# A Novel Spectral Conversion Based Approach for Noisy Speech Enhancement

Huy-Khoi DO, Trung-Nghia PHUNG, Huu-Cong NGUYEN, Van-Tao NGUYEN, and Quang-Vinh THAI, *Members, IACSIT*

***Abstract*—Present noisy speech enhancements algorithms are efficiently used for additive noise but not very good for convolutive noise as reverberation. And even for additive noise, the estimation of noise, when only one microphone source is provided, is based on the assumption of a slowly varying noise environment, commonly assumed as stationary noise. However, real noise is non-stationary noise, which difficult to be efficiently estimated. Spectral conversion can be used for predicting the vocal tract (spectral envelope) parameters of noisy speech without estimating the parameters of the noise source. Therefore, it can be applied to a general speech enhancement model, for both stationary and non-stationary additive noise environment, as well as convolutive noise environment, when only one microphone source is provided. In this paper, we propose a spectral conversion based speech enhancement method. The experimental results show that our method outperforms traditional methods.**

***Index Terms*—Speech Enhancement, speech denoising, spectral conversion, LP model**

## I. Introduction

Present single microphone speech enhancement algorithms are efficiently used for additive noise (white and colored) but not very good for convolutive noise as reverberation.

And even for additive noise, the estimation of noise, when only one microphone source is provided, is based on the assumption of a slowly varying noise environment, commonly assumed as stationary noise. However, real noise is non-stationary noise, which difficult to be efficiently estimated.

Although, multi-microphone models outperform single-channel models, the requirement of having more than one microphone in multi-microphone speech enhancement is not always impractical.

Therefore, developing a model for speech enhancement for both stationary and non-stationary additive noise environment, as well as convolutive noise environment, when only one microphone source is provided, is an important and interesting topic.

There are not many present models and algorithms can solve efficiently in this topic.

Spectral conversion is usually used in voice conversion methods. State of the art voice conversion is the GMM-based voice conversion, presented in section III.

Spectral conversion can be used for predicting the vocal tract (spectral envelope) parameters of noisy speech without estimating the parameters of the noise source [1]–[4]. Therefore, it can be applied to a general speech enhancement model, for both stationary and non-stationary additive noise environment, as well as convolutive noise environment, when only one microphone source is provided. Spectral conversion based speech enhancement was proposed in [5, 6], and developed in [1, 2, 3, 4].

Although spectral conversion is one promising method for speech enhancement, this kind of approach showed the two main drawbacks, making it has not attracted many researchers up to now.

The first drawback is the difficulty of source (*F0*) estimation in noisy environment, making it difficult to synthesize the enhanced speech. Therefore, it is difficult to directly use the spectral conversion concept in noisy speech enhancement methods

Vocal tract parameters normally can be combined with source parameters to synthesize the enhanced speech. In [6], the authors applied their model to alaryngeal speech, in which the source of distorted is easily estimated from the source of original speech. They did not apply their method for noisy speech enhancement because of the difficulty of source (F0) estimation in noisy environment.

Also due to the difficulty of estimating the source parameters in noisy environment, in [5], predicted vocal tract parameters are just used as a means for estimating the parameters of an "optimal" linear filter. The optimal filters, Wiener filter and Kalman filter, then are used in their speech enhancement method.

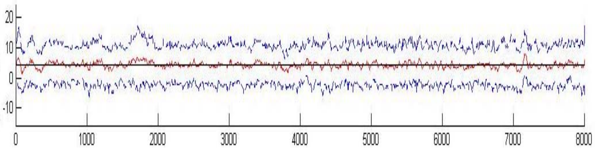


Fig. 1. Residual Gain Changing in Frequency Domain

The first drawback of spectral conversion based noisy speech enhancement can be overcome by using the method in [1, 2, 3], in which, instead of using traditional source/filter synthesis method to synthesize the restored speech, the BC speech (likes noisy speech) is filtered to AC speech (likes

Manuscript received November, 12, 2011; revised November 23, 2011.

Huy-Khoi Do and Van-Tao NGUYEN are with the Thai Nguyen University of Information and Communication Technology, Thai Nguyen, Vietnam (e-mail: dohuykhoi1981@yahoo.com , nvtao@ictu.edu.vn ).

Trung-Nghia PHUNG is with Japan Advanced Institute of Science and Technology, Ishikawa, Japan (email: ptnghia@jaist.ac.jp)

Huu-Cong NGUYEN is with Thai Nguyen University, Thai Nguyen, Vietnam (email: conghn@tnu.edu.vn).

Quang-Vinh THAI is with Institute of Information Technology, Vietnam Academy of Science and Technology (email: tqvinh@ioit.ac.vn ).

clean speech) by LP parameters of the source and the target. This method does not use any source (*F0*) estimation for BC speech; therefore, it has efficiently restored the AC speech from BC speech. Detail of this method will be presented in section II.

The second drawback of spectral conversion based speech enhancement is the common issue in learning-based speech enhancement, in which the spectral parameters are not robust and the trained parameters therefore are largely different between noisy conditions. It causes the inefficiency of enhancement systems when the testing conditions are not match with the training conditions.

To solve this problem, in this paper, we propose the use of the perceptual LSF instead of using LPC and LSF spectral parameter. The detail will be presented in section IV. The experimental results show that our method outperforms traditional methods.

Some rest issues of spectral conversion based speech enhancement are the need of large training data and speaker independent adaptation, which are known as the common issues of speech recognition systems, will not be solved in this paper.

## II. LP-based Model for Noisy Speech Enhancement

In this section, we present the LP-based model proposed for BC speech restoration in [1, 2, 3] and applied for noisy speech enhancement in this paper.

Let *x(t)* and *y(t)* be clean (AC in [4]) and associated noisy (BC in [4]) speech. Using LP analysis, we can present the discrete *x(n), y(n)* as,

$$\hat{x}(n) = -\sum_{i=1}^{p} a_i x(n-i), \hat{y}(n) = -\sum_{i=1}^{p} a_i y(n-i) \quad (1)$$

where $\hat{x}(n)$ is the predicted signal value, *x(n − i)* the previous observed values, and ai the predictor coefficients (the LPC coefficients). The residual is obtained by the error between the current and the predicted samples

$$g_x(n) = x(n) - \hat{x}(n), g_y(n) = y(n) - \hat{y}(n) \quad (2)$$

The two correspondent discrete *x(n)* and *y(n)* are represented by LP model in the Z-domain as below.

$$-G_x(z) = X(z)\sum_{i=0}^{P} a_x(i)z^{-i}, a_x(0) = -1 \quad (3)$$

$$-G_y(z) = Y(z)\sum_{i=0}^{P} a_y(i)z^{-i}, a_y(0) = -1 \quad (4)$$

where X(z) and Y(z) are the Z-transforms of *x(n)* and *y(n)*, P are the LP orders, and *ax(i), ai(y)* are the *ith* LPC coefficients. Here, *Gx(z)* and *Gy(z)* are the Z-transforms of the LP residuals *rx(n)* and *ry(n)*.

We defined the residuals ratio of *x(n)* and *y(n)* in Z (or frequency domain) as the gain k

$$k = G_x(z)/G_y(z) \text{ or } k = G_x(w)/G_y(w) \quad (5)$$

Since the LP residuals *gx(n)*, *gy(n)* are related to the glottal source information of *x(n)* and *y(n)*, this kind of information may unchanged between clean (AC) and associated noisy (BC) speech signals. Figure 1 shows the values of the gain k in frequency domain of a speech sample. It reveals that this gain is approximate 5 dB around its average value over all frequencies. This result suggests that the residuals ratio in frequency domain is quite constant and can be fixed as an average constant.

Let us assume that the mathematical description of transfer function h(n) from x(n) to y(n) is an M-order FIR filter. In the Z domain, it is represented as

$$H(z) = \frac{Y(z)}{X(z)} = \sum_{i=0}^{M} h(i)z^{-i} \quad (6)$$

We can obtain the equation for H-1(z) as

$$H^{-1}(z) = \frac{1}{H(z)} = k.\frac{\sum_{i=0}^{P} a_y(i)z^{-i}}{\sum_{i=0}^{P} a_x(i)z^{-i}} \quad (7)$$

The noisy (BC) speech can be restored to the associated clean (AC) speech by using this inverse filtering function $H^{-1}(z)$.

LSF is used to encode LP spectral information more efficiently than other LP parameters. Let A(z) be a general LP filter on an LP representation, the LSF coefficients can be derived from a symmetric polynomial and an anti-symmetric polynomial, U(z) and V(z), as the phase of conjugated zeros.

$$A(z) = \sum_{i=0}^{P} a(i)z^{-i}, a(0) = -1 \quad (8)$$

$$U(z) = A(z) + z^{-(P+1)}A(z^{-1}), \quad (9)$$

$$V(z) = A(z) - z^{-(P+1)}A(z^{-1}), \quad (10)$$

Substituting Eqs. (8)-(10) into Eq. (7), we can obtain the equation for the inverse filtering as

$$H^{-1}(z) = k\frac{U_y(z) + V_y(z)}{U_x(z) + V_x(z)} \quad (11)$$

Here, *(Uy(z), Vy(z))* and *(Ux(z), Vx(z))* are symmetric polynomial and an anti-symmetric polynomial for noisy (BC) and clean (AC) speech that are determined from LSF coefficients.

The inverse filtering therefore depends on the LSF coefficients of clean and noisy speech and the gain k.

## III. The GMM-based Spectral Conversion

GMM and Neural Network (NN) are the most efficiently used for training in spectral conversion [1, 2, 3, 4, 7, 8].

It is impractical to use NN for training in huge corpus. Due to the over-training problem of NN, NN seems not suitable to train various kinds of noisy speech in one training session. Another problem of NN training is that it is difficult to adapt the NN model to unknown noise kinds of noisy speech. This problem makes difficult to build the BC restoration for open dataset noisy speech.

The results in [2, 3] are also confirmed that GMM training outperforms NN training. Therefore, in this research, we used GMM for training the joint spectral vectors of noisy and clean speech. In this section, we present the training and predicting procedure using GMM based voice conversion that we use for our noise speech enhancement.

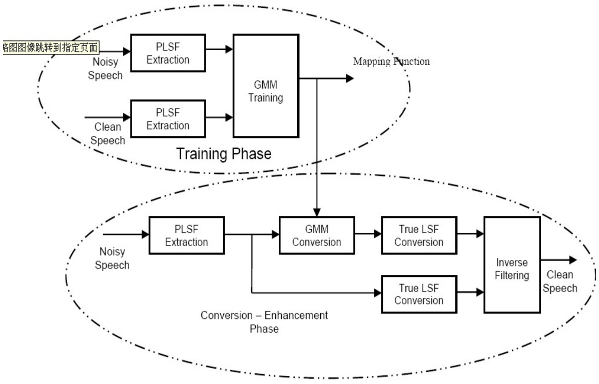


Fig. 2. General Diagram

### A. Training Procedure

The source speech is represented by a time sequence X = [*x1, x2,..., xn*], where xi is a D dimensional feature vector for the i-th frame,i.e. xi =[*x1,x2,...,xD*]$^T$. The target speech is represented by a time sequence Y =[*y1, y2,..., yn*], where yj =[*y1,y2,...,yD*]$^T$. The joint source-target vector Z =[*z1, z2,..., zn*] where zq =[*xTi* , *yTj* ]$^T$.

The distribution of Z is modeled by Gaussian mixture model, as in Eq.(12).

$$p(z) = \sum_{m=1}^{M} \alpha_m \mathrm{N}(\mathrm{z};\mu_{\mathrm{m}},\textstyle\sum_{\mathrm{m}}) = p(x,y) \quad (12)$$

where M is the number of Gaussian components. N(z; μm, Σm) denotes the 2D dimension normal distribution with the mean μm and the covariance matrix Σm. αm is the prior probability of z having been generated by component m, and it satisfies 0 ≤ *αm* ≤ 1, $\sum_{m=1}^{M} \alpha_m = 1$ .The parameters (*αm,μm,Σm*) for the joint density p(*x,y*) can be estimated using the expectation maximization (EM) algorithm.

### B. Predicting Procedure

The transformation function that converts source feature x to target feature y is given by Eq.(13).

$$F(x) = E(y\,|\,x) = \int yp(y\,|\,x)dy \quad (13)$$

So,

$$F(x) = \sum_{m=1}^{M} p_m(x)(\mu_m^y + \textstyle\sum_m^{yx}(\sum_m^{xx})^{-1}(x-\mu_m^x)), \quad (14)$$

where $p_m(x) = \dfrac{\alpha_m N(x;\mu_m^x,\sum_m^{xx})}{\sum_{m=1}^{M}\alpha_m N(x;\mu_m^x,\sum_m^{xx})}$ , $\mu_m = \begin{pmatrix} \mu_m^x \\ \mu_m^y \end{pmatrix}, \sum_m = \begin{pmatrix} \mu_m^{xx} & \mu_m^{xy} \\ \mu_m^{yx} & \mu_m^{yy} \end{pmatrix}$ and p$_m$(x) is the probability of x belonging to the m$^{th}$ Gaussian component.

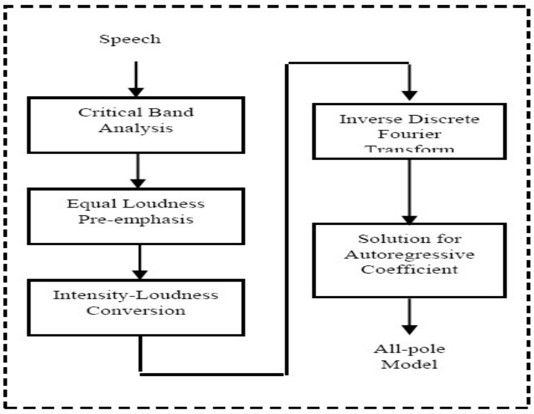

Speech
Critical Band Analysis
Equal Loudness Pre-emphasis
Intensity-Loudness Conversion
Inverse Discrete Fourier Transform
Solution for Autoregressive Coefficient
All-pole Model

Fig. 3. Perceptual Linear Preditive (PLP)

## IV. Perceptual Line Spectral Frequency (PLSF)

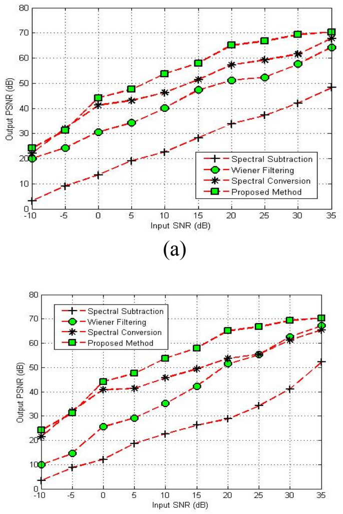
In spectral conversion of noisy speech, if the training is on data with an environment identical to that for the test data, it can usually convert spectral parameters of noisy speech to those of clean speech for both additive and convolutive noise. Unfortunately, the noise is seldom known in advance.

When the data from different environments is used in training and test, the same recognizer typically performs much worse. Our goal is to understand and eliminate variance in the speech signal due to the environmental changes and thus ultimately avoid the need for extensive training of the conversion system in different environments.

The original PLP (Perceptual Linear Predictive) was proposed in [8], in which the true spectrum is converted to the form closed with human hearing before computing the LPC coefficients. Two main techniques are used, including critical-band spectral resolution and equal-loudness pre-emphasis and compression. The PLP algorithm is shown in the figure.3.

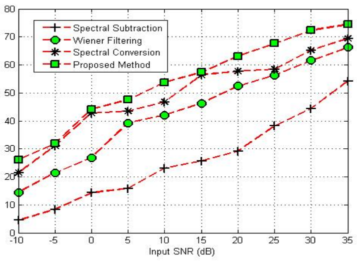
The PLP has been shown as more robust to noise than original LPC. However, same as the original LPC, PLP coefficients seem to be inappropriate for statistical models as GMM because of their relatively large dynamic range.

Therefore, in this paper, we convert the PLP into LSF representation in order to encode LP spectral information more efficient in statistical GMM conversion.



(a)

(b)



(c)

Fig. 4. Objective Evaluation Results for (a) white (b) pink (c) factory noisy speech

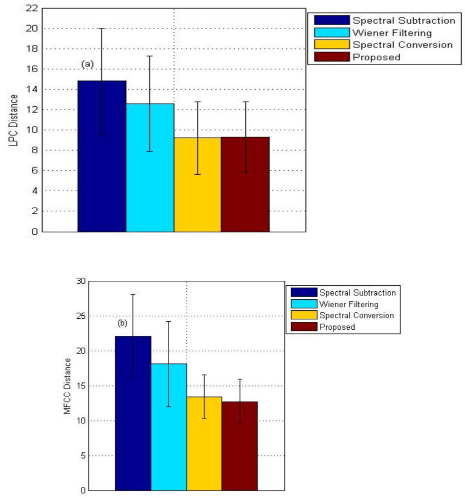


Fig 5. a, LPC Distance and b, MFCC distance

## V. Implementation and Evaluations

### *A. Implementation*

The general diagram of our proposed method is depicted in the figure.2. We use the PLSF parameters for LP-based model. In our experiments, the number of Gaussian components M, which should be chosen large enough if we have enough data for training, is chosen as 15, the order of LP analysis P is chosen as 20.

### *B. Data Preparation*

The clean speech data used in our evaluation is the TIMIT English database and our Vietnamese dataset, which recorded from 20 speakers, including 10 males and 10 females. All speakers are native Vietnamese students. To build the noisy speech database, we add the noise from the NOISEX-92 database to our clean speech database. The noisy data includes three kinds of noise, which is factory, pink and white noise. The data used for objective evaluation is the TIMIT. Due to the time consuming limitation, we just evaluated the recognition scores of the LP-based methods with the white noise, SNR = -10dB, used for Vietnamese dataset.

### *C. Objective Evaluations for Voice Quality of Enhanced Noisy Speech*

To evaluate our method, we implement and compare our method with the standard spectral subtraction method [9] and Wiener-filter based method [10], in which all of them are non-learning based methods. We also compare our method with the standard spectral conversion method [7] applying for noisy speech enhancement.

We used PSNR (Peak Signal to Noise Ratio) to objectively comparatively evaluate proposed method.

The PSNR is defined as:

$$PSNR = 10 * \log(\frac{NX^2}{\|x-r\|^2}) \qquad (15)$$

where N is the length of the enhanced signal, X is the maximum absolute square value of the signal x and $\|x-r\|^2$ is the energy of the difference between original and enhanced signals. The results are show in the figure.4 a, b, c. The results reveal that our proposed method outperforms the others.

### *D. Objective Evaluations for performance in ASR of Enhanced Noisy Speech*

One of interesting application of noisy speech enhancement is pre-processor for noisy automatic speech recognition (ASR). Because state of the art ASR use LPC and MFCC features [11], we defined the LPC distance (LCD) and MFCC distance (MCD) similar to those in [1], to evaluate the denoised performance of our proposed method in ASR.

$$LCD = \sqrt{\sum_{i=1}^{P}(a_x(i) - a_y(i))^2} \qquad (16)$$

$$MCD = \sqrt{\sum_{i=1}^{Q}(c_x(i) - c_y(i))^2} \qquad (17)$$

where $a_x(i)$ and $a_y(i)$ are the LPCs, $c_x(i)$ and $c_y(i)$ are the MFCCs of the denoised and the original clean speech for evaluation. P is the LP order and Q is the cepstral order.

State of the art ASRs use not only primary LPC, MFCC coefficients but also those delta coefficients. Therefore, we compare both of mean and variance of LCD, MCD of the methods. The experiment is conducted in only factory noisy speech environment

The results are show in the figure .5 a, b. The results reveal that spectral conversion based methods outperform the traditional spectral subtraction and wiener filtering. Our proposed seems not improving the ASR performance in comparison with standard spectral conversion method.

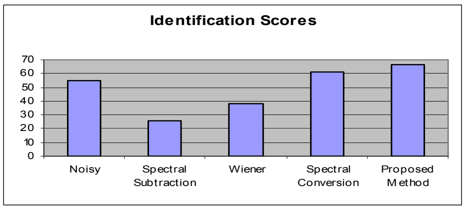
### *E. Subjective Evaluations for Speech Intelligibility*

For English TIMIT database, we carried out the subjective tests with five subjects who had normal hearing. All are native English foreign teachers in our university. For Vietnamese dataset, we carried out the subjective tests with ten subjects who had normal hearing; all are native Vietnamese student in our university.

The speech signals of English and Vietnamese syllables were played in random order in the tests. The subjects had not heard these syllables previously and had not been trained before the experiment. They were asked to listen to each word only once and write down what they heard. Speech intelligibility could generally be evaluated using the average recognition accuracy scored by all subjects.

The results are shown in the figure.6 for English (a) and

Vietnamese (b). It reveals that for both of English and Vietnamese, while non-learning based methods degrade intelligibility of speech, learning based methods improve intelligibility. In addition, our proposed method improves the speech intelligibility of enhanced speech in comparison with standard spectral conversion method.



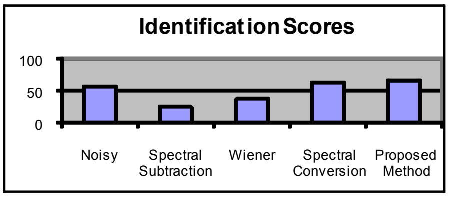


Fig. 6 Subjective Evaluation Results for (a) English (b) Vietnamese

## VI. Conclusion

In this paper, we proposed a spectral conversion based noisy speech enhancement method using our LP based model. We conducted both objective and subjective evaluations. The experimental results show that our proposed method improves both of voice quality (PSNR) and intelligibility of speech.

## References

[1] T. T. Vu, K. Kimura, M. Unoki, and M. Akagi, "A Study on Restoration of Bone-conducted Speech with MTF-based and LP-based Models," "*Journal of Signal Processing*", vol. 10, no. 6, pp. 407-417, 2006.

[2] T. N. Phung, M. Unoki, and M. Akagi , "Comparative Evaluation of Bone-conducted-speech Restoration based on Linear Prediction Scheme," "*IEICE Technical Report*", vol. 110, no. 71, pp. 53-58, June, 2010.

[3] T. N. Phung, M. Unoki, and M. Akagi, "Improving Bone-Conducted Speech Restoration in noisy environment based on LP scheme," "*Proc. APSIPA 2010*", Dec, Singapore 2010.

[4] H. K. DO and Q. V. THAI, "A new approach for speech denoising using spectral conversion," "*Proc. ICSPS 2011*", August, Yantai, China.

[5] A. Mouchtaris, J. V. Spiegel, P. Mueller, and Panagiotis Tsakalides, "A Spectral Conversion Approach to Single-Channel Speech Enhancement," "*IEEE Trans On Audio, Speech, And Language Processing*", vol.15, no.4, May 2007

[6] N. Bi and Y. Qi, "Application of Speech Conversion to Alaryngeal Speech Enhancement," "*IEEE Transactions On Speech And Audio Processing*", vol.5, no.2, March 1997.

[7] A. Kain and M. W. Macon, "Spectral Voice Conversion For Text-To-Speech Synthesis," "*Proc. ICASSP 1998*", vol. 1, pp. 285-288, 1998

[8] H. Hermansky., Perceptual linear predictive (PLP) analysis for speech. "*J. Acoust. Soc. Am*", pp. 1738-1752, 1990.

[9] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," "*IEEE Trans on Acoustics, Speech and Signal Processing*", vol 27 Issue:2, pp. 113 – 120, Apr 1979,.

[10] J. S. Lim and A. V. Oppenheim, "Enhancement and band width compression of noisy speech," "*Proc. of the IEEE*", Vol. 67, No. 12, 1586–1604, Dec. 1979.

[11] L. Rabiner and B. H. Juang, "*Fundamental of Speech Recognition*", Copyright 1993 by AT&T.

**Huy-Khoi DO** received his B.E in Electronics and Telecommunications from Vietnam Nationnal University – Hanoi in 2003 and his M.E in Electronics and Telecommunications from Le Quy Don Technical University in 2006. He is currently a lecturer of Thai Nguyen University of Information and Communication Technology. He is also a PhD candidate at the Institute of Information Technology, Vietnam Academy of Science and Technology. He research interest is signal processing in automatic control and communication.

**Trung-Nghia PHUNG** received his B.E in Electronics and Telecommunications from Ha Noi University of Technology in 2002, his M.Sc in Telecommunications from Vietnam National University - Hanoi. in 2007. He has been a lecturer at the Thai Nguyen University of Information and Communication Technology. He is currently a PhD candidate at Japan Advanced Institute of Science and Technology. He research interests are speech and acoustic signal processing and applications.

**Huu-Cong NGUYEN** received his M.Sc in Automatic Control in 1997, and his PhD from Hanoi University of Technology on control theory and optimal control, in 2003. He is currently and Associate Professor and Vise Director of Thai Nguyen University. He research interests are optimal control of distributed parameter systems, soft computing.

**Van-Tao NGUYEN** received his B.S in Mathematics from Thai Nguyen University of Pedagogy in 1994, his Msc from Vietnam National University Hanoi in 2000 and his PhD from Vietnam Academy of Science and Technology in 2009. He is currently Vise Rector of Thai Nguyen University of Information and Communication Technology. He research interests are image, audio watermarking and processing, software engineering,, and information security.

**Quang-Vinh THAI** received his Engineering degree in Automatic Control from Technical University of Odessa, USSR in 1977, and his PhD degree from Moscow Power Engineering Institute, Moscow, Russia in 1991. He is currently Professor and Dean of Institute of Information Technology, Vietnam Academy of Science and Technology. He research interests are fuzzy control, automatic control and automation application.