environment noise and the variation of speaker's voice with a higher recognition rate of voice signal to some extent.

II. Principle of System Design

Speaker identification is the problem of pattern classification which recognizes a correct result after classifying the features of different speakers' speech. Fig. 1 shows the flow chart of a complete speaker identification system, it consists of the following steps:

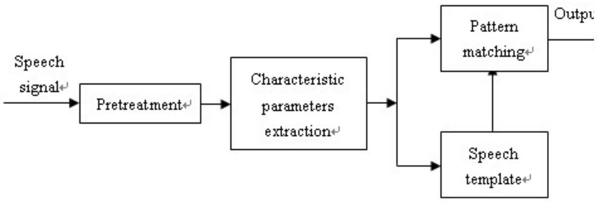


Fig. 1. Speaker identification system flowchart

(1) Pretreatment/Feature extraction: This step generally consists of three sub-processes. First, some form of speech activity detection is performed to remove non-speech portions from the signal. Next, features conveying speaker information are extracted from the speech. From the source-filter theory of speech production, it is known that the speech spectrum shape encodes information about the speaker's vocal tract shape via resonances. So some form of spectral based features is used in most speaker identification systems. Short-term analysis, typically with 20 ms frames generated every 10 ms, is used to compute a sequence of magnitude spectra using either LPC or FFT analysis. Most commonly the magnitude spectra are then converted to cepstral features after passing through a mel-frequency filterbank and time-differential (delta) cepstra are appended. The final process in pretreatment/feature extraction is some form of channel compensation. It is well known that different input devices will impose different spectral characteristics on the speech signal, such as bandlimiting and shaping. Channel compensation aims at removing these channel effects. Most commonly some form of linear channel compensation, such as long- and short-term cepstral mean subtraction, are applied to features.

(2) Speech template: The speech from each known, verified speaker, for all speakers that need to identified, is acquired to build (train) the speech template for that speaker. Usually this is carried out off-line as part of the system configuration and before the system is deployed. Speech from a speaker is passed through the pretreatment/feature extraction steps described above and the feature vectors are used to create a speaker speech template. Desirable attributes of a speaker speech template are: (a) a theoretical underpinning so one can understand model behavior and mathematically approach extensions and improvements; (b) generalizable to new data so that model does not over fit the enrollment data and can match new data; (c) parsimonious representation in both size and computation.

(3) Speaker classification: Speaker classification operation of the system is carried out where the speech from an unknown utterance is compared against each of the trained speech template in order to achieve speaker identification.It acts as a normalization to help minimize non-speaker related variability (e.g., text, microphone, noise) in the likelihood ratio score. There are two dominant approaches used for representing speaker classification in the likelihood ratio test. The first approach, known as likelihood sets, cohorts or background sets, use a collection of other speaker models to compute the imposter match score. The imposter match score is usually computed as a function, such as the max or average, of the match scores from a set of non-claimant speaker models. The non-claimant speaker models can come from other enrolled speakers or as fixed models from a different corpus. The second approach, known as general, world or universal background modeling, uses a single speaker-independent model trained on speech from a large number of speakers to represent speaker-independent speech.

There are many techniques, such as, dynamic time-warping (DTW), hidden Markov models (HMMs), neural networks (NNs), and vector quantization (VQ), have some or all of these attributes and have been used in speaker verification/identification systems.

In the training mode of the DTW approach [6,7], the speaker templates, which are the sequences of feature vectors obtained from the text-dependent speech waveforms, are created. In the testing mode, matching scores are produced by using DTW to align and measure the similarities between the test waveform and the speaker templates.

In the HMMs approach [8-10], the sequences of feature vectors, which are extracted from the speech waveforms, are assumed to be a Markov process and can be modeled with an HMM. During the training mode, HMMs' parameters are estimated from the speech waveforms. In the testing mode, the likelihood of the test feature sequence is computed based on the speaker's HMMs.

In the neural networks-based method [11-12], each speaker has a personalized neural network that is trained to be activated only by that speaker's utterances. The testing waveforms are tested by the speakers' personalized neural networks to make Speaker Identification decisions.

In the VQ Speaker Identification approach [13-14], in the training mode, a codebook for each speaker is obtained as a reference template for the speaker. In the testing mode, Speaker Identification is usually performed by finding the codebook and its corresponding speaker that gives the smallest average VQ distortion to represent the unknown speaker's waveform. The average VQ distortion here shows the similarity between the unknown speaker's speech and the reference template. The smaller average VQ distortion, the better match between testing speech and reference template is. The lack of time warping in the VQ approach greatly simplifies the system. However, some speaker-dependent temporal information, which is present in the waveforms, is neglected in VQ Speaker Identification.

III.FEATURE EXTRACTION

The most fundamental process common to all forms of speaker and speech recognition systems is that of extracting vectors of features uniformly spaced across time from the time-domain sampled acoustic waveform. Mel-frequency Cepstral Coefficients (MFCC) [15-16], based on short-time spectral analysis, are commonly used feature vectors for speaker identification. Fig. 2 illustrates the computation of MFCC feature extraction flowchart:

PRE-ENHANCEMENT: A high-pass filter is applied to the waveform. This emphasizes the higher frequencies and compensates for the human speech production process which tends to attenuate high frequencies. Let the input speech signal $s(n)$ pass through 1st order high-pass filter

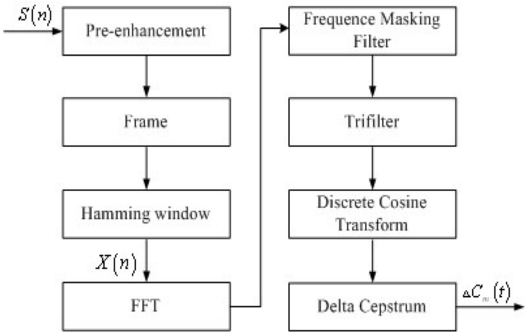


Fig. 2. MFCC flow chart

$$H(z) = 1 - a \times z^{-1} \quad 0.1 \le a \le 0.9 \tag{1}$$

The filtered signal function is:

.
$$s_1(n) = s(n) - a \times s(n-1) \tag{2}$$

FRAMING: Combine N sample spots to a conservative unit called frame(we set N 256).To prevent the drastic change of neighboring two frames for a smooth short-time voice features and sequence spectral, we make some overlaps which includes M sample spots(here we set M=N/2)between the neighboring two frames.

WINDOWING: Each frame is multiplied by a window function such as hamming window. The window function is needed to smooth the effect of using a finite-sized segment for the subsequent feature extraction by tapering each frame at the beginning and end edges. Supposing the signal after frame process is $s(n), n = 0,1,...,N-1$, after multiplied with hamming window is $x(n) = s(n) \times w(n)$, the form of $w(n)$ as follow:

$$W(n,\alpha) = (1-\alpha) - \alpha\cos\left(2\pi n / (N-1)\right), 0 \le n \le N-1 \tag{3}$$

Different α will produce different hamming window. The hamming window offers the familiar bell-shaped weighting function but does not bring the signal to zero at the edges of the window. It minimizes the spectral distortion.

FOURIER TRANSFORM AND HEARING MASKING: A Fast Fourier Transform (FFT) operation is applied to each frame to yield complex spectral values. Here, the phase information is ignored and only the FFT magnitude spectrum is considered. Human can exactly identify the speech which has low signal noise ratio even with jamming voice. This mainly depends on the input function of both ears. To decrease the influence of noise signal to speech signal, a hearing masking effect based masker is applied in the frequency domain to speech signals [16].

TRIFLER: Multiply spectrum energy by a group of dozen triflers(1-7 order using low-frequency MFCC,8-13 order using MidMFCC,14-20 order using IMFCC[17]) in order to find out the logarithmic energy of each filter's output. The frequency of triangle's two down points in the each filter equals to the center frequency of the neighboring two filters which means the transition belt of every two neighboring filters overlap. The correspondence of Hz-Mel frequency as follows:

$$f_{MFCC} = 2595 \times \log_{10}(1 + f / 700)$$

$$f_{IMFCC} = 2146.1 - 1127 \times \ln\left(1 + \frac{4000 - f}{700}\right) \tag{4}$$

$$f_{MIDMFCC} = \begin{cases} 1073.05 - 527 \times \ln\left(1 + \frac{2000 - f}{300}\right), 0 < f \le 2000 \\ 1073.05 + 527 \times \ln\left(1 + \frac{f - 2000}{300}\right), 2000 < f \le 4000 \end{cases}$$

DISCRETE COSIN TRANSFORM: Put the above-mentioned 20 logarithmic energy into the discrete cosine transform formula in order to get the L-order Mel-Scale Cesptral parameter. Here, L is usually set 12, N adopts 20. Discrete cosine transforms formula as follow:

$$y(k) = \alpha(k)\sum_{n=0}^{N-1} x(n)\cos\left(\frac{\pi(2n+1)k}{2N}\right), k = 0,1,...,L-1 \tag{5}$$

Where $\alpha(k) = \begin{cases} \sqrt{\frac{1}{N}} & k = 0 \\ \sqrt{\frac{2}{N}} & k \ne 0 \end{cases}$.

DELTA CEPSTRUM: Although we've got 12 feature parameters, we usually alternatively add Delta cepstrum parameter in practical application of voice recognition to show the variations of Delta cepstrum parameter with respect to the time which means the differentiation of Delta cepstrum parameter with respect to the time. In other words, it represents the dynamic change of Delta cepstrum in time. Formula as follows:

$$\Delta C_m(t) = \left\lfloor \sum_{j=-M}^{M} C_m(t+\tau) \right\rfloor / \left\lfloor \sum_{j=-M}^{M} \tau^2 \right\rfloor \tag{6}$$

So far, the MFCC feature extraction is finished. From the point of overall frame, actually we first process the voice signal by STFT (Short-time Fourier Transform), then process the energy spectrum by Bandpass filter using group filters, finally make Delta cepstrum calculation. In the practical application, we can show the feature vector in different dimension according to the test as required.

IV. Speech Pattern Recognition

For speaker identification system, after feature extraction we need to build speech template for the speaker recognize which one is the speaker through the feature classification. The so-called speech template is a recognition model for presenting the speaker's speech characteristics' distribution in feature space. Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), Dynamic Time warping (DTW) and Vector Quantization (VQ) are prevalent techniques for patterns matching in speaker identification systems[18,19]. The VQ technique includes two steps: training and testing. In the training phase, required number of highly representative code vectors for each speaker is achieved by using VQ. Vector quantization is implemented through Binary-Splitting Linde-Buzo-Gray algorithm. The collection of these code vectors is called a codebook. A codebook (acoustic model) for each speaker is constructed in the same way. In the testing phase, an unknown voice after extracting voice feature vectors will be compared with the codebook of each speaker in the speaker database and distortion will be computed in order to identify who is the speaker.

A. Lbg Algorithm

LBG algorithm is a codebook training recursive algorithm with the principle of nearest algorithm and minimize average distortion algorithm, which generates the codebook whose performance has much to do with the original codebook [20]. The codebook is clustered from the MFCC feature vector of speaker's training series. LBG algorithm as follows:

(1) Give all reference vectors X required in tainting VQ codebook. Let S represent the class of X; Set quantitative series, distortion control threshold (here we adopts 0.01), maximum iteration times L and original codebook $\left\{Y_1^{(0)},\ \ Y_2^{(0)},...,Y_N^{(0)}\right\}$; Set overall distortion $D^{(0)}=\infty$; Initialize the iteration time m=1

(2) In terms of nearest principle we divide S into N subsets

$$\left\{S_1^{(m)},S_2^{(m)},...,S_n^{(m)}\right\} \quad d\left(X,Y_l^{(m-1)}\right)\le d\left(X,Y_i^{(m-1)}\right),\forall i,i=l \quad \left(x\in S_1^{(m)}\right) \qquad (7)$$

(3) Calculate distortion:

$$. \qquad D^{(m)}=\sum_{i=1}^{N}\sum_{X\in S_l^{(m)}} d(X,Y_l^{(m-1)}) \qquad (8)$$

(4) Calculate new code:

$$Y_1^{(m)},Y_2^{(m)},...,Y_N^{(m)}. \qquad Y_i^{(m)}=\frac{1}{N_i}\sum_{X\in S_i^{(m)}} X \qquad (9)$$

(5) Calculate relative distortion:

$$. \qquad \delta^{(m)}=\frac{\left|D^{(m-1)}-D^{(m)}\right|}{D^{(m)}} \qquad (10)$$

Compare $\delta^{(m)}$ with distortion threshold δ. If $\delta^{(m)}\le\delta$, it turns to step (4), else turns to step (3). Fig 3. Shows some sample of speech VQ feature vectors calculated by LBG algorithm.

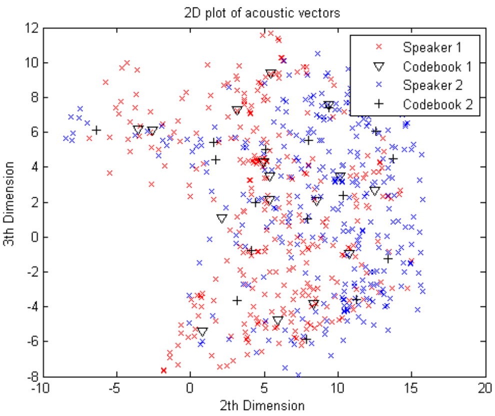


Fig. 3. Speech VQ vectors

B. Vq Voice Recognition

In this phase, an unknown voice after extracting voice feature vectors will be compared with the codebook of each speaker in the speaker database and distortion will be computed. Unknown voice will have minimum distortion with the true speaker. By the following method, system will provide the identity of the speaker.

Given $X=\left\{x_1,x_2,...,x_T\right\}$ is an uncertain speaker's feature vector of T frames, $\left\{B^1,B^2,...,B^N\right\}$ is the codebook in training phase (N is the number of speakers).Specific step of recognition as follows:

(1)Find out $\min\limits_{m\in M} d\left(x_j,B_m^i\right)$, thereby, x_j is the feature

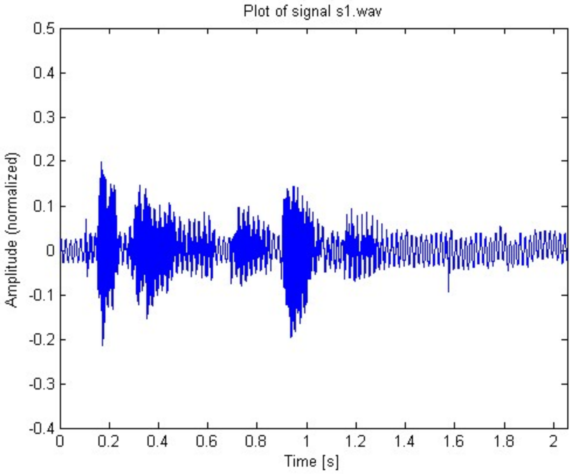


Fig. 4. Voice of "Zhejiang Sci-tech University" time domain waveform figure

vector in j th frame, B_m^i represents the m th code of the i th speaker, with the overall M codes, here d is the Euclidean distance measure.

(2)Calculate average quantitative distortion

$$D_i = \frac{1}{T}\sum_j \min_{1\le m\le n}\left[d\left(x_j, B_m^i\right)\right]. \qquad (11)$$

(3)The index i of minimum distortion in $\{D_1, D_2,..., D_N\}$ corresponds to the speaker.

V. Experimental Results

The voice samples are recorded in a room (window and door open) contains about 12 people, one people recorded a random sentence for several seconds, and the others doing work on themselves. The voice signals are acquired by a PC sound card at sampling frequency 11025 Hz and resolution 16 bit. There are some uncertain noisy generated by the activities of the other people in the room and indoor and outdoor environment acoustic noisy. Fig. 4 shows the speech signal collected by the system when one speaker utters the name of Zhejiang Sci-Tech University in Chinese.

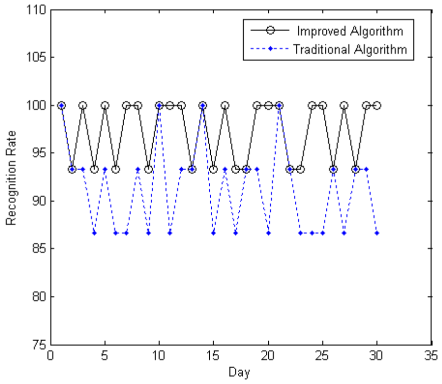


Fig. 5. Comparison of two speaker identification algorithm recognition rate variation over time.

TABLE 1.
INDIVIDUAL RECOGNITION ACCURACY RATES OF TRADITIONAL METHOD AND OUR PROPOSED METHOD.

Speaker ID	Recognition Accuracy Rate (%)		Speaker ID	Recognition Accuracy Rate (%)	
	Traditional	Proposed		Traditional	Proposed
1	96.7	100	2	86.7	100
3	96.7	96.7	4	83.3	93.3
5	93.3	100	6	90	100
7	100	96.7	8	90	100
9	83.3	96.7	10	93.3	100
11	96.7	96.7	12	90	93.3
13	93.3	93.3	14	90	96.7
15	90	96.7			

In a month, every speaker reordered one different sentence everyday. The environment during sound recording is detected with some noises (No 4 sample environment with stronger noise). Thus, we collect a voice database includes more than 450 voice samples of 15 (13 males and 2 females) different speakers. The digitized speech signals are blocked into consecutive overlapping frames, during of each frames is 23 ms and a new frame contains the last 11.5ms of the previous frame's data. In other words, a frame has 256 samples and every new frame has 128 samples of previous frame. The performance of the proposed speaker identification is evaluated by performing two experiments on this voice database. Following are the voice samples for the both experiments.

In the first experiments, the voice samples collected on first day are used as training data, and the other data are used to testing. The performance of the proposed method is compared with traditional speaker identification system using MFCC Features with VQ Technique. Here, we set $\alpha = 0.46$ to produce hamming window

TABLE 2.
RECOGNITION ACCURACY RATES USING SAMPLES COLLECTED ON DIFFERENT DAY AS TRAINING DATA

Training Data	Recognition Accuracy Rate (%)	Training Data	Recognition Accuracy Rate (%)
1st day	97.8	2nd day	97.3
3rd day	97.6	4th day	97.1
5th day	98.0	10th day	96.7
15th day	97.3	20th day	97.8
25th day	98.0	30th day	96.7

In the second experiments, we selected the voice samples collected on different one day as training data, and the other voice samples used as testing data. It is clearly that the noisy contained in the voice samples on different day and the physical states of the speakers are not same. In other word, there time-varying noisy in the collected voice sample. Results of Experiment 1 are shown in Fig. 5 and Table 1. The comparison results show that the accuracy and robustness of speaker identification are improved by introducing a hearing masking effect based masker and a group of dozen triflers during the MFCC features extraction process. Results of Experiment 2 as illustrated in Table 2, which show that the recognition accuracy rates are unrelated to the training data selects.

VI. Conclusion

In this paper, we proposed a new MFCC based speaker identification system with VQ modeling technique. Its performances are investigated under an unconstrained situation, includes variable noise condition and text-independent speech. Results show that the proposed speaker identification system has very good identification accuracy and therefore, it is robust against time varying noise.

In addition to the low-level spectrum features used by current systems, there are many other sources of speaker information in the speech signal that can be used. High-level features not only offer the potential to improve accuracy, they may also help improve robustness since they should be less susceptible to noise. In future, we will focus on exploitation higher-level of information to the identification accuracy and effort to overcome the more difficult issues in unconstrained situations.

ACKNOWLEDGMENTS

This project is supported by the National Natural Science Foundation of China under grants (60873020, 61070063, 50875245), and supported by Zhejiang Provincial Natural Science Foundation of China under grants (Z1080702, Y107558).

REFERENCES

[1] B. S. Atal, “Speaker Recognition: Tutorial,” *Proc. IEEE*, vol. 64, 1976, pp. 460-475.

[2] J. R. Deller, J. L. Hansen, and J. G. Proakis, “Discrete-Time Processing of Speech Signals,” *IEEE Press*, NY , 2000.

[3] M. Grimaldi, F. Cummins, “Speaker identification using instantaneous frequencies,” *IEEE Transactions on Audio, Speech and Language Processing*, vol 16 (6), 2008, pp. 1097-1111.

[4] X. Lu, J. Dang, “An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification,” *Speech Communication*, Vol. 50(4), 2008, pp. 312-322.

[5] C. Sandipan, R. Anindya, and M. Sourav, “Capturing complementary information via reversed filter bank and parallel implementation with MFCC for improved text-independent speaker identification,” *IEEE International Conference on Computing: Theory and Application*, India, 2007, pp. 463-466.

[6] J. P. Campbell, “Speaker recognition: a tutorial,” *Proc. IEEE*, 1997, 85, pp. 1437-1462.

[7] H. Sakoe, S. Chiba, “Dynamic programming algorithm optimization for spoken word recognition,” *IEEE Trans. Acoust. Speech Signal Process*, 1978, 26, pp. 43-49.

[8] T. Matsui, S. Furui, “Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMM’s’,” *IEEE Trans. Speech Audio Process*, 1994, 2, (3), pp. 456-459.

[9] N. Z. Tishby, “On the application of mixture AR hidden Markov models to text independent speaker recognition,” *IEEE Trans. Acoust. Speech Signal Process.*, 1991, pp. 563-570.

[10] K. Yu, J. Mason, J. Oglesby, “Speaker recognition using hidden Markov models, dynamic time warping and vector quantization,” *IEEE Proc. Vis. Image Signal Process.*, 1995, 142, (5), pp. 313-318.

[11] K. R. Farrell, R. J. Mammone, K. T. Assaleh, “Speaker recognition using neural networks and conventional classifiers,” *IEEE Trans. Speech Audio Process.*, 1994, 2, (1), pp. 194-205.

[12] J. Oglesby, J. S. Mason, “Optimisation of neural models for speaker identification,” *ICASSP-90*, 1990, pp. 261-264.

[13] W. B. Mikhael, P. Premakanthan, “Speaker identification employing redundant vector quantisers,” *Electron Lett.* 2002, 38, pp. 1396-1398.

[14] G. Zhou, W. B. Mikhael, B. Myers, “A novel discriminative vector quantisation approach for speaker identification,” *J. Circuits. Syst. Comput.*, 2005, 14, (3), pp. 581-596.

[15] S. Hayakawa, F. Itakura, “Text-dependent speaker recognition using the information in the higher frequency band,” *IEEE International Conference on Acoustic, Speech and signal Processing*, Adelaide, Australia, 1994, pp. 137-149.

[16] X. Luo, I. Y. Soon, C. K. Yeo, “An auditory model for robust speech recognition,” *IEEE International Conference on Audio, Language and Image Processing*, Shanghai, China, 2008, pp. 1105-1109.

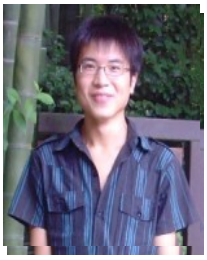
[17] Z. Qian, L. Y. Liu, X. Y. Li, “Speaker identification based on MFCC and IMFCC,” *International Conference on Information Science and Engineering*, 2009, pp. 5416-5419.

[18] R. Togneri, D. Pullella, “An overview of speaker identification: Accuracy and robustness issues,” *IEEE Circuits and Systems Magazine*, Vol. 11(2), 2011, pp. 23-61.

[19] M. Faundez-Zanuy, E. Monte-Moreno, “State-of-the art in speaker recognition,” *IEEE Aerospace and Electronic Systems Magazine*, Vol. 20(5), 2005, pp. 7-12.

[20] A. Zulfiqar, A. Muhammad, A. M. Martinez Enriquez, “A speaker identification system using MFCC features with VQ technique,” *3rd International Symposium on Intelligent Information Technology Application*, 2009, pp. 115-118.

Yaming Wang obtained his Ph. D. degree in biomedical engineering from Zhejiang University, China. He is currently a professor of computer science at Zhejiang Sci-Tech University, Zhejiang, China. He had been Visiting Researcher and Visiting Scientist to Hong Kong University of Science & Technology, HKUST. His current title is Dean of College of Information and Electronics. His research interests include computer vision, pattern recognition and signal processing, computer vision, medical image processing.



Fuqian Tang obtained his Bachelor degree in electronic and information engineering from Linyin Normal University, Shandong, China, in July, 2009. He is currently pursuing the Master degree in signal and information processing at Zhejiang Sci-Tech University, Zhejiang, China. His research interests include pattern recognition, artificial intelligent and signal processing.

Junbao Zheng obtained his Ph. D. degree in biomedical engineering from Zhejiang University, China. His first degree is Bachelor of biomedical engineering and awarded in Zhejiang University. He is currently an associate professor of computer science at Zhejiang Sci-Tech University, Zhejiang, China, since July 2008. His research interests include pattern recognition and signal processing, computer vision, sensor, artificial intelligent, measurement science and technology.