International Journal of Reconfigurable and Embedded Systems (IJRES)
Vol. 9, No. 2, July 2020, pp. 109~115
ISSN: 2089-4864, DOI: 10.11591/ijres.v9.i2.pp109-115

Efficient robust speech recognition with empirical mode decomposition using an FPGA chip with dual core

Shing-Tai Pan[1], Ching-Fa Chen[2], Wen-Sin Tseng[3]
[1,3]Department of Computer Science and Information, National University of Kaohsiung, Taiwan
[2]Department of Electronic Engineering, Kao Yuan University, Taiwan

Article Info

Article history:

Received Jan 01, 2020
Revised Feb 14, 2020
Accepted Feb 28, 2020

Keywords:

Field Programmable Gate Array (FPGA)
Multi-Core Embedded System
Empirical Mode Decomposition
Hidden Markov Model
Speech Recognition

ABSTRACT

The purpose of this paper is to accelate the computing speed of Empirical Mode Decomposition (EMD) based on multi-core embedded systems for robust speech recognition. A reconfigurable chip, Field Programmable Gate Array (FPGA), is used for the implementation of the designed system. This paper applies EMD to discompose some noised speech signals into several Intrinsic Mode Functions (IMFs). These IMFs will be combined to recover the original speech by multiplying their corresponding weights which were trained by Genetic Algorithms (GA). After applying Empirical Mode Decomposition (EMD), we obtain a cleaner speech for recognition. Due to the complexity of the computation of the EMD, a dual-core architecture of embedded system on FPGA is proposed to accelerate the computing speed of EMD for robust speech recognition. This will enhance the efficiency of embedded speech recognition system.

Corresponding Author:

Shing-Tai Pan,
Department of Computer Science and Information
Engineering, National University of Kaohsiung,
700 Kaohsiung University Road, Nanzih Dist., Kaohsiung City 811, Taiwan, ROC.
Email: stpan@nuk.edu.tw

1. INTRODUCTION

It has been a long time for the development of speech recognition. However, speech recognition for the speech subjects to environmental noise is still an open problem. The most important problem in the robust speech recognition is the mismatch problem arising from the mismatch of the training and application environment due to the noises. Consequently, a speech sensor with the ability of noises cancellation is important for the realization of robust speech recognition. The methods for handling the mismatch problem can be classified into two categories: feature-based method and model-based method. Feature-based methods focus on the feature parameters rather than on model parameters for speech or noise [1-7]. Model-based methods exploit prior knowledge about the distributions of speech and noise for speech feature enhancement [8-13]. In this paper, the noised speech signals will be processed by eliminating the noise components before capturing the features through Mel-Frequency Cepstrum Coefficient (MFCC). Hence, the speech features become cleaner when they are fed into the speech recognition platform for recognition. A better recognition rate for the noised speech signal can then be obtained.

This study applies the EMD to decompose noised speech signals into the components including speech signals or noises. EMD is first proposed by Prof. Huang to combine the Hilbert Transform (HT) to analyze the nonlinear and non-stationary time series. The combination of EMD and HT is then called Hilbert Huang Transform (HHT) [14]. The EMD was applied initially on the signal analysis of in the area of geoscience, strength analysis of material structure and the trend analysis of the stock market, etc. Hence, it is

of the goal of this paper to find the weights corresponding to different IMFs and combines these weighted IMFs to recover the original speech signals. The weights for each IMF are trained by GA to find an optimal combination of IMFs. However, since EMD process will cost a lot of computation time, another goal of this paper is to implement a dual-cores architecture on an FPGA to accelerate the operation of EMD.

2. EMPIRICAL MODE DECOMPOSITION (EMD)

In this section, the procedure for performing EMD is introduced. Besides, a strategy based on GA and EMD to the robust speech recognition is proposed.

2.1. Procedure of EMD operations

The main step to perform EMD operation is to divide a speech signal into several intrinsic mode functions (IMFs). The condition for the data series to be an IMF can be described as follows [14] .Let the original signal is X (t) and Temp(t) = X (t) .

Step 1: Find the upper envelop U (t) and lower envelop L(t) of the signal Temp(t). Calculate the mean of the two envelops m(t) = [U (t) + L(t)]/ 2 . The component of Temp(t) is obtained by the equation

$$h(t) = Temp(t) - m(t). \quad (1)$$

Step 2: Check whether the signal $h(t)$ satisfies the conditions of IMF or not. If it is, then the first IMF is obtained as $imf_1(t) = h(t)$ and go to next step, else assign the signal $h(t)$ as $Temp(t)$ and go to Step 1

Step 3: Calculate the residue $r_1(t)$ as

$$r_1(t) = Temp(t) - imf_1(t) \quad (2)$$

Assign the signal $r_1(t)$ as $X(t)$ and repeat Step 1 and Step 2 to find $imf_2(t)$

Step 4: Repeat Step 3 to find the subsequent IMFs as follows.

$$r_n(t) = r_{n-1}(t) - imf_n(t) \text{ , } n = 2, 3, 4, \ldots. \quad (3)$$

This step is end when the signal $r_n(t)$ is constant or a monotone function. After the EMD procedure Step 1~ Step 4 is finished, the following decomposition of $X(t)$ is obtained.

$$X(t) = \sum_{i=1}^{n} imf_i(t) + r_n(t).$$

2.2. Combining GA with EMD for noise separation

In order to illustrate the effect of noises on IMFs, the EMD for a clean speech signal is first performed and the obtained IMFs are shown in Figure 1 [15]. In this figure, the leading five IMFs are shown, since the speech signal almost totally exists in these IMFs. Beyond these IMFs, it is hardly to find any speech signal components. It can be seen that the later the order of IMF is the lower the frequencies is. In order to examine that which IMF the noise or the speech signal will exist, a white noise is added into the clean speech signal. And then the IMFs, obtained from the EMD for a noised speech signal, are shown in Figure 2 [15]. Based on Figure 2, it is easy to find that the noise almost exists in the 1st IMF. Moreover, comparing Figure 1 to Figure 2, we can find that the 2nd and 3rd IMFs of the noised speech are very similar to the corresponding IMFs of the clean speech. So, we conclude that the speech signal mostly exists in the 2nd and 3rd IMFs. However, some experiments reveal that there are still some components of speech exist in later IMFs than 2nd and 3rd IMFs. Indeed, from a numerical experimental results from my previous work shown in Table 1 [15], it can be seen that the speech components exists in the later IMFs more evidently when the magnitude of the added noise becomes larger. This experiment reveals that the previous works on EMD for speech signal, which used only 2nd and 3rd IMFs to recover the original speech signal will lose some speech components in later IMFs. Thus, this paper asserts that the later IMFs should be included by multiplying some weights to recover the original signal. Actually, the weights for the each IMFs to recover the original speech are variant for different SNR of a noised speech. Consequently, this paper proposed a strategy which uses GA to train the optimal weighting of IMFs to recover the speech signal subject to various strength of noise. In the training phase of the weights for each IMFs, the chromosomes in GA are defined as $Chrm = [w_1\ w_2\ \ldots\ w_n]$ and the recovered speech is then expressed as:

$$X_{en}(t) = \sum_{i}^{n}{}_{=1} w_i \cdot imf_i(t). \quad (4)$$

The fitness function for the GA used in this study is defined as follows.

$$fitness = \frac{1}{MSE+1} \quad (5)$$

$$MSE = \frac{1}{k_u \times i_u \times t_u} \sum_{k=1}^{k_u} \sum_{i=0}^{i_u} \sum_{t=1}^{t_u} \left| E_{k,i,t} \right|^2 \quad (6)$$

in which $E_{i,k,t}$ means the output error for t^{th} record of i^{th} literal by k^{th} person

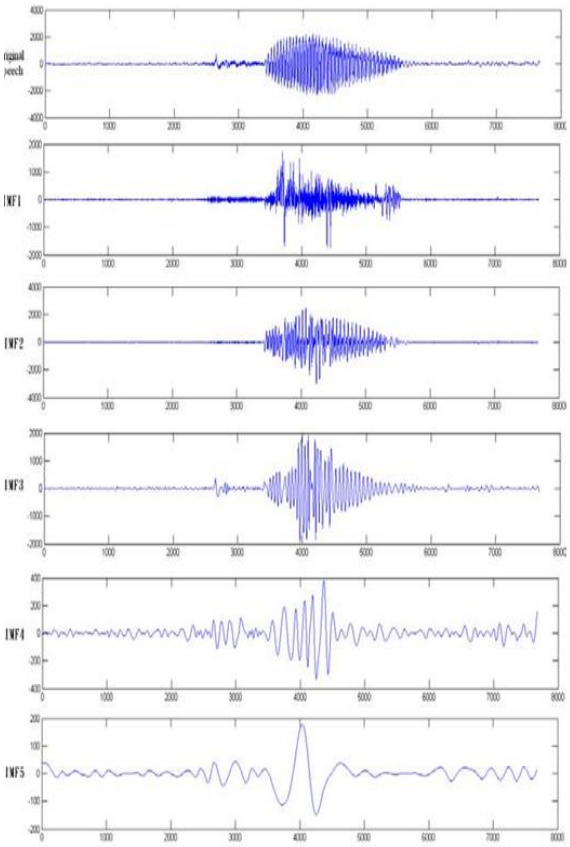

Figure 1. IMFs for clean speech signal [15]

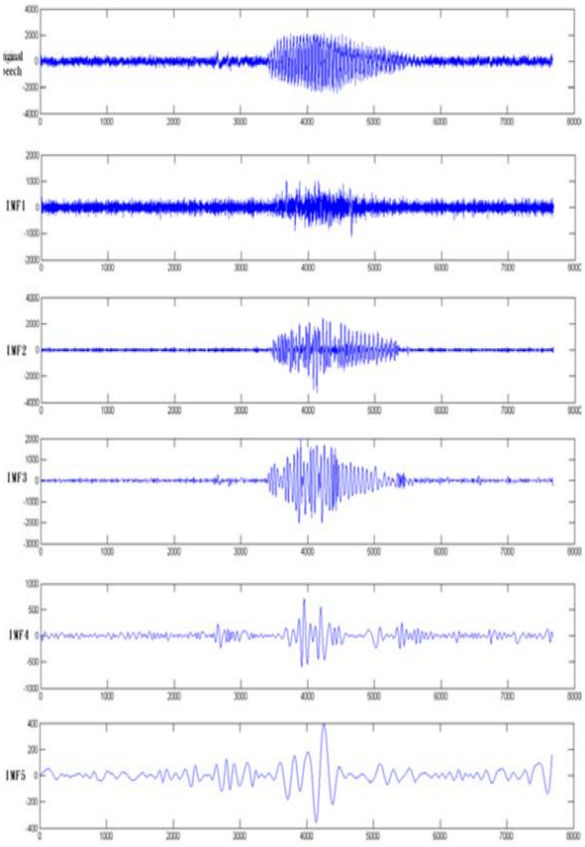

Figure 2. IMFs for noised speech signal [15]

Table 1. Speech recognition rates of noised speech signals for various SNR versus various IMFs combinations

SNR(dB) / Recog. Rates / Comb of IMFs	∞	25	20	15	10	5	0
without EMD	90.0	75.3	68.6	61.3	40.3	15.3	8.0
imf_1	82.6	52.3	38.6	24.6	17.3	12.0	10.0
imf_2	92.6	88.6	78.3	63.0	56.0	42.6	28.0
imf_3	64.0	53.3	44.0	34.6	26.0	19.3	14.6
imf_1+imf_2	96.0	92.0	80.6	61.3	40.0	22.0	11.7
imf_2i+mf_3	94.6	86.0	80.6	67.3	56.6	51.3	36.6
imf_2 +..+imf_5	93.3	88.6	79.3	69.0	56.6	52.6	37.3

3. IMPLEMENT THE DESIGNED SPEECH RECOGNITION SYSTEM ON FPGA

In this paper, the developed noises cancellation method for speech sensors and the speech recognition system was implemented on a FPGA-based SOC embedded platform. The block diagram of the SOC architecture used in this study is shown in Figure 3. An Altera develop board DE2-70 in which a Cyclone FPGA chip is included is used for this experiment. The constraint on the development board for this experiment is that there is no operation system in the FPGA chip, only single-threaded procedure is available. This will slow down the computation speed of the speech recognition systems. On the board a push button is used for the control of the starting and ending of the voice record and a Toggle switch is used for controlling the sampling rate of AUDIO codec. The EDA tools Quartus II, SOPC Builder and Nios II are used to develop and simulate the system. In the hardware implementation, SRAM and Flash RAM are used for the storage of source code and testing signal, respectively. The I2C Protocol is used to control the register of the platform. Besides, AUDIO Controller is used to receive speech data and SEG7 is used for the display of recognition results. The standard control IPs which are supported by SOC Builder are adopted for driving the necessary elements SDRAM, SRAM and LCD. The push button and Toggle switch are connected by the built-in PIO. Moreover, SEG7 Controller and AUDIO Controller are user-defined. In the experiment, PLL is adjusted to have a frequency of 100Mhz and a delay of 3ns, and then support the system's clock

The specs of the FPGA is as follows.

- Cyclone II 2C70 FPGA
- 70,000 LEs
- 2-Mbyte SSRAM
- Two 32-Mbyte SDRAM
- 8-Mbyte Flash memory

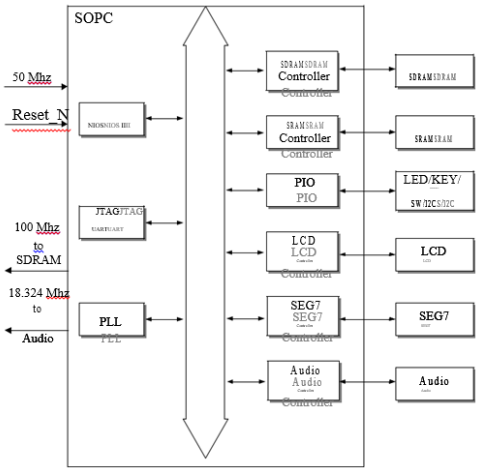

Figure 3. Block diagram of the SOC system in this experiment

3.1. Dual Core Realization on FPGA

This paper used two Nios II/f fast cores with 100MHz clock for each CPU to implement the proposed system. The specs of NIOS II can be seen in Figure 4.

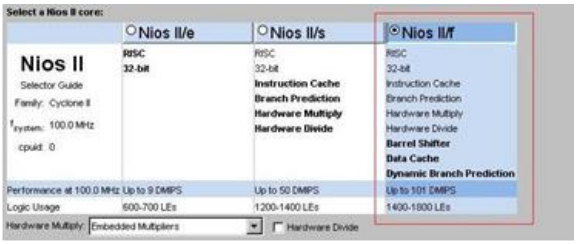

Figure 4. The specs of NIOS II/f in a FPGA

There are two 32-Mbyte SDRAMs in the embedded platform used in this experiment. The test speeches and all the parameters for speech recognition, that is the parameters for the classifier (HMM here), the EMD and the GA, were stored in SDRAM 1 while the sharing data which are shared for the two CPUs were stored in SDRAM 2. The memory allocation of SDRAM 1 for the booting program of the two CPUs, called CPU1 and CPU2, is depicted in Figure 5. Moreover, the memory allocation of SDRAM for the parameters, functions and data which are necessary for the operations of CPU1 and CPU2 is depicted in Figure 6. The details for each segment are listed as follows:

- .text－ the execution codes
- .rodata－ the read-only data
- .rwdata－ variables and pointers
- .heap－ dynamic allocation of memory
- .stack－parameters of function call and data for temporary variables

The shared memory of the two CPUs is managed by a software MUTEX CORE which is supported by SOPC TOOLS of Altera Company.

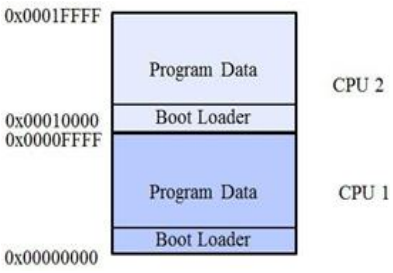

Figure 5. Memory allocation for the booting programs of the two CPUs

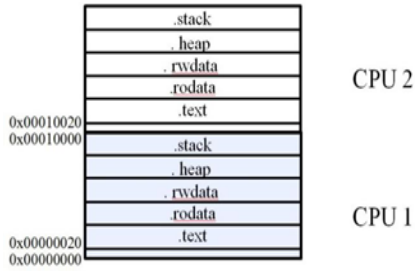

Figure 6. Memory allocation for the parameters, functions and data of the two CPUs

The parallel process of EMD by the two CPUs is fulfilled by saparating a speech signal into two parts. The first part and the second part are sent to CPU1 and CPU2 for performing the EMD process, respectively. After the EMD process for the two parts are completed, CPU1 accesses the share memory for the EMD results and then performs the speech recognition by the algprithm implemented in FPGA. The cooperation of the two CPUs is depicted in Figure 7.

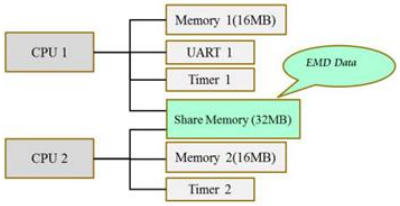

Figure 7. The cooperation of dual cores

4. EXPERIMENTAL RESULTS

Ten speeches 0~9 are recorded with 8kHz, 16 bits length and monotone. For GA, each generation has 16 chromosomes, the survival rate for each chromosome is 0.5, and the mutation rate for each gene is 0.05. Besides, the chromosomes in parent generation are randomly crossover to generate the chromosomes of the next generation. The tables, Table 2 and Table 3 reveal the time cost for the EMD and speech recognition by using single core and dual core, respectively. Moreover, the word-by-word comparisons of computation time by using single-core architecture and dual-cores architecture for EMD and speech recognition are depicted in Figure 8 and Figure 9, respectively. It is obviously that the time cost by using dual cores is much less than that by using single core. The percentages for time saving for each speech are listed in Table 4. According to Table 4, the percentage for saving time by using dual core for EMD process are in the range from 23.55% to 49.85%, and that for recognition are in the range from 18.22% to 45.20%. The average

saving percentage for EMD is 41.39% and is 35.88% for recognition. This shows that a dual-cores architecture can speed up the EMD process and speech recognition.

Table 2. Time costs for EMD process and speech recognition by using single core

	Time cost (average)		
Speeches	**EMD (sec.)**	**Recognition (sec)**	**Recognition Rate**
0	30.80	32.97	96%
1	50.78	53.22	85%
2	34.97	36.85	98%
3	12.23	13.65	100%
4	33.34	35.23	100%
5	28.69	31.18	98%
6	18.77	20.50	96%
7	28.81	30.65	92%
8	18.84	20.52	86%
9	20.77	22.65	98%
Average	***27.80***	***29.74***	***94.9%***

Table 3. Time costs for EMD process and speech recognition by using dual core

	Time cost (average)		
Speeches	**EDM (sec.)**	**Recognition**	**Recognition Rate**
0	17.46	20.40	96%
1	26.35	29.71	81%
2	17.98	20.59	98%
3	8.60	10.59	98%
4	16.72	19.30	98%
5	15.94	19.34	98%
6	14.35	16.77	94%
7	15.43	17.99	98%
8	11.42	13.72	88%
9	12.35	14.93	96%
Average	***15.66***	***18.33***	***94.5%***

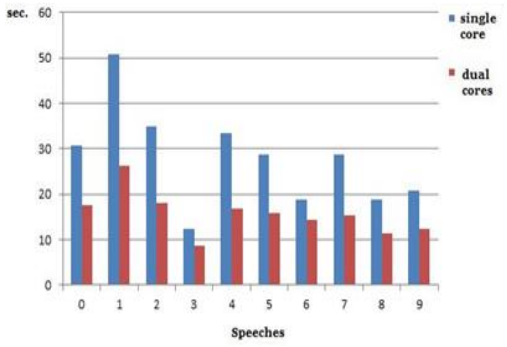

Figure 8. Comparison of computation time for EMD using single core and dual cores

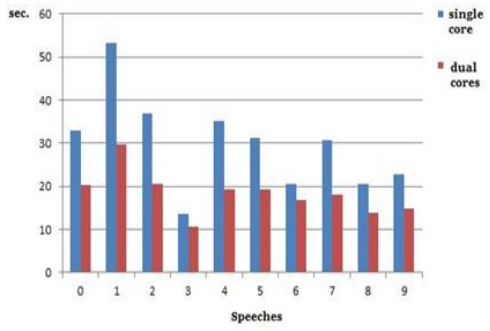

Figure 9. Comparison of computation time for speech recognition using single core and dual cores

Table 4. Percentage for saving time by using dual core for EMD process and recognition

Speeches	**Saving Percentage for**	**Saving Percentage for**
0	43.32%	38.12%
1	48.11%	44.17%
2	48.59%	44.13%
3	29.65%	22.41%
4	49.85%	45.20%
5	44.43%	37.98%
6	23.55%	18.22%
7	46.44%	41.31%
8	39.39%	33.13%
9	40.53%	34.12%
Average	***41.39%***	***35.88%***

5. CONCLUSION

The dual core architecture for accelerating the computation time of EMD is proposed in this paper. The EMD combination with GA is used here for noise separation from a contaminated speech. Ten speeches are recorded for experiments in this paper. Experimental results show that the proposed dual core architecture implemented on an FPGA can save a lot of computation time without degrading the speech recognition rates. However, since the computation time is still too much to real-time applications, more cores are necessary to be integrated to increase the computation ability for an FPGA in the future.

ACKNOWLEDGEMENTS

This research work was supported by the Ministry of Science and Technology of the Republic of China under contract MOST 108-2221-E-390 -018

REFERENCES

[1] B. Milner and J. Darch, "Robust Acoustic Speech Feature Prediction from Noisy Mel-Frequency Cepstral Coefficients," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 2, pp. 338-347, 2011.
[2] L. Buera, A. Miguel, O. Saz, A. Ortega, and E. Lleida, "Unsupervised Data-Driven Feature Vector Normalization With Acoustic Model Adaptation for Robust Speech Recognition," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 296-309, 2010.
[3] J.W. Hung and W.H. Tu, "Incorporating Codebook and Utterance Information in Cepstral Statistics Normalization Techniques for Robust Speech Recognition in Additive Noise Environments," *IEEE Signal Processing Letters*, vol. 16, no. 6, pp. 473-476, 2009.
[4] L.D. Persia, D. Milone, H.L. Rufiner, and M. Yanagida, "Perceptual evaluation of blind source separation for robust speech recognition," *Signal Processing*, vol. 88, no.10, 2578-2583, 2008.
[5] J.T. Chien and M.S. Lin, "Frame-synchronous noise compensation for hands-free speech recognition in car environments," *IEE Proceedings - Vision, Image and Signal Processing*, vol. 147, no.6, pp. 508-515, 2000.
[6] Y. Zhan, H. Leung, K.C. Kwak, and H. Yoon, "Automated Speaker Recognition for Home Service Robots Using Genetic Algorithm and Dempster–Shafer Fusion Technique," *IEEE Trans. on Instrumentation and Measurement*, vol. 58, no. 9, pp. 3058-3068, 2009.
[7] C.T. Ishi, S. Matsuda, T. Kanda, T. Jitsuhiro, H. Ishiguro, S. Nakamura, and N. Hagita, "A Robust Speech Recognition System for Communication Robots in Noisy Environments," *IEEE Trans. on Robotics*, vol. 24, no. 30, pp. 759-763, 2008.
[8] C.W. Hsu and L.S. Lee, "Higher Order Cepstral Moment Normalization for Improved Robust Speech Recognition," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 17, no. 2, pp. 205-219 2009.
[9] A. Sankar and C.H. Lee, "A maximum-likelihood approach to stochastic matching for robust speech recognition," *IEEE Trans. Speech Audio Process*, vol.4, no. 3, pp.190-2021996.
[10] C. H. Lee, "On stochastic feature and model compensation approaches to robust speech recognition," *Speech Commun.*, vol. 25, pp. 29-47, 1998.
[11] M.J.F. Gales and S.J. Young, "Robust speech recognition in additive and convolutional noise using parallel model combination," *Comput. Speech Lang.*, vol. 9, pp. 289-307, 1995.
[12] Y. Tsao and C.H. Lee, "An Ensemble Speaker and Speaking Environment Modeling Approach to Robust Speech Recognition," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 17, no. 5, pp. 1025-1037, 2009.
[13] S. Windmann and R. Haeb-Umbach, "Parameter Estimation of a State-Space Model of Noise for Robust Speech Recognition," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 17, no. 8, pp. 1577-1590, 2009.
[14] N.E. Huang, "The empirical mod e decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proc. R. Soc. London*, pp. 903-995, 1998.
[15] S.T. Pan and X.Y. Li, "An FPGA-Based Embedded Robust Speech Recognition System Designed by Combining EMD and a Genetic Algorithm," *IEEE Transactions on Instrumentation & Measurement*, vol. 61, no. 9, pp. 2560-2572, 2012.

BIOGRAPHIES OF AUTHORS

Shing-Tai Pan was born in Pingtung, Taiwan, on November 4, 1966. He received the M.S. degree in electrical engineering from National Sun Yat-Sen University, Kaohsiung, Taiwan, in 1992 and the Ph. D. degree from National Chiao Tung University, Hsinchu, Taiwan, in 1996. In 2006, he joined the Department of Computer Science and Information Engineering, National University of Kaohsiung, Kaohsiung, Taiwan. He is now a Professor in the department and is the Dean of Office of Research and Development in NUK. He is a member of Taiwanese Association for Artificial Intelligence (TAAI) and Chinese Automatic Control Society (CACS). He is also a member of The Association for Computational Linguistics and Chinese Language Processing (ACLCLP). His current research interests are in the area of biomedical signal process, digital signal process, speech recognition, evolutionary computations, applications of artificial intelligent and intelligent control systems design