13

The Research of Noise-Robust Speech Recognition Based on Frequency Warping Wavelet

Xueying Zhang and Wenjun Meng
Taiyuan University of Technology, Taiyuan University of Science & Technology
China

1. Introduction

The main task of speech recognition is to enable computer to understand human languages (Lawrence, 1999; Jingwei et al., 2006). This makes it possible that machine can communicate with human. Usually, speech recognition includes three parts: pre-processing, feature extraction and training (recognition) network. In this paper, the speech recognition system is described as Fig. 1. It consists of filter bank, feature extraction and training (recognition) network. The function of filter bank is dividing speech signal into different frequency band to be good for extraction feature. The good feature can improve the system recognition rate. The training (recognition) network trains (recognizes) the feature vectors according to feature mode and outputs recognition results.

The research on noise-robust capability of speech recognition system is a difficult problem that has been limiting the practical application of the speech recognition system (Tianbing et al., 2001). Because human ear has strong noise-robust capability, it is very important to abstract the features of fitting auditory characters of human ear for improving system noise-robust performance. The warping wavelet overcomes the disadvantage that the common wavelet divides frequency band in octave band and it is more suitable to the auditory characters of human ear. Bark wavelet is a warping wavelet that divides frequency band according to critical band (Qiang et al., 2000). At the same time, MFCC (Mel Frequency Cepstrum Coefficients) (Lawrence, 1999) and ZCPA (Zero-Crossing with Peak Amplitude) (Doh-suk et al., 1999) features themselves have noise-robust performance. HMM is classical recognition network, and wavelet neural network is also popular recognition network (Tianbing et al., 2001). So considering above three parts of speech recognition system, the paper used the two kinds of filters: FIR filter and Bark wavelet filter; two kinds of features:

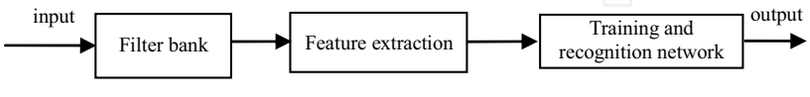


Figure 1. The speech recognition system

Source: Robust Speech Recognition and Understanding, Book edited by: Michael Grimm and Kristian Kroschel,
ISBN 987-3-90213-08-0, pp.460, I-Tech, Vienna, Austria, June 2007

MFCC and ZCPA; two kinds of recognition networks: HMM and WNN (Wavelet Neural Network). Limited by article length, the paper selected a few of composite modes, described their principles and presented experimental results. The three parts of Fig.1 have effect on one another. Their different combination can get different results. In practical application we can select optimum combination mode. The combination modes selected in the research are listed in Table 1.The article includes five sections: the first section is introduction; the second section describes the principle of ZCPA and implementation method of combination mode 1 and 2 in Table 1; and the third section describes the principle of Bark wavelet and implementation method of combination mode 3 and 4 in Table 1; the fourth section is the experimental results and discussion; the fifth section is conclusions.

Combination modes index	Filter	Feature	Training and recognition network
1	FIR	ZCPA	HMM
2	FIR	ZCPA	WNN
3	Bark wavelet	ZCPA	HMM
4	Bark wavelet	MFCC	HMM

Table 1. The combination modes of speech recognition system in the paper

2. The Principle of ZCPA and Implementation Methods of Combination Mode 1 and 2

2.1 The principle of ZCPA

The human auditory system consists of outer ear, middle ear and inner ear. Speech signals are transformed into mechanical vibrations of the eardrum at the outer ear, and then are transmitted to the cochlea of the inner ear through the middle ear. The role of the middle ear is known as impedance matching between the outer ear and the inner ear. The speech signals are mainly processed in the inner ear, especially in the cochlear of the inner ear. The basilar membrane of the cochlear has the function of frequency choice and tune. Speech signals transmitted through the oval window at the base of the cochlear are converted into travelling waves of the basilar membrane. The site of maximum excursion of the travelling wave on the basilar membrane is dependent on frequency. High frequencies show maximum excursion near the base while low frequencies near the apex. Frequencies are distributed according to logarithm relationship along the basilar membrane over 800Hz. The frequency-position relationship can be expressed as Equation (1) (Doh-suk et al., 1999):

$$F = A(10^{ax} - 1) \tag{1}$$

Where F is frequency in Hz, and x is the normalized distance along the basilar membrane with a value of from zero to one. A and a are the constants. A=165.4 and a=2.1.

The cochlear takes a very important role in the auditory system, which can apperceive and transmit the speech signals. In fact, with the function of series-parallel conversion, it corresponds to a bank of parallel band-pass filters. Signals imported by series are decomposed and exported by parallel. Then it provides evidence to some extent for cochlear filter model.

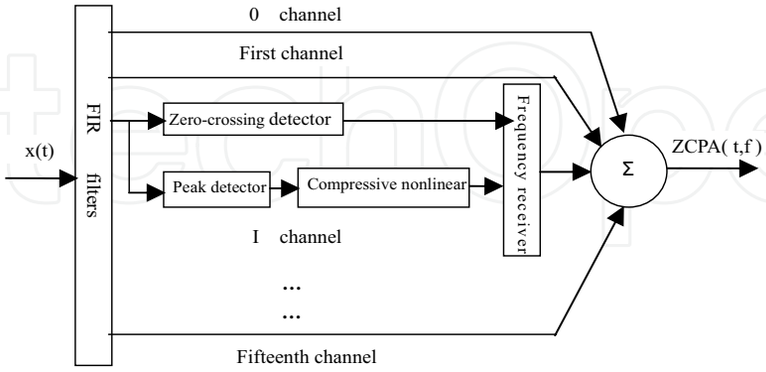


Figure 2. ZCPA feature extraction scheme

Fig. 2 shows the feature extraction block scheme of ZCPA (Doh-suk et al., 1999). This system consists of a bank of band-pass filters, zero-crossing detector, peak detector, nonlinear compression and frequency receiver. The filter bank consists of 16 FIR band-pass filters that are designed to simulate the basilar membrane of cochlear. And zero-crossing detector, peak detector and nonlinear compression simulate the auditory nerve fibers. The frequency information is obtained by computing the zero-crossing intervals of speech signal from zero-crossing detector. And the intensity information is obtained by detecting peak amplitudes between the intervals and making nonlinear amplitude compression from peak detector and the nonlinear compression. The frequency receiver combines the frequency with the peak information. Finally, this information is compounded to form the feature output of speech signals.

2.1.1 The design of the filters

Because the basilar membrane of cochlear corresponds to a bank of parallel band-pass filters, we can choose 16 points along the basilar membrane and get 16 FIR filters, whose frequencies change from 200Hz to 4000Hz. The centre frequency of every filter can be obtained from the Equation (1). And bandwidths are set to be proportional to the equivalent rectangular bandwidth (ERB) (Oded, 1992; Doh-suk et al., 1999).

$$ERB = 6.23F^2 + 93.39F + 28.52 \quad (2)$$

Where F is the centre frequency of each filter in Hz. Table 2 shows the centre frequencies and the bandwidths of each FIR filter, f_i is the centre frequency, Δf_i is the bandwidth.

Filter No.	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
f_i(Hz)	200	264	340	429	534	657	802	1011	1172	1408	1685	2011	2395	2845	3376	4000
Δf_i(Hz)	46	53	61	71	82	95	111	129	151	176	206	241	283	331	389	456

Table 2. The relationship between centre frequency and bandwidth of FIR filters

2.1.2 The theory of the zero-crossing

When the two adjacent samples have the different sign, we name this phenomenon as zero-crossing. And the up-going zero-crossing is that the current sample value is bigger than zero and the former sample value is smaller than zero. Because the number of the zero-crossing is different for different frequency's signals, that is to say, high frequency signals have more zero-crossings than low frequency signals. So the up-going zero-crossing rates can reflect the frequency information of the speech signals.

2.1.3 Extraction of the intensity information

To simulate the relation between the phase-locking degree of the auditory nerve fibers and the intensity of the stimulus, the maximal peak value between the two adjacent zero-crossing samples need be detected. This relation can be described by a monotonic function. That is expressed as Equation (3) :

$$g(x) = \lg(1.0+20x) \tag{3}$$

Where x is the maximal peak value, and g(x) is the result of the nonlinear compression.

2.1.4 Frequency receiver

The frequency band of speech signal is mainly between 200Hz and 4000Hz. For obtaining low dimension and high effect feature parameters, we used ERB-rate (Oded, 1992) scale to divide the frequency band into 16 bands. And each band is named as a frequency bin. By computing the up-going zero-crossing number of the signal duration, it can be known that the signal duration belong to some frequency bins. The frequency information included in the speech signals can be reflected consequently.The frequency information and the intensity information across all channels are combined to obtain the frequency histograms (i.e. frequency bins) by the frequency receiver (Oded, 1994). The course can be described by Equation (4).

$$zcpa(m,\ i) = \sum_{k=1}^{N_{ch}} \sum_{l=1}^{Z_k-1} \delta_{ij_l} g(p_{kl}),\ 1 \le i \le N \tag{4}$$

Where, zcpa(m, i) are output features , m is frame index , i is the index of frequency bin. And N is the number of frequency bin, which is equal to 16. N_{ch} is the number of FIR filters , here N_{ch}=16. Each of filters is a signal processing channel and k is its index. Z_k denotes the number of the up-going zero-crossings at frame m and channel k. l is the index of up-going zero-crossings, l = 1, 2, …, Z_k-1. And p_{kl} denotes the peak amplitude between the l-th and (l+1)-th up-going zero-crossings. δ_{ij_l} is the Kronecker delta (When i= j_l, δ_{ij_l} =1; When i ≠ j_l, δ_{ij_l} =0). j_l is the frequency bin index mapped by the interval between the l-th and(l+1)-th up-going zero-crossings , $1 \le j_l \le N$.

2.1.5 Time and amplitude normalization

The feature form obtained is zcpa(m, i). In most case, because m is of different values, the number of feature vectors obtained is also different. Thus, it is necessary to normalize time. In the paper, the dimension of feature vectors is 64×16 after time normalization. The time normalization method used is Non-Linear Partition method (Zhiping et al., 2005). It is also necessary to normalize amplitude for latter processing convenience.

2.2 The implementation method of combination mode 1: FIR+ZCPA+HMM

2.2.1 The experimental principle

An isolated word speech recognition system with FIR filter, ZCPA feature and discrete HMM is implemented on Windows operating system in C++ language in this paper. The system diagram is shown in Fig. 3. In the experiment, speech data with different SNRs of 50 words 16 persons is used (including data of 15dB, 20dB, 25dB, 30dB and clean). Each person says each word 3 times. The model is trained by speech data (a certain SNR) of 9 persons and the recognition is carried out by speech data (under the same SNR) of other 7 persons, so the recognition result is obtained under this SNR. The parameters of HMM for each word is trained by 27 samples (9 persons×3 times), and the number of test data file depends on the number of word using in the experiment.

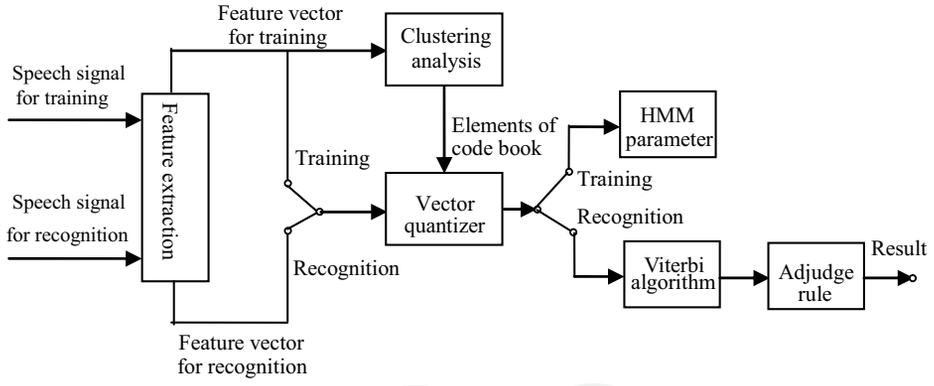


Figure 3. Isolated word speech recognition diagram based on HMM

2.2.2 The experimental step

Step 1: Feature extraction.

Two problems commonly need to be solved: one is to extract representative appropriate feature parameters from speech signal; the other is to compress properly data. The feature parameter of speech is extracted once each frame, and each frame feature parameter commonly constitutes a vector, so speech feature is a vector sequence. At the front end of the system, sampling rate of speech signal is 11.025 kHz, frame length is 10ms, sampling point number is 110, and frame shift is 5ms. Uniform 64×16 (or 1024) dimensional feature vector sequence is obtained through time and amplitude normalizing.

Step 2: Vector quantization.
The extracted feature need be quantized in vector way to satisfy the demands of discrete HMM adopted as recognition network. Vector quantization is an effective data compressing technology, and the codebook is obtained by LBG clustering method in this article. The feature extracted from speech signal becomes speech pattern after data compressing. Obviously, that speech pattern is of representative is one of the main factors to improve speech recognition rate.
Firstly, all the features of training words form a large set of speech feature vectors (for example, the number of vectors of 50 words is 50×27× (1024/4)), which is used to train codebook. In the system, the size of the codebook is 128, and the dimension of the code word is 4. These vectors are distributed to 128 classes, and each class has a tab (from 1 to 128). Secondly, 1024 feature values of each word can form 256 feature vectors with 4 dimensions to enter vector quantizer that has been trained. According to the nearest neighbourhood rule, 256 vectors are quantized and each vector is represented by the tab of class which the vector belongs to. Finally, the tab of codeword replaces former vector and becomes speech pattern of each word as input signal for the next processing.
Step 3: Training HMM.
After above processing, the feature tab of the word is obtained as the input sequence $\{O=O_1, O_2, \ldots, O_T\}$ of discrete HMM. For discrete HMM, each word is represented by a HMM ($\lambda = (A, B, \pi)$) (Lawrence, 1999), and is trained by 27 sampling sequences. The HMM is from left to right mode without span, and each word model has 5 states (shown in Fig.4). After trained, the state transfer matrix of a certain word is shown in Fig.5. Because the size of codebook is 128, the number of observation sign is M=128.
About the assumption of initial value, three parameters in this system are set as equal probability values. Training method is classical Baum-Welch algorithm, and training terminate condition is that according to the logarithm of probability for one pronunciation, the absolute value changing before and after the revaluation of the parameter is less than a certain threshold value (in the experiment the threshold is 0.01).

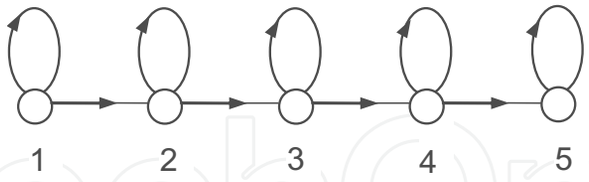


Figure 4. From left to right model without span

$$A = \begin{bmatrix} 0.97922 & 0.02078 & 0 & 0 & 0 \\ 0 & 0.985461 & 0.014539 & 0 & 0 \\ 0 & 0 & 0.97241 & 0.02759 & 0 \\ 0 & 0 & 0 & 0.969138 & 0.030862 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Figure 5. State transfer matrix

Step 4: The recognition of word.
After previous training, each word has its own model parameter. It is similar to quantization of training data set that feature vector used in test (speech data of another 7 persons in a certain SNR) enters vector quantizer formed in step 2. Thus, each word is quantized into a codeword tab sequence, which is used in computing probability through all the parameters for HMM of all the words, and Viterbi algorithm is used in this process. The criterion in this algorithm is searching the single best state sequence, in other words, making condition probability P (Q | O, λ) maximum, which also equals to making P (Q, O| λ) maximum (Shuyan et al., 2005). Therefore, the model corresponding to the maximum probability is the recognition result. The recognition rate is the ratio between the number of correctly recognized words and the number of all the test words.
Table 5 shows the recognition results of different words using the combination of FIR+ZCPA+HMM under different SNRs.

2.3 The implementation method of combination mode 2: FIR+ZCPA+WNN

2.3.1 The structure of WNN

In the structure of FIR+ZCPA+WNN, the training and recognition network uses WNN instead of HMM , other condition is unchanged, as shown in Fig. 6. The theory base of wavelet neural network is the reconstructing theory of wavelet function. It ensures the continuous wavelet basis has the ability of approximating any function, so we can take place the Sigmoid or Gaussian function of neural network by wavelet basis to construct a new feed-forward neural network. This system constructed the WNN according to the wavelet basis fitting. It is known to us that a signal function f(t) can be fitted via linear combination of selected wavelet basis (Zhigang et al., 2003):

$$\hat{f}(t)=\sum_{k=1}^{K} w_k \phi\left(\frac{t-b_k}{a_k}\right) \tag{5}$$

Where b_k is position factor, a_k is scale factor and k is the number of basis function. The network topology can refer to Fig. 7.

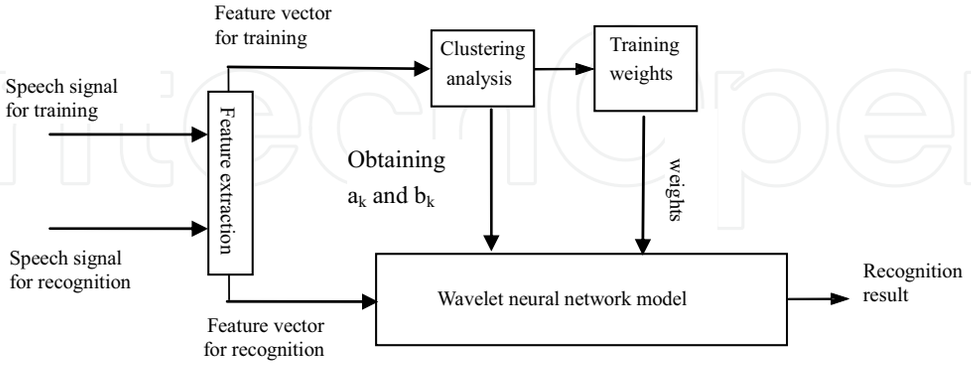


Figure 6. Diagram of speech recognition based on wavelet neural network

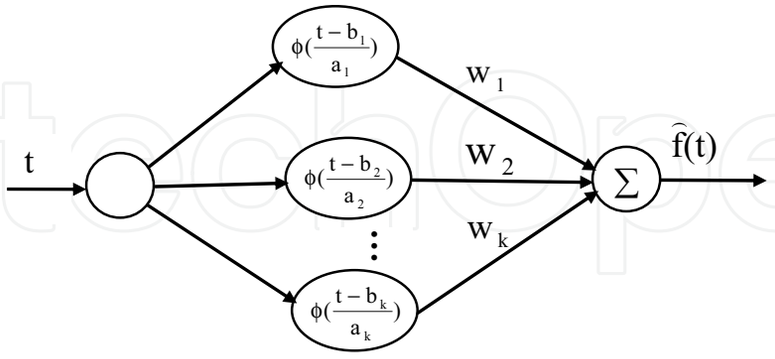


Figure 7. Wavelet neural network topology with single input and single output

Where, a_k and b_k are variables, only w_k acts as the weight between the hidden and output layer, and $\phi\left(\frac{t-b_k}{a_k}\right)$ can be regarded as the output value of input nodes. The network structure is similar to the traditional forward multi-layer perceptron; the key difference depends on the hidden layer function. The system this paper presented is a three-layer neural network with multi-input and multi-output. The inputs are the feature parameters of every word, and their dimension is 1024, so the number of input nodes is 1024. The number of hidden nodes depends on words number recognized. Here, 10 nodes for 10 words and 50 nodes for 50 words.

It is a total connection with weight value 1 between input layer and hidden layer, and the same as that between hidden layer and output layer, while they have weights. The nodes of output layer equal to the classification number of words. So this system has same nodes in hidden and output layer. The form of the basis function of every hidden node is uniform, while the scale and position parameters vary with nodes. Strictly speaking, this network is not based on the wavelet mathematical analysis; in fact, it was used to approximate function in wavelet combination of certain form. In this system, it was for word recognition. As for the selecting method of wavelet basis function, there are no uniform rules in theories (Johnstone, 1999). Generally, it can be determined by experience and practical application. Here, Mexican Hat wavelet was selected. The experiment showed that Mexican Hat wavelet certainly has excellent characteristics in recognition.

2.3.2 The determining of WNN parameters

The parameters to be determined in the system are number of hidden nodes, scale factor a_k, position factor b_k and connection weights w_k between all hidden and output nodes. Estimation of the number of hidden nodes also has no uniform rules in theories. In this paper, the hidden nodes number equal to that of the output layer, i.e. the word classification number. Otherwise, a bias should be added to hidden layer, it has fixed value of 1. This bias factor also should be connected to all the output nodes in order to estimate weight values.

Several methods can be used to estimate the three parameters of wavelet network, such as BP network training method and orthogonal least square which optimized these three parameters simultaneously. The network topology of this paper makes it possible that the training of scale and position can be separated from the weights training. One of common

methods of estimating a_k and b_k is clustering. The clustering algorithm just assigns the given vectors into several finite classes according to certain distortion measure. However, this method does not take full advantage of the information of training samples. The clustering algorithm of K-Means has been used in this experiment to estimate the parameters, but the recognition results are not satisfying.

Because the given training samples have involved the corresponding classification information of every training feature, this information can be used to estimate the position parameter b_k. The hidden exciting function of wavelet network is a local function; it has strong approximation ability for the function with big difference in localness. The same as the feature with big difference, they can be classified properly. In recognition network, we hope the output of all training samples corresponding to a certain word via the hidden node which is determined by its position factor can get the biggest value. In other words, the more adjacent to position factor b_k, the bigger the output of the k-th node will be. Hence, for all the training samples corresponding to a certain word, their centroid can be calculated to be a position factor. Once a position factor has been estimated, there will be a scale factor under the condition of corresponding to it. Here Equation (6) was used to calculate scale factor knowing position factor (Musavi et al., 1992):

$$a_k = \sqrt{\frac{1}{1+\sqrt{2}} \sum_{k=1}^{K} \left\| x_k - b_k \right\|^2} \tag{6}$$

Where x_k is feature vector, b_k and x_k have same dimension.

LMS method was adopted to train the weights between hidden and output layer in this paper. Training weight by LMS only needs to compute several matrix multiplications, so it need less time to train. And the hidden nodes can be added to meet the practical requirement. The added nodes will not have an obvious effect on the training time. In this system, the number of input nodes is 1024. Generally, wavelet network with high dimension will lead to a "dimension disaster", which means with the increasing of dimension of input and training sample, the network converge speed will descend severely. In this paper, the calculating of position and scale factor was separated from weights training, so it will avoid effectively the "dimension disaster" because of high dimension.

2.3.3 The experimental steps

Step 1: ZCPA feature extraction.

After pre-processing speech signal we obtained ZCPA feature with 1024 dimension of every word. It can be directly inputted to following wavelet network.

Step 2: Confirming network structure.

The node number of input layer equals feature vector dimension, or 1024. The node number of hide layer or output layer equals classification words number. Hidden layer sets a bias node which output value is 1. It was also connected to all nodes of output layer to take part in weight training.

Step 3: Confirming position factor and scale factor.

Input features were divided into N classes by using clustering method supervised. Where, N is the classification number of words. For each classification, the position factor was obtained by calculating the centroids of all training samples of classification. And corresponding scale factor was calculated by Equation (6).

Step 4: Calculating weight values.

Using LMS method to calculate the weight values from hidden layer to output. Because LMS method has not the course of iterative operation, its convergence speed is very fast and it fits to real-time speech recogintion system.

Table 5 shows also the recognition results of different words using the combination of FIR+ZCPA+WNN under different SNRs.

3. The Principle of Bark Wavelet and Implementation Methods of Combination mode 3 and 4

3.1 The principle of Bark wavelet

The binary wavelet used commonly divides frequency band in octave band way (Gowdy et al., 2000). This does not fit entirely to the auditory character of human ear to speech. Bark wavelet was put forward on the basis of hearing perception, so it should have better function than binary wavelet.

The basilar membrane of the cochlear has the function of the frequency choice and tune. For different centre frequencies, the signals of corresponding critical frequencies band can arouse the different place librations of the basilar membrane. So from 20Hz to 16kHz, these frequencies can divide into 24 bands. The different frequency speech signals of the same place of the basilar membrane are added to evaluate, that is to say, the perception of the human auditory system to speech frequency is a nonlinear mapping relation with actual frequency. So this introduces the conception of the Bark scale, Traunmular (Zhiping et al., 2005) presents the relation of the linear frequency and Bark frequency. That is:

$$b = 13\arctan(0.76f) + 3.5\arctan(f/7.5)^2 \tag{7}$$

Where b is Bark frequency, and f is the linear frequency in Hz.

The basic thought of constructing Bark wavelet is: firstly, because of the same importance of the time and frequency information in the speech analysis, wavelet mother function selected should satisfy time and bandwidth product least; secondly, for being consistent with conception of the frequency group, mother wavelet should has the equal bandwidth in the Bark domain; furthermore, their bandwidth are the unit bandwidthes, namely 1 Bark (Qiang et al., 2000).

According to the above analysis, the formation of wavelet function selected is Equation (8) in the Bark domain.

$$W(b) = e^{-c_1 b^2} \tag{8}$$

Furthermore, when the bandwidth is the 3dB, the constant c_1 is selected 4ln2. It is easy to prove that

$$\int_{-\infty}^{\infty} \frac{\left|e^{-c_1 b^2}\right|^2}{b} db < \infty \tag{9}$$

Simultaneously, since the Equation (7) is a monotonic function, the transformation from the Bark frequency to linear frequency do not influence the Equation (8). Therefore Bark wavelet

satisfies "admissible condition", that is to say, Bark wavelet transform can perfectly reconstruction.

Supposing the speech signals analyzed is s(t), whose linear frequency bandwidth satisfies $|f| \in [f_1, f_2]$, and corresponding Bark frequency bandwidth is $[b_1, b_2]$. Thus wavelet function in Bark domain can be defined as:

$$W_k(b)=W(b-b_1-k\Delta b),\ k=0, 1, 2, \ldots, K-1 \tag{10}$$

Where , Δb is the translation step-length and k is the scale parameter. According to the equal bandwidth principia, there has $\Delta b = \frac{(b_2 - b_1)}{K-1}$.

Then substitute Equation (8) into (10) , we can get

$$W_k(b) = e^{-4\ln 2(b-b_1-k\Delta b)^2} = 2^{-4(b-b_1-k\Delta b)^2},\ k = 0,1,2\cdots\cdots K-1 \tag{11}$$

Then substitute Equation (7) into (11), so in the linear frequency the Bark wavelet function can be described as:

$$W_k(f) = c_2 \cdot 2^{-4[13\arctan(0.76f)+3.5\arctan(f/7.5)^2-(b_1+k\Delta b)]^2} \tag{12}$$

In the Equation (12), c_2 is the normalization factor, and c_2 can be obtained through Equation

$$c_2 \sum_{k=0}^{K-1} W_k(b) = 1,\ 0 < b_1 \le b \le b_2 \tag{13}$$

From Equation (12), in the frequency domain Bark wavelet transform can be expressed as:

$$s_k(t) = \int_{-\infty}^{\infty} S(f) \cdot W_k(f) \cdot e^{j2\pi f} df \tag{14}$$

Where, S(f) is the spectrum of speech signal s(t) analyzed, $s_k(t)$ is the signal of the k-th channel, which had been transformed by Bark wavelet.

Notice that the Equation (13) is also correct for linear frequencies, so there has

$$\sum_{k=0}^{K-1} s_k(t) = \sum_{k=0}^{K-1} \int_{-\infty}^{\infty} W_k(f) \cdot S(f) \cdot e^{j2\pi ft} df = \int_{-\infty}^{\infty} \sum_{k=0}^{K-1} W_k(f) \cdot S(f) \cdot e^{j2\pi ft} df \tag{15}$$

In Equation (13), let c_2=1, we get

$$\int_{-\infty}^{\infty} S(f) \cdot e^{j2\pi ft} df = s(t) \tag{16}$$

Therefore, the Equation (16) is called as the engineering perfect reconstruction condition of the Bark wavelet.

3.2 The realization method of combination mode 3：Bark+ZCPA+HMM

3.2.1 The design of Bark wavelet filter

The sub-section uses Bark wavelet filter to replace FIR filter forming the mode: Bark+ZCPA+HMM. Fig.8 shows the use method of Bark wavelet filter in preprocessing of

ZCPA feature extraction. Where, S(f) is the spectrum of s(t) and W(f) is Bark wavelet function. They meet the relations as S(f)=FFT[s(t)], Y(f)=S(f)W(f) and $\hat{s}(t) = IFFT[Y(f)]$.

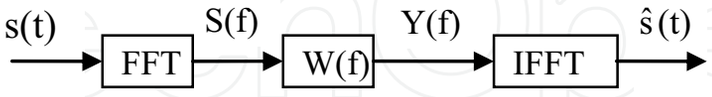


Figure 8. The scheme of the Bark wavelet filters used in preprocessing

Bark scale can be seen as the mapping from linear frequency field to perception frequency field. By using Equation (7) the mapping relation may be realized. This builds the base of critical band filter bank analysis technology. The frequency between 20Hz~16kHz can be divided into 24 fields(or frequency groups).Their pass band bandwidth are equal to the critical band of corresponding frequency, that have become the standard in fact (See Table 3).

Bark No	Low (Hz)	High (Hz)	Bandwidth (Hz)	Bark No.	Low (Hz)	High (Hz)	Bandwidth (Hz)
1	20	100	80	13	1720	2000	280
2	100	200	100	14	2000	2320	320
3	200	300	100	15	2320	2700	380
4	300	400	100	16	2700	3150	450
5	400	510	110	17	3150	3700	550
6	510	630	120	18	3700	4400	700
7	630	770	140	19	4400	5300	900
8	770	920	150	20	5300	6400	1100
9	920	1080	160	21	6400	7700	1300
10	1080	1270	190	22	7700	9500	1800
11	1270	1480	210	23	9500	12000	2500
12	1480	1720	240	24	12000	15500	3500

Table 3. The allocation about 24 critical frequency bands

The critical bandwidth changes with its centre frequency f. The higher f is, the wider the critical bandwidth is. Their relation is as Equation (17).

$$BW_{critical} = 25 + 75 \times [1 + 1.4(\frac{f}{1000})^2]^{0.69} \ (Hz) \tag{17}$$

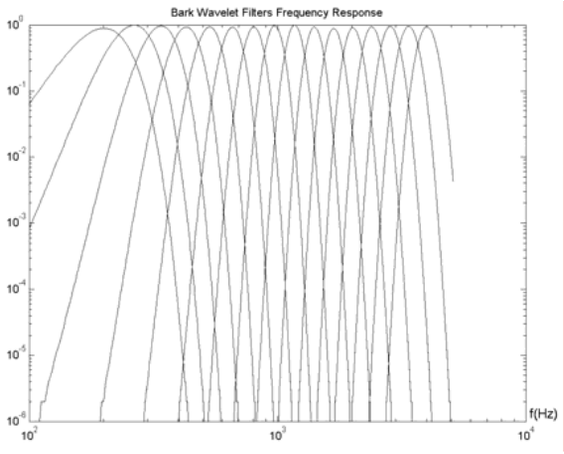


Figure 9. The frequency responses of the 16 Bark wavelet filters

From Table 4, we can see that in Bark field each filter bandwidth follows the equal bandwidth principle, or 3 Bark. During the course of experiment, 1 Bark and 2 Bark bandwidth were seleted, but the results were not good. The parameter k was seleted such as followings. From Equation (1) the centre frequencies of 16 Bark wavelet filters were calculated and from Equation (12) the maxmium responses at centre frequencies can be got, so k values can be determined. From Equation (17) the bandwidthes of all filters can be determined.Having the parameters of Table 4, Bark wavelet filters can be realized.

b_1	b_2	k	Bandwidth (Hz)	b_1	b_2	k	Bandwidth (Hz)
1	4	7	100	9	12	4	180
2	5	5	100	10	13	6	210
3	6	2	100	11	14	7	250
4	7	1	110	12	15	9	300
5	8	0	120	13	16	10	360
6	9	1	130	14	17	10	450
7	10	1	140	15	18	10	550
8	11	4	160	16	19	10	680

Table 4. Confirmation about all the Bark wavelet parameters

3.2.2 The calculation steps of the Bark wavelet filters

Step 1: Read the original speech datum, and estimate the speech data length.
Step 2: Frame the speech data, so that we can obtain the time domain signals of the every frame. Use the 20ms as the frame length, and 10ms as the frame shift.
Step 3: Then go along the FFT transform to obtain the spectrum information of the speech signals, and the width of the window is 256.

Step 4: Initialize the Bark wavelet different parameters.
Step 5: Begin the Bark wavelet transform. Because we used the computer to simulate this transform, there need the discrete Bark wavelet transform. It can be described as:

$$W_k(N-i) = W_k(i-1) = W_k(i \cdot \frac{f_s}{N}),\ i = 1 \cdots, \frac{N}{2} \tag{18}$$

Where N is the last speech data and f_s is sampling frequency.
Step 6: The transform to the input speech signals of the every filter with Bark wavelet should also adopt the discrete Equation (19). As follows:

$$s_k(n) = \sum_{l=0}^{N-1} S(l) W_k(l) e^{\frac{j2\pi n1}{N}} \tag{19}$$

Step 7: Then continue to operate the next frame speech signal as above. Finally we can obtain all the values through Bark wavelet transform.
Through the above steps, we can obtain the results of signal passing Bark wavelet filters, and then we can use the ZCPA theory to extract the features.
Table 7 shows the recognition rates comparison of different words using the ZCPA and BWZCPA (Bark Wavelet ZCPA) features in different SNRs.

3.3 The realization method of combination mode 4：Bark+MFCC+HMM

3.3.1 The principle of the MFCC feature extraction

MFCC feature extraction flow chart is showed as Fig. 10. The working process is:

1. Pre-emphasis of the speech signal, frame, adding window, then make the FFT to obtain the frequency information.
2. Pass the signal through the Mel frequency coordinate triangle filter bank to mimic the human hearing mechanism and the human hearing sensibility to different speech spectrum.
3. Calculate the logarithm value of the signal through the Mel filters to obtain the logarithmic spectrum.
4. Make the discrete cosine transform to the signal and obtain the MFCC feature. In the MFCC algorithm, we use the FFT to calculate the frequency spectrum of the signal in the front-end, while in the back-end we use DCT to further reduce the speech signal's

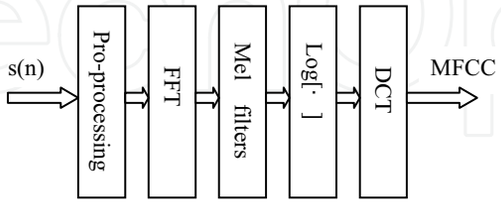


Figure 10. The feature extraction of the MFCC

redundant information, and reach the aim of normalizing speech feature coefficients with small dimensions. The analysis method is based on the assumption that the speech signal is short time stationary signal, so using the fixed window's Fourier transform, we can get the local time-frequency information. As to the any speech segment, the time-frequency resolution is fixed. Based on the uncertainty principle we can find that the time-frequency resolution can't be both high simultaneously. This will make the speech signal details fuzzy, especially to those speech segments with non-stationary, like the explodent sound and spirant, it will definitely lose significant information.

DCT is the FFT with zero valued imaginary part. The feature vectors based on the DCT cover all the frequency band, if only one frequency segment is destroyed by noise interference, then all of MFCC coefficients will be strongly interfered. A speech frame may include two adjacent phonemes, suppose one is sonant and the other is surd, this will definitely blur the two phonemes' information and reduce the speech recognition rate. But if the speech spectrum can be divided into several sub-frequency bands, the situation will be different. Fixed window DFT feature vectors in the time domain and frequency domain all have the same resolutions, and unable to meet requirement of the variant time-frequency resolution which is needed by the feature vectors. Therefore, the FFT and DCT algorithm is unsuitable for the usage in non-stationary speech signal analysis.

3.3.2 The extraction of Bark wavelet MFCC feature

Having analysed the drawbacks of the MFCC, we presented the improving method using Bark wavelet, just as the Fig.11. After the signal is pre-processed and before making FFT, Bark wavelet filtering is inserted. Because the wavelet has the property of multi-resolution, it can divide the signal into some parts with different time-frequency resolution that is suitable to speech signal processing. Here, the front-end Bark wavelet divides signal frequency into 25 sub-bands and we calculate the FFT of each band. Then the spectrum combination, Mel filter processing and the logarithmic energy calculating are performed. Finally the DCT of the former algorithm is replaced by Bark wavelet transform and we obtain the BWMFCC features.

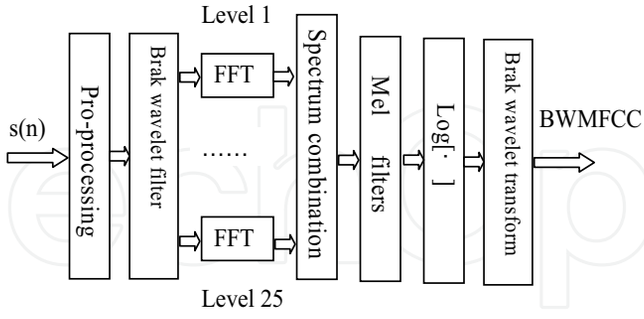


Figure 11. The feature extraction of the Bark wavelet MFCC

The calculation process is as follows:

1. Making pre-emphasis, framing and window adding processes to the original speech data file s(n).
2. Making the Bark wavelet filtering to each frame signal, by using Equation (20).

$$S_k(n) = W_k(n) \cdot S(n),\ 0 \le k \le K-1 \tag{20}$$

Where, $W_k(n)$ is the discrete form of Equation (12), S(n) is the speech signal frequency spectrum, $S_k(n)$ is the k-th sub-band's speech spectrum.

3. The spectrum combination is performed by using Equation (21) to obtain:

$$S'(n) = \sum_{k=0}^{K-1} S_k(n) \tag{21}$$

4. Passing the $S'(n)$ signal through Mel filter bank. The Mel filter bank can smooth the frequency spectrum, and reduce the harmonic, and emphasis the original formant of the speech signal. Therefore, the tone and the pitch of the sound will not appear in the feature coefficients, the speech signal recognition will not be interfered by the different pitch of the input signal.
5. Using the Equation (22) computing the logarithmic energy D(n).

$$D(n) = \log\left(\sum_{k=0}^{N-1} |S'(k)|^2 H_n(k)\right),\ (0 \le n < N) \tag{22}$$

6. Finally, passing the D(n) through Bark filter to get the final speech features BWMFCC(m) by using Equation (23) .

$$BWMFCC(m) = \sum_{n=0}^{N-1} W_m(n) \cdot D(n) \quad (0 \le m \le M-1,\ 0 \le n \le N) \tag{23}$$

Where, N is the number of Mel filters. $H_n(k)$ is Mel frequency filter. In our experiment, N=26, M is the sub-band's number of back-end Bark wavelet, here M=16.

Table 8 shows the recognition rates comparison of different words using the MFCC and BWMFCC features in different SNRs.

4. Results and Analysis

The section will present the experimental results and discussion of the second section and the third section. The exprimental condition and the database used are same. The ZCPA or MFCC feature of every word is normalized into speech feature vector sequence of 64X16 demension. In the experiments, speech data with different SNRs of 50 words 16 persons is used (including data of 15dB, 20dB, 25dB, 30dB and clean). Each person says each word 3 times. The model is trained by speech data (a certain SNR) of 9 persons and the recognition is carried out by speech data (under the same SNR) of other 7 persons, so the recognition result is obtained under this SNR.

4.1 The combination mode 1: FIR+ZCPA+HMM and mode 2: FIR+ZCPA+WNN

Table 5 shows the recognition rates of different words using the combination of FIR+ZCPA+HMM or FIR+ZCPA+WNN under different SNRs. We analysis these experimental results in the tables as follows.

1. HMM is the statistical model based on the time sequence structure of speech signal, and it can simulate reasonably speech time changing process, and describe the whole non-stationary and local stationary of speech well. But a shortcoming of HMM is that

distinguish ability is not strong enough. HMM is influenced by the number of training samples, along with the increase in number of samples, the recognition rate will be improved greatly (see Table 6). Table 6 is the recognition rates using FIR+ZCPA+HMM after adding training samples from 9 to 16, but test samples are unchanged.

2. Wavelet neural network (WNN).
   This network has great advantages in recognition. Not only the structure is simple, algorithm is easy to be implemented and the recognition rate is high, but also linear least square method is used to train weight value, the convergence speed is fast, and training time is only a few minutes. From the table we can see that, along with the increase in the number of words, the number of hidden nodes of network increases, and the recognition rate rises gradually. When reaching a certain number of words, the recognition rate descends but descends little. Although the number of hidden nodes increases, training time is not obviously influenced. Besides, under noisy environment, the recognition rate of the system decreases little, for the speech under the SNR of 15dB, the recognition rate of 50 words is near 90%. It is sufficiently showed that the system combined of ZCPA feature parameter with WNN has good robust performance, so this recognition system can fulfill the task of large number of words, non-special person, real time speech recognition and has wide application foreground.

Noumber of words	SNR / Feature	15dB	20dB	25dB	30dB	Clean
10	HMM	85.7	84.7	86.2	85.7	89.1
	WNN	87.1	90.5	90.5	91.4	92.9
20	HMM	76.6	81.2	82.4	81.7	85.7
	WNN	89.5	92.1	93.3	93.1	94.5
30	HMM	77.1	81.9	83.1	82.9	83.5
	WNN	92.1	93.2	93.3	94.3	94.0
40	HMM	76.6	79.0	81.3	82.6	83.0
	WNN	91.9	93.3	94.3	94.1	94.4
50	HMM	72.1	74.5	80.1	79.0	81.7
	WNN	89.7	91.7	93.3	93.4	94.3

Table 5. The recognition rates using FIR+ZCPA+HMM or FIR+ZCPA+WNN mode (%)

SNR / Number of words	15 dB	20 dB	25 dB	30 dB	Clean
10	88.0	88.7	90.7	91.3	92.0
20	86.0	87.7	90.3	89.3	91.7
30	84.2	87.3	89.1	89.6	90.4
40	82.8	87.7	88.7	90.7	90.8
50	81.7	85.6	87.7	86.7	89.3

Table 6. The recognition rates using FIR+ZCPA+HMM after adding training samples (%)

4.2 The combination mode 3: Bark+ZCPA+HMM and mode 4: Bark+MFCC+HMM

4.2.1 The combination mode 3 : Bark+ZCPA+HMM

As showed in Table 7, ZCPA means the recognition results of using FIR as preprocessing and BWZCPA means the recognition results of using Bark wavelet as preprocessing.
From Table 7 we can see that the recogntion results with Bark wavelet filter are better than the ones with FIR filter in bigger words and higher noise environment. Especially, the system function was improved with increasing of words number. This illustrated Bark wavelet more closer to the hearing perception of human ear than common wavelet.

Noumber of words	SNR / Feature	15dB	20dB	25dB	30dB	Clean
10	ZCPA	85.71	84.76	86.19	85.71	89.05
	BWZCPA	84.00	87.14	90.00	90.48	90.00
20	ZCPA	76.6	81.19	82.38	81.67	85.71
	BWZCPA	77.14	83.57	87.56	84.29	88.20
30	ZCPA	77.14	81.90	83.17	82.86	83.49
	BWZCPA	78.42	83.31	85.71	85.56	87.98
40	ZCPA	76.55	78.26	81.31	82.62	82.98
	BWZCPA	77.50	81.48	84.76	85.00	87.14
50	ZCPA	72.10	74.48	80.09	78.95	81.71
	BWZCPA	73.14	78.20	83.71	85.52	85.24

Table 7. The recognition rates using ZCPA and BWZCPA features in different SNRs (%)

4.2.2 The combination mode 4: Bark+MFCC+HMM

Table 8 shows the recognition rates comparison using MFCC and BWMFCC features in different SNRs. From the table, we can see the recognition rates are significantly increased by using the BWMFCC. The analysis reasons are as follows.

1. By using the wavelet transform in front-end MFCC, we can extract and separate the speech in different transform scales. These in the frequency domain are similar to the frequency segmental processing, and make the subsequent speech analysis to be of natural local property. Thus, the analysis is more delicate.
2. By using wavelet instead of DCT, we overcome the shortage of the DCT. If some frequency segment is destroyed, it only interferes fewer coefficients not all coefficients. We can remove these destroyed coefficients so as to not making strong influence to the speech recognition system.
3. The Bark wavelet transform is designed especially for the speech signal processing. The changing of analysis scale is based on the concept of critical band, and makes wavelet band of each scale be a frequency group. By using the method, we obtain a model which is more close to the human hearing mechanism. The method based on wavelet transform has better robust feature. The result shows that BWMFCC feature can remain high recognition rate under low SNR and large vocabulary conditions. It makes the practical speech recognition system become possible.

Noumber of words	SNR / Feature	15dB	20dB	25dB	30dB	Clean
10	MFCC	86.67	91.90	92.86	93.33	95.24
	BWMFCC	95.72	97.14	96.19	97.14	99.05
20	MFCC	83.80	88.57	90.47	91.47	93.57
	BWMFCC	93.33	94.52	96.19	96.19	96.67
30	MFCC	83.33	87.73	90.32	90.48	93.74
	BWMFCC	94.76	96.19	97.14	97.30	96.35
40	MFCC	82.76	87.57	90.00	91.47	93.57
	BWMFCC	93.69	94.88	95.71	96.07	96.31
50	MFCC	81.19	86.66	89.90	92.28	92.85
	BWMFCC	92.29	93.43	94.19	94.67	94.29

Table 8. The recognition rates using MFCC and BWMFCC features in different SNRs (%)

5. Conclusion

The later conculsions can be obtained from the experimental results of the fourth section.

1. The three parts of speech recognition are conjunct one another and exist the relation restricted among themselves. Bark wavelet was used in improving the feature of ZCPA and MFCC，the latter effect is obviously better than the former.It illustrates that Bark wavelet and the speech character described by MFCC feature are more closer than Bark wavelet and the speech character described by ZCPA feature. The fact is also as such. Bark wavelet is constructed directly according to the hearing perception of human ear, and MFCC is the cepstrum coefficents on the basis of Mel frequency. While Mel frequency is just the hearing frequency of human ear. Though the frequency bins of ZCPA are divided according to the hearing perception, the zero-crossing rate and peak amplitude are time-domain parameters，which are transformed nonlinearly mapping to frequency bin. This kind of nonlinear transform may affect the consistency of ZCPA and hearing frequency, that results in decreasing in function.
2. If the selection of training or recognition network is different, they have different effect on the results. Furthermore, the function of recogntion network has direct relationship with front-end filter and feature extracted. This point can be seen from the experimental results of combination mode1（FIR+ZCPA+HMM）and mode 2（FIR+ZCPA+WNN）. Comparing the two modes, the former two parts are same and the third part is different from using HMM or WNN，the results obtained have much more different.The wavelet neural network has bright foreground for speech recognition.Its training speed is fast, which is good for implementation in real time. Further,it has also good recognition rates under no noise or noise environment and the number of recogintion words is larger.
3. The paper researched some kinds combination modes aiming to the three parts of speech recognition system in Fig. 1. For other combination modes, such as Bark+MFCC+WNN，Bark+ZCPA+WNN and so on, we will research them in later work. Which of combination ever is optimal? This needs considering practical application case. We hope the research can be refered by interesting researcher and get to the purpose of communication mutually and progress.

6. Acknowledgements

The project is sponsored by the Natural Science Foundation of China (No. 60472094), Shanxi Province Scientific Research Foundation for Returned Overseas Chinese Scholars (No. 2006-28), Shanxi Province Scientific Research Foundation for University Young Scholars ([2004] No.13), Shanxi Province Natural Science Foundation.(No. 20051039), Shanxi Province Natural Science Foundation (No. 2006011064) and Ph.D. Start-up Fund of TYUST, China. The authors gratefully acknowledge them.

7. References

Doh-suk, Kim , Soo-Young, Lee & Rhee M., Kil. (1999). Auditory Processing of Speech Signal for Robust Speech Recognition in Real-World Noisy Environments. *IEEE Transactions on speech and audio processing,* Vol.7, No.1, (Jan. 1999) 55-68, 1063-6676

Gowdy J., N; Tufekci Z. (2000). Mel-Scaled Discrete Wavelet Coefficients for Speech Recognition, *Proceedings of IEEE ICASSP'2000*, pp.1351-1354, 0-7803-6293-4, Turkey, Jun. 2000, Publisher, Istanbul

Jingwei, Liu ; Xi, Xiao. (2006). Research and Prospect on Robustness Technology in Real-environment Speech Recognition. *Computer Engineering and Applications*, Vol.42 No.24 , (Aug. 2006) 7-12, 1002-8331

Lawrence, Rabiner. (1999). *Fundamentals of Speech Recognition,* Tsinghua University Press, 7302036403, Beijing, China

Lain M., Johnstone. (1999). Wavelets and the Theory of Non-parametric Function Estimation. *Philosophical Transactions*: *Mathematical, Physical and Engineering Sciences,* Vol.357, No.1760, (Sept. 1999) 2475-2493, Phil.Trans. R. Soc. Lond. A, 1364503X

Musavi, M.; et al. (1992). On the Training of Radial Basis Function Classifiers. *Neural Networks e,* Vol.5, No.4, (July 1992) 595-603, 0893-6080

Oded, Ghitza. (1992). Auditory Nerve Representation as a Basis for Speech Processing, In: *Advances in speech signal processing, S. Furui and M. M. Sondhi*, 453-485, Marcel Dekker, New York

Oded, Ghitza. (1994). Auditory Models and Human Performance in Tasks Related to Speech Coding and Speech Recognition. *IEEE Transactions on Speech and Audio Processing,* Vol.2, No.1, (Jan. 1994) 113-131, 1063-6676

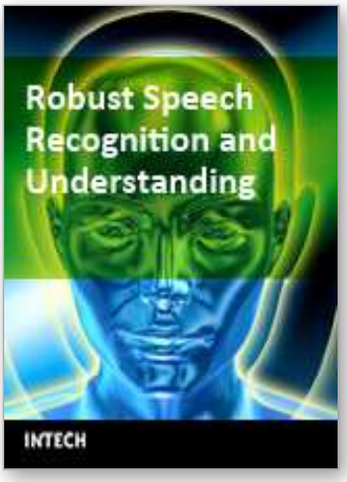
Qiang, Fu; Kechu, Yi. (2000). Bark Wavelet Transform of Speech and its Application in Speech Recognition. *Journal of Electronics,* Vol.28, No.10, (Oct. 2000) 102-105, 0372-2112

Shuyan,Zhao; et al. (2005). A Speech Recognition System of Isolated Words Based on ZCPA and DHMM, Vol.36, No.3, (May 2005) 246-249, 1007-9432

Tianbing,Yao; Tianren, Yao & Tao, Han. (2001). Development and Prospect of Robust Speech Recognition. *Signal Processing,* Vol.17, No.6, (Dec. 2001) 484-497, 1003-0530

Zhiping, Jiao; et al. (2005).A Noise-Robust Feature Extration Method Based on Auditory Model in Speech Recognition, Vol.36, No.1, (Jan. 2005) 13-15, 1007-9432

Zhigang, Liu; Xiaoru, Wang & Qingquan, Qian. (2003). A Review of Wavelet Networks and Their Applications. *Automation of Electric Power Systems,* Vol.27, No.6, (Mar. 2003) 73-79, 1000-1026



Robust Speech Recognition and Understanding

Edited by Michael Grimm and Kristian Kroschel

ISBN 978-3-902613-08-0
Hard cover, 460 pages
Publisher I-Tech Education and Publishing
Published online 01, June, 2007
Published in print edition June, 2007

This book on Robust Speech Recognition and Understanding brings together many different aspects of the current research on automatic speech recognition and language understanding. The first four chapters address the task of voice activity detection which is considered an important issue for all speech recognition systems. The next chapters give several extensions to state-of-the-art HMM methods. Furthermore, a number of chapters particularly address the task of robust ASR under noisy conditions. Two chapters on the automatic recognition of a speaker's emotional state highlight the importance of natural speech understanding and interpretation in voice-driven systems. The last chapters of the book address the application of conversational systems on robots, as well as the autonomous acquisition of vocalization skills.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen