Speaker Verification System with Hybrid Model Improved by Adapted Continuous Wavelet Transform

Hyoungsoo Kim*, Sung-il Yang*, Younghun Kwon** and Kyungjoon Cha***

* This work was supported by the fund of Hanyang University

Abstract

In this paper, we develop a hybrid speaker recognition system [1] enhanced by pre-recognizer and post-recognizer. The pre-recognizer consists of general speech recognition systems and the post-recognizer is a pitch detection system using adapted continuous wavelet transform (ACWT) to improve the performance of the hybrid speaker recognition system. Two schemes to design ACWT is considered. One is the scheme to search basis library covering the whole band of speech fundamental frequency (speech pitch). The other is the scheme to determine which one is the best basis. Information cost functional is used for the criterion for the latter. ACWT is robust enough to classify the pitch of speech very well, even though the speech signal is badly damaged by environmental noises.

Ⅰ. Introduction

For speaker recognition, feature vectors that exhibit high speaker discrimination power and high interspeaker variability are desired. In most cases, LPC, Cepstrum, LSP, and Filterbank energies are used as feature vectors [2, 3]. Many forms of pattern matching and models are possible for recognizers. Pattern-matching methods include dynamic time warping (DTW), hidden Markov model (HMM), multilayer perceptron (MLP), and vector quantization (VQ). Template models are used in DTW, statistical models in HMM, and time delayed neural network(TDNN). These features and models have respectively individual characteristics. For each model, features from the same speaker do not yield the same results. We compensate errors from poor features and poor models by hybrid method of several features and several recognizers. In this system, hybrid method are engaged for pre-recognizer. If result-scores of pre-recognizer are low, another recognizer, as post-recognizer, may be used. We use each speaker's pitch as an individual identity and post-recognizer compensates for low scores.

Voiced speech is produced through the excitation of vocal tract with quasi-periodic vibrations (pitch) of the vocal folds at the glottis. The pitch period is one of the important factors for verifying speakers. We obtain the pitch period using Adapted Continuous Wavelet Transformation (ACWT). We also survey ACWT and display basis library designed for detecting speech pitch. It is imperative to understand the errors made by speaker verification system (SVS). There are two types of errors: false rejection (FR) of a valid user and false acceptance (FA) of an invalid user. FA errors are fatal for high-security speaker-verification applications. Two types of FA errors are existed: Invalid user with invalid password (Denial-1), and invalid user with valid password (Denial-2).

In some instances, the formants of the vocal tract can alter significantly the structure of the glottal waveform so that the actual pitch period may be difficult to detect [4]. Such interactions generally are most deleterious to pitch detection during rapid movements of the articulators when the formants are also changing rapidly.

In practical application, the background ambient noise may also affect the performance of the pitch detector seriously. This is especially serious in mobile communication environments where a high level of noise

* Dept. of Control & Instrumentation Engineering, Hanyang University.
** Dept. of Physics, Hanyang University.
*** Dept. of Mathematics, Hanyang University.
Manuscript Received : July 6, 1999

is present.

Our new PDA may improve the performance of SVS. The method is to detect the nearly accurate pitch period considering the local statistics of speech. Information cost functional, which is used in ACWT, adapts the basis of ACWT to the local acoustic characteristic of speech.

Ⅱ. Pitch Detection Method with Adapted Continuous Wavelet Transform (ACWT)

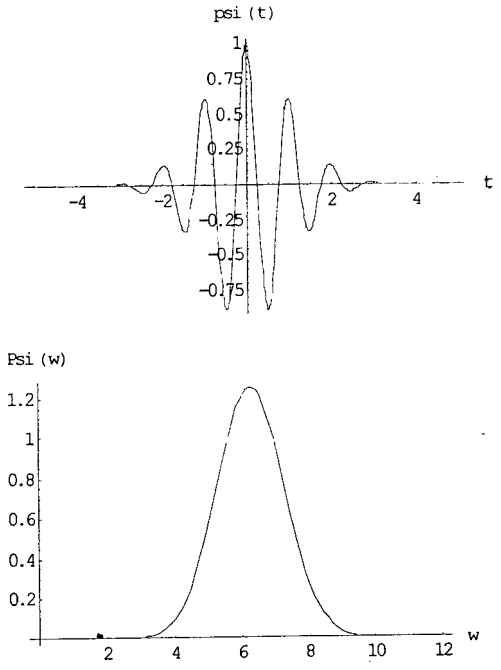


Figure 1. Mother wavelet function in Time and Frequency domain.

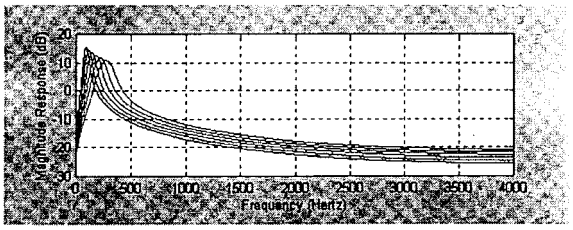


Figure 2. Some of bases ACWT's basis library.

The Gauss wavelet tracks well pitch in lower resolution scales. From these concepts, we analyze pitches from speech by ACWT based on Modulated Gaussian Wavelet (Morlet). The basic scheme of post-recognizer is to extract adaptively the fundamental frequency (pitch) of speech with CWT. The basis function is Morlet wavelet basis [5, 6], which is a real version

$$\psi(t) = e^{-t^2/2\sigma^2}\cos(2\pi f_0 t) \tag{1}$$

This mother wavelet's center frequency and bandwidth are respectively approximately 6.2831 Hz and 1.6623 Hz. The parameters of Eq. (1), such as σ and f_0, can be changed by application's characteristics. Speech pitch period is generally located from 3 ms to 20 ms, corresponding to 300 to 50 Hz in the fundamental frequency.

We make basis library that covers the entire pitch band and use an information cost functional [7] for selecting the best basis to determine the input speech pitch, and try to arrange efficiently the basis library with least number of bases to consider computational load. Fig. 2 displays some of wavelet bases for the basis library.

The basis library is designed according to the following process. At first, we consider the worst case that input speaker's pitch is located between two discretized basis components (ψ_j and ψ_{j+1}) in the basis library, to validate our designed basis library. In this case, distance between a center frequency of two basis components and a fundamental frequency is to be longest. The center frequency of discretized basis component is F_s/α where α is a scale parameter and to be a real number.

$$c_\tau = \sum_{n=-(N-1)/2}^{(N-1)/2} x(n)\psi_\alpha(\tau - n)$$

where α = scale parameter $\in R$, $0 < \tau < L-1$.

The criterion to select the scale parameters can be defined as follows.

$$NE_\alpha = -\sum_{\tau=-(L-1)/2}^{(L-1)/2} \left(\frac{c_\tau}{\sum_{i=-(L-1)/2}^{(L-1)/2} c_i} \right)^2 \log\left(\left(\frac{c_\tau}{\sum_{i=-(L-1)/2}^{(L-1)/2} c_i} \right)^2 \right)$$

$\Leftrightarrow NE_\alpha <$ Selection Threshold

where ψ_α is a sampled Morlet wavelet basis. (2)

We determine scale parameters for the basis library by calculating the longest frequency distance which is satisfying Selection Threshold constraint, changing the center frequency of ψ_α. The our basis library that consists of four basis components and covers whole speech pitch

band is constructed by Eq. (2).

ACWT does not require preemphasis and any other enhancing method in pitch detection. It is so robust to noises that we can catch exact pitches under very noisy environment. However, wide analysis interval may yield some errors. If the pitch periods of some input speeches fluctuate severely, their bands may be beyond the frequency bands of bases of CWT. Composed speeches make few errors in rather large analysis interval because of their stable pitch periods. Even though input speech is very time-varying, the errors can be compensated for by using narrow analysis interval or culling the proper bases from the other calibrated bases of ACWT. Four scale outputs of ACWT are extracted from input speech by Eq. (3).

We use the non-normalized Shannon entropy, as an information cost functional [7].

$$E_j(x) = -\sum_i c_i^2 \log(c_i^2), \quad j = 1,2,3,4$$

with the convention $0\log(0) = 0$ (4)

The optimal scale output is decided by comparing each information cost functional value of $E_1(x)E_2(x)E_3(x)$ and $E_4(x)$ calculated from four scale outputs of CWT. If $E_3(x)$ is the smallest number, a set of output sequence $\{c_3\}$ is the output of optimal basis and is post-processed by peak detector. Peak detector consists of a center clipper and a peak tracer. A set of output sequence $\{c_i\}$ is normalized before using the peak detector because output level is fluctuated with the energy of input speech. This procedure

Where L is the length of filter taps and N is the length of input sequence.

$$\begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_N \end{bmatrix} = \frac{1}{\sqrt{a}} \begin{bmatrix} \psi_0 & \psi_1 & \psi_2 \cdots\cdots \psi_{\frac{(L-1)}{2}} & 0 & 0 & 0 & 0 \cdots\cdots \\ \psi_{-1} & \psi_0 & \psi_1 \cdots\cdots \psi_{\frac{(L-1)}{2}-1} & \psi_{\frac{(L-1)}{2}} & 0 & 0 & 0 \cdots\cdots \\ \vdots & \vdots & \vdots & \ddots & & & \\ 0 & 0 & 0 \cdots\cdots \psi_{-\frac{(L-1)}{2}} & \psi_{-\frac{(L-1)}{2}-1} & \cdots\cdots\cdots\cdots & \psi_{-2} & \psi_{-1} \quad \psi_0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} \quad (3)$$

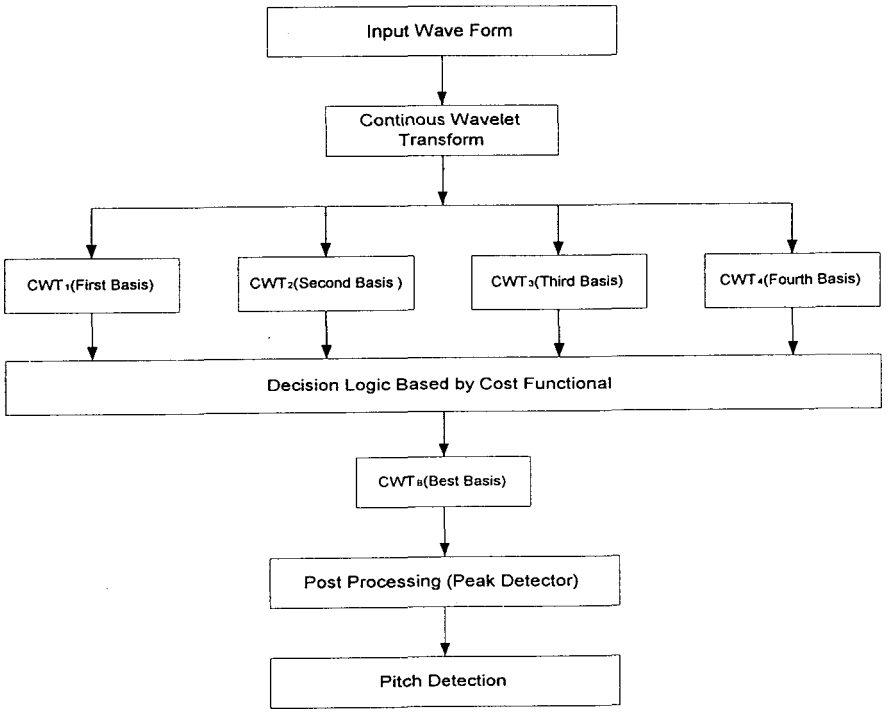


Figure 3. Block diagram of adapted continuous wavelet transform for pitch detection.

is depicted in Figure 3.

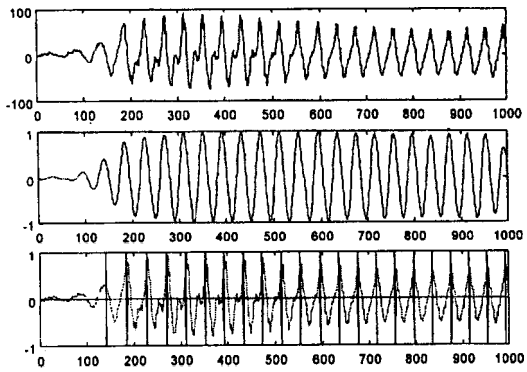


Figure 4. Wavelet filtered output of speech A (/seyo/). Pitch mean = 40.35. Pitch variance = 1.33.

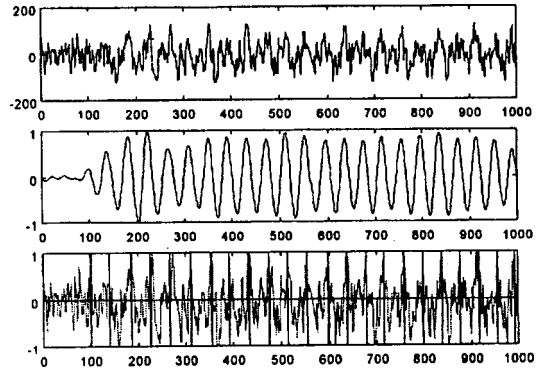


Figure 5. Wavelet filtered output of speech A with environmental noises. SNR=0.76. Pitch mean = 40.29. Pitch variance = 2.78.

Figs. 4 and 5 are results of same speech sources, but Fig. 5 is added with environmental noises, such as car noises and click sounds. The results show some consistency of ACWT under noisy environment.

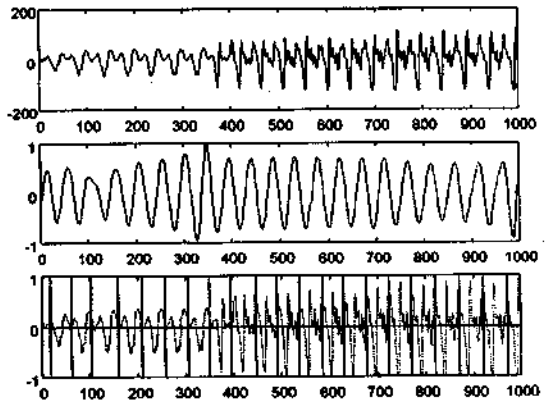


Figure 6. Wavelet filtered output of /man/. Pitch mean = 46.40.

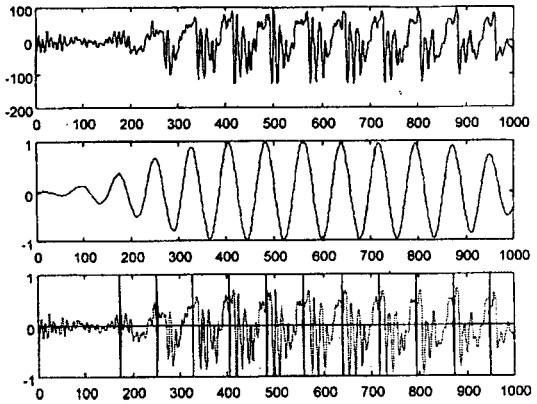


Figure 7. Wavelet filtered output of speech /ban/. Pitch mean = 77.33.

Figs. 6 and 7 show the results of ACWT in speeches with voiced consonants.

Ⅲ. System Structure

3.1 Pre-Recognizer

Cepstrum, LSP (Line Spectrum Pair), and Filterbank energies are considered as feature vectors for speaker recognition. For pattern-matching methods, DTW, HMM, MLP, and VQ are usually engaged. These features and models have respectively individual characteristics. For each model, features from the same speaker do not yield the same results. A hybrid system engaging 3 recognizers such as HMM, MLP and DTW may reduce the recognition error rates caused by a recognizer.

3.2 Post-Recognizer

If result-scores of pre-recognizer are low, post-recognizer should be used as another recognizer. We use each speaker's pitch as an individual identity. Voiced speech is produced through vocal tract excitation with quasi-periodic vibrations (pitch) of the vocal folds at the glottis. The pitch period is one of the important factors for verifying. We obtain the pitch period using ACWT. Section 2 gives a full detail of post-recognizer.

3.3 Overall System Structure

This speaker verification system consists of HMM, MLP, DTW and Pitch Detection Part. HMM is discrete HMM. The structure of the system is shown in Fig. 8.

Three kinds of features individually are fed into three recognizers. Then, score is determined by the results of three recognizers. The result from the pitch detection part

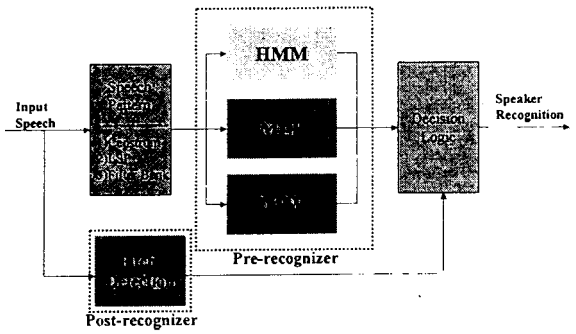


Figure 8. A block diagram of Hybrid Recognizer.

is also the criterion in scoring. The score in the decision logic part is calculated according to the following equation:

$$A=\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}=\begin{bmatrix} \alpha^H X_{HC} & \beta^H X_{HL} & \gamma^H X_{HF} \\ \alpha^M X_{MC} & \beta^M X_{ML} & \gamma^M X_{MF} \\ \alpha^D X_{DC} & \beta^D X_{DL} & \gamma^D X_{DF} \end{bmatrix} \quad (5)$$

where A is the output matrix of pre-recognizer.

$$D=\sum_{i,j=1}^{3} a_{ij} \quad (6)$$

where D is the score of pre-recognizer at the decision logic part.

X is the output of each pattern matching method in Eq. (5). The superscripts of weighting coefficients (α,β,γ), H, M and D stand for HMM, MLP and DTW, respectively. So do the first subscripts of X. The second subscripts of X stand for Cepstrum, LSP and Filterbank energy, respectively. The weighting coefficients are determined by experiments. The experiment procedure of hybrid SVS is described in section 4.1.

Ⅳ. Experimental Results

4.1 Procedure

1) Pitch detection by ACWT: calculating mean (P_m) and variance (P_v) of pitch of input speech.
2) If P_v < Threshold1 then goto next step, else goto step 1. (Small P_v means composed speech and the speech having little noise.)
3) Feature extraction (Cepstrum, LSP, and Filterbank energy)
4) Experiment of pre-recognizer (HMM, MLP, and DTW) with the three features: If all the 9 outputs of recognizers are same, then Valid Speaker, else goto next step.
5) If the outputs of more than Threshold2 are same, then compare the input speech pitch mean(P_{im}) and the enrolled speaker pitch mean (P_{em}).
6) If ($| P_{em} - P_{im} |/P_{em} \times 100$ < Threshold3) then Valid Speaker, else Invalid Speaker.

4.2 Experiments

Experimental data were sampled and quantized at 8 bits/sample. Three kinds of experiments were taken and each recognizer was trained with five speakers. System confidence was checked in three tests: FR, Denial-1, Denial-2. As for experiments, individual recognizer showed some error rate, but Hybrid recognizer yielded 0% error rate.

(1) 1st experiment: False Rejection

The 1st experiment checked the access of five valid speakers with valid password. The speakers should be accepted.

Table1. False Rejection experiment.

Speaker	Pre-recognizer		Post-recognizer		Hybrid System Total Error
	# of error	%	# of error	%	%
A	3/15	20%	0/3	0%	0%
B	0/15	0%			0%
C	0/15	0%			0%
D	0/15	0%			0%
E	0/15	0%			0%

Speaker A's speeches caused some error rate in the pre-recognizer, i.e., 3 speeches were rejected. The post-recognizer compensated the error and the hybrid system yielded no error.

(2) 2nd experiment : False Acceptance-Denial 1

The 2nd experiment checked the denial of invalid speakers with invalid password. The speakers should be rejected.

Table 2. False Acceptance-Denial 1.

Speaker	Pre-recognizer		Post-recognizer		Hybrid System Total Error
	# of error	%	# of error	%	%
F	0/25	0%			0%
G	2/25	8%	0/2	0%	0%
H	0/25	0%			0%
I	0/25	0%			0%
J	0/25	0%			0%
K	0/25	0%			0%
L	0/25	0%			0%
M	0/25	0%			0%

Speaker G's two speeches passed wrongly, but the post-recognizer rejected the passed speeches.

(3) 3rd experiment : False Acceptance-Denial 2

The 3rd experiment checked the denial of invalid speakers with valid password. The speakers should be rejected.

Table 3. False Acceptance-Denial 2.

Pre-recognizer		Post-recognizer		Total Error(Access)
# of error	%	# of error	%	%
0/130	100 %			0%

This total system confidence is shown as follows.

Table 4. Total System Confidence.

Success		Fail	
# of recognition	%	# of fail	%
405/405	100%	0/405	0%

Hybrid system produced no error in total 405 trials.

V. Conclusion

Some speech features such as Cepstrum, LSP, Filterbank energies, etc. represent their typical speech characteristics. And some recognizer models such as HMM, MLP, DTW, TDNN, etc. have their specific characteristics in recognition process. In designing speaker-verification system, hybrid recognizer is considered to be more reliable than single recognizer because it compensates for the recognition errors with the hybrid model.

In this study, a hybrid recognizer is engaged. It has two components: one is pre-recognizer with three features (Cepstrum, LSP, and Filterbank energies) and three recognizers (HMM, MLP, and DTW). The other is post-recognizer with pitch obtained by wavelet analysis as features.

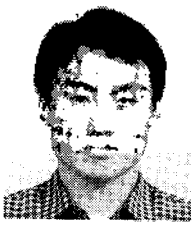
System confidence is checked in three kinds of tests: false rejection of a valid user with valid password, false acceptance of an invalid user with invalid password, and false acceptance of an invalid user with valid password. Individual recognizer shows some error rate but the hybrid recognizer results in 0% error rate. Its high feasibility is tested by experimental results from the task of hybrid recognizer using pre-recognizer and post-recognizer.

Furthermore, we show the designing method and the performance of adapted continuous wavelet transform to detect the pitches of input speech. This convincingly illustrates the utility of adapted continuous wavelet transform in a wide range of applications and will provide us with a perspective on pitch detection method.

References

1. H.J. Nam, H.S. Kim, Y. Kwon and S. Yang, "Speaker verification system using hybrid model with pitch detection by wavelets," *Proc. of the IEEE-SP Int. Sym. on TFTS*, pp. 153-156, October 1998.
2. L.R. Rabiner and B.H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993.
3. L. R. Rabiner and R.W. Schafer, *Digital Processing of Speech Signals*, Prentice Hall, 1978.
4. A. M. Kondoz, *Digital Speech, Coding for Low Bit Rate Communications Systems*, John Wiley & Sons, 1994.
5. P. Goupillaud, A. Grossman, and J. Morlet. *Cycle-octave and Related Transforms in Seismic Signal Analysis*. Geoexploration, vol. 23, Elsevier Science Pub., pp. 85-102, 1984/85.
6. A. Grossmann and J. Morlet, "Decomposition of Hardy functions into square integrable wavelets of constant shape," *SIAM Journ. of Math. Anal.*, vol. 15, no. 4, pp. 723-736, July 1984.
7. M.V. Wickerhauser, *Adapted Wavelet Analysis from Theory to Software*, AK Peters Wellesley, Massachusetts, 1994.

▲Hyoungsoo Kim



He received the B. S. and M. S. degrees in Control and Instrumentation Engineering from Hanyang University, Korea in 1997 and 1999, respectively. His research interests include digital speech processing and digital image processing with wavelet transform.

▲Sung-Il Yang

He was born in Goesan, Chung-buk, Korea in 1956. He received B. S. degree in Electronics Engineering with the greatest honors from Hanyang University, in Seoul, Korea, and his M. S. and Ph. D. degrees in Electrical & Computer Engineering from the University of Texas at Austin, Austin, Texas, in 1986 and 1989, respectively.

Since 1990, he has been with Dept. of Control and Instrumentation Engineering, Hanyang University. He is now an associate professor and his current research interests include speech recognition, and digital signal processing.

He is also a member of IEEE, Korea Institute of Telematics and Electronics, and the Acoustical Society of Korea.

▲Younghun Kwon

He was born in Seoul, Korea in 1961. He received B. S. degree in Mathematics from Hanyang University, in Seoul, Korea, and his M. S. and Ph. D. degrees in Physics from the University of Rochester at Rochester, NewYork, in 1986 and 1987, respectively.

Since 1995, he has been with Dept. of Physics, Hanyang University, Ansan, Korea. He is now an associate professor and his current research interests include mathematical physics, geometry, and artificial intelligence.

He is also a member of the Korean Physical Society, and the Korean Mathematical Society.

▲Kyung-Joon Cha

He was born in Seoul, Korea in 1961. He received B. S. degree in Mathematics from Hanyang University, in Seoul, Korea, and M. S. degree in Statistics from University of Wisconsin-Madison in 1985, and Ph. D. degree in Statistics from Southern Methodist University in 1990.

He had worked at Dept. of Statistics in Sejong University as associate professor and since 1993, he has been with Dept. of Mathematics in Hanyang University and currently he is an associate professor and his current interests include statistical computing, statistical modeling and neural network.

He is a member of the Korean Statistical Society.