distribution

In this section, the irregular decomposition based on speakers' energy distribution is proposed. In Section 4.3, experimental results showed that a part of energy is conspicuous in the specific frequency region. Fig. 9 shows the energy distribution of conventional WPT which illustrates no matter male or female speakers, voice energies appeared not only in the low frequency region, i.e., it reveals that the existence of the voice energy in higher frequency region. Hence the decomposition concentrated on these regions; in the other hand, the resolution on lesser energy regions was decreased. In Table 1 and Table 2, the center frequencies of WP feature sets spaced in a similar Mel scale and the proposed approach are given for the purpose of comparison. If the consideration included the complexity of the system or time spent in the stage, then the number of parameters is significantly important, because they directly affect the computational load. Using the

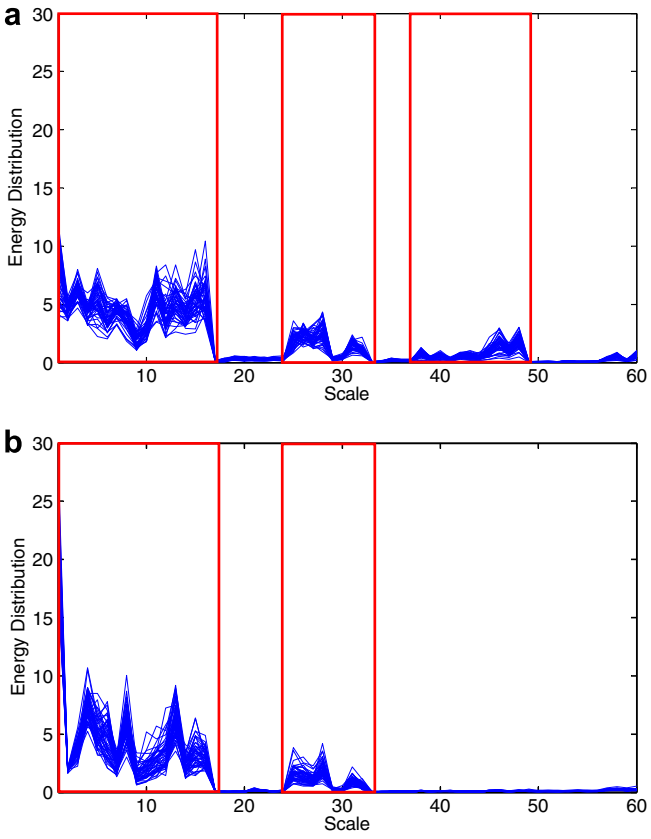

Fig. 9. Energy distribution of speakers: (a) male and (b) female.

Table 1
Center frequency of WPT filters spaced in Mel scale

Filter	Center frequency (Hz)	Filter	Center frequency (Hz)	Filter	Center frequency (Hz)
1	31	21	1281	41	3064
2	94	22	1344	42	3188
3	156	23	1406	43	3314
4	219	24	1469	44	3439
5	281	25	1531	45	3564
6	344	26	1594	46	3689
7	406	27	1656	47	3814
8	469	28	1719	48	3939
9	531	29	1781	49	4125
10	594	30	1844	50	4375
11	656	31	1906	51	4625
12	719	32	1969	52	4875
13	781	33	2063	53	5125
14	844	34	2188	54	5375
15	906	35	2313	55	5625
16	969	36	2438	56	5875
17	1031	37	2563	57	6250
18	1094	38	2688	58	6750
19	1156	39	2813	59	7250
20	1219	40	2939	60	7750

Table 2
Center frequency of WPT filters with irregular decomposition

Filter	Center frequency (Hz)	Filter	Center frequency (Hz)	Filter	Center frequency (Hz)
1	47	20	1594	39	3219
2	94	21	1656	40	3281
3	156	22	1719	41	3344
4	219	23	1781	42	3406
5	281	24	1844	43	3469
6	344	25	1906	44	3531
7	406	26	1969	45	3594
8	469	27	2125	46	3656
9	531	28	2375	47	3718
10	594	29	2563	48	3781
11	656	30	2656	49	3844
12	719	31	2719	50	3906
13	781	32	2781	51	3969
14	844	33	2844	52	5625
15	906	34	2906	53	5875
16	969	35	2969	54	6250
17	1125	36	3031	55	6750
18	1375	37	3094	56	7250
19	1531	38	3156	57	7750

method in this section, the computational load can be eased off without sacrificing the recognition rate. Fig. 10 is the demonstration of the proposed WP feature sets with irregular decomposition.

4.6. Experimental results

Table 3 shows the experimental results of all approaches used in the experimental investigation. The recognition rate of DWT with eight feature sets reached the lowest value. In WP experiments, it was found that the recognition rates improved upon increasing the number of feature sets. However, the improvement implies a tradeoff between the recognition rate and extracting time. It is seen that the recognition rate has improved from 71.6% to 97.8%, but the number of feature sets has increased four times, from 32 to 128. On the other hand, the growth of extracting time indicated that the computational load has been burdened. Moreover, the recognition rate improved only 1% when the number of feature sets is doubled in the 7-level WPT. From the viewpoint of real applications, it may be uneconomic. The cause of this phenomenon can be attributed to the fact that the background noise is being detailed while the decomposition is further detailed.

WP feature sets spaced in Mel scale had good performance on recognition rate, and they interpreted the achievements of Mel cepstral theory. The proposed irregular decomposition method shows better recognition rate than conventional WPT and WPT in Mel scale. This is because the energy in speakers' pronunciation concentrates in some specific region, and the decomposition will be detailed on them. In addition, for the lesser-energy regions, the resolution of WPT will be decreased. In this paper, the recognition rate results from various feature extraction methods are compared; the effectiveness of the proposed approach is demonstrated using experimental data. These results pointed out that the proposed approach could

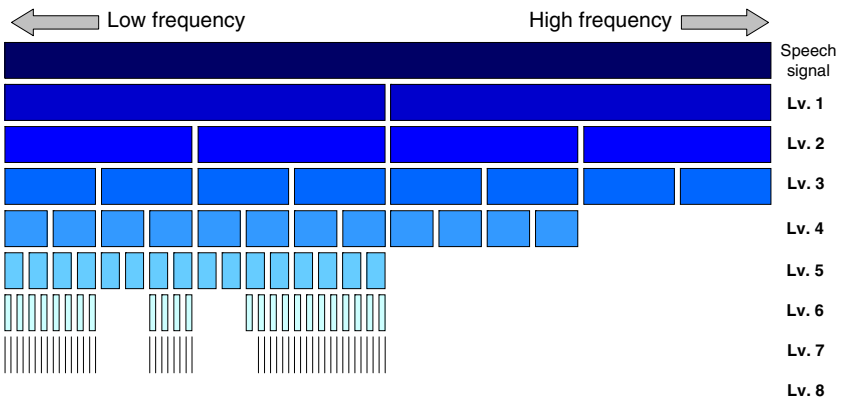

Fig. 10. New proposed WPT parameters with irregular decomposition.

Table 3
Comparison of all feature extracting approaches

Approaches	Number of feature sets	Recognition rate (%)	Extracting time (s)
DWT	8	70.8	1.02
5-level WPT	32	71.6	2.27
6-level WPT	64	94.6	3.42
7-level WPT	128	97.8	6.67
WPT in Mel scale	60	94.4	3.73
Proposed irregular decomposition of WPT	57	96.6	3.61

achieve good recognition rate with fewer feature sets with less computational load.

5. Conclusions

An expert system for speaker identification using irregular decomposition of WPT is presented in this paper. Apart from DWT, conventional WPT and WPT in Mel Scale, the irregular decomposition method has an improved recognition rate without increasing the extracting time. Although the experimental results show that the feature sets of these approaches can be used for speaker identification, the proposed approach worked well with fewer feature sets and the accuracy rate is superior to other methods with similar number of feature sets. It also indicated that the energy distribution of speakers' utterances is uneven; hence, the analysis can be focused on energy-centralized part to prevent unnecessary operations. In the future, the first task is to further reduce the number of feature sets and fit for any speaker identification applications.

Acknowledgements

The study was supported by the National Science Council of Taiwan, Republic of China, under Project No NSC-96-2221-E-018-015.

References

Adami, A. G., & Barone, D. A. C. (2001). A speaker identification system using a model of artificial neural networks for an elevator application. *Information Sciences, 138*, 1–5.

Alkhaldi, W., Fakhr, W., & Hamdy, N. (2002). Multi-band based recognition of spoken Arabic numerals using wavelet transform. In *Proceedings of the 19th national radio science conference* (pp. 224–229). Egypt.

Avci, E., & Akpolat, Z. H. (2006). Speech recognition using a wavelet packet adaptive network based fuzzy inference system. *Expert Systems with Applications, 31*, 495–503.

Avci, E., Hanbay, D., & Varol, A. (2006). An expert discrete wavelet adaptive network based fuzzy inference system for digital modulation recognition. *Expert System with Applications, 33*, 582–589.

Behroozmand, R., & Almasganj, F. (2007). Optimal selection of wavelet-packet-based features using genetic algorithm in pathological assessment of patients' speech signal with unilateral vocal fold paralysis. *Computers in Biology and Medicine, 37*, 474–485.

Davis, S. B., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustic Speech and Signal Processing, 28*, 357–366.

Farooq, O., & Datta, S. (2001). Mel filter-like admissible wavelet packet structure for speech recognition. *IEEE Signal Processing Letters, 8*(7), 196–198.

Fonseca, E. S., Guido, R. C., Scalassara, P. R., Maciel, C. D., & Pereira, J. C. (2007). Wavelet time-frequency analysis and least squares support vector machines for the identification of voice disorders. *Computers in Biology and Medicine, 37*, 571–578.

Goupillaud, P., Grossman, A., & Morlet, J. (1984). Cycle-octave and related transforms in seismic signal analysis. *Geoexploration, 23*, 85–102.

Gowdy, J. N., & Tufekci, Z. (2000). Mel-scaled discrete wavelet coefficients for speech recognition. In *Proceedings of the acoustics speech and signal processing, ICASSP '00. IEEE international conference*. Istanbul.

Kanedera, N., Arai, T., Hermansky, H., & Pavel, M. (1999). On the relative importance of various components of the modulation spectrum for automatic speech recognition. *Speech Communication, 28*, 43–55.

Karam, J. R., Phillips, W. J., & Robertson, W. (2000). New low rate wavelet models for the recognition of single spoken digits. In *Proceedings of the electrical and computer engineering, 2000 Canadian conference*. Canada.

Kotnik, B., Kacic, Z., & Horvat, B. (2003). The usage of wavelet packet transform in automatic noisy speech recognition systems. In *Proceedings of the IEEE EUROCON 2003* (pp. 131–134). Slovenia.

Lou, X., & Loparo, K. A. (2004). Bearing fault diagnosis on wavelet transform and fuzzy inference. *Mechanical System and Signal Processing, 18*, 1077–1095.

Louis, A. K., Maass, D., & Rieder, A. (1997). *Wavelets-theory and applications*. Hoboken, NJ: Wiley.

Lung, S. Y. (2006). Wavelet feature selection based neural networks with application to the text independent speaker identification. *Pattern Recognition, 39*, 1518–1521.

Mashao, D. J., & Skosan, M. (2006). Combining classifier decisions for robust speaker identification. *Pattern Recognition, 39*, 147–155.

Sarikaya, R., & Hansen, J. H. L. (2000). High resolution speech feature parametrization for monophone-based stressed speech recognition. *IEEE Signal Processing Letters, 7*(7), 182–185.

Sarikaya, R., Pellom, B. L., & Hansen, J. H. L. (1998). Wavelet packet transform features with application to speaker identification. In *Proceedings of the IEEE Nordic signal processing symposium* (pp. 81–84). Denmark.

Specht, D. F. (1991). A general regression neural network. *IEEE Transactions on Neural Networks, 2*(6), 568–576.

Sroka, J. J., & Braida, L. D. (2005). Human and machine consonant recognition. *Speech Communication, 45*, 401–423.

Tajima, K., Port, R., & Dalby, J. (1997). Effects of temporal correction on intelligibility of foreign-accented English. *Journal of Phonetics, 25*, 1–24.

Torres, H. M., & Rufiner, H. L. (2000). Automatic speaker identification by means of Mel cepstrum, wavelets and wavelets packets. In *Proceedings of the 22nd annual EMBS international conference* (pp. 978–981). Chicago.

Zhang, X., & Jiao, Z. (2004). Speech recognition based on auditory wavelet packet filter. In *Proceedings of the signal processing, ICSP '04. Seventh international conference*. Beijing.