Fourth International Conference on Computational Intelligence, Robotics and Autonomous Systems
November 28-30, 2007, Palmerston North, New Zealand

Cerebellar-based Text-Dependent Speaker Verification

S. D. Teddy[1], E. M-K. Lai[2] and C. Quek[1]
[1]Centre for Computational Intelligence, School of Computer Engineering,
Nanyang Technological University, Nanyang Avenue, Singapore.
[2]Institute of Information Sciences and Technology, Massey University, Wellington, New Zealand.
sdt@pmail.ntu.edu.sg, e.lai@massey.ac.nz, ashcquek@ntu.edu.sg

Abstract

Speaker verification via the use of sampled speech belongs to a class of biometric recognition problems that offered a promising alternative approach to the traditional techniques for automatic person authentication. As the primary objective of a computerized speaker verification system is to be able to efficiently discern between an authentic speaker and an impostor, the accuracy of the speaker model employed to capture the speaker-specific characteristics extracted from the speech samples determines the performance level of the verification system. In this paper, a cerebellar-based approach to the text-dependent speaker verification problem is presented. The proposed technique employs the novel PSECMAC network to model the speaker-specific voice characteristics extracted via an MFCC analysis. Experiments performed on ten recruited speakers yielded an average frame-by-frame classification EER of 11.4%. The verification performances of the proposed system are encouraging.

Keywords: speaker verification, text dependent, PSECMAC, CMAC, cerebellum

1 Introduction

The ability to accurately differentiate a person from another is the key requirement in applications that involve the control or authorization of access to secured areas or materials. Some of these include automated banking, computer network security and retrieval of confidential information. Biometrics, or biometric recognition, refers to the kaleidoscope of technologies that employs biologically measurable and unique human physiological or behavioral characteristics for person recognition purposes. Examples of the biometric data used for person recognition include fingerprints, palm prints, iris and retinal scans, DNA, facial characteristics, and voice. Biometric-based recognition has many advantages over the traditional person recognition methods. Biometric characteristics cannot be forgotten or easily stolen and therefore a biometric-based person recognition system is robust against impostor attacks. The use of biometric-based authentication is also becoming socially acceptable since it is inexpensive and convenient to use.

Speaker recognition is a person recognition method using voice-induced biometrics information. Voice-based person recognition is highly economical and convenient as the users are only required to input a spoken phrase to the system in order to have their identity verified. The individual-specific voice characteristics are derived from the uniqueness in the behavioral and physiological aspects of the speaker's speech production system. Phonetics and linguistics research has established that the main differentiating physiological aspect of the human speech production system is the vocal tract [1]. Since no two vocal tracts are exactly the same, each individual's voice has certain acoustic peculiarities that characterize his/her vocal tract. The objective of a speaker recognition system is therefore to capture and exploit these differentiating features to discern between speakers. Speaker recognition can be classified into two problems [2]: (1) speaker identification and (2) speaker verification (authentication). Speaker identification is the problem of determining the identity of a speaker from a closed set of candidates. Speaker verification, on the other hand, refers to the problem of verifying the identity claim of the speaker. Both speaker identification and speaker verification systems can be further classified into the text-dependent and text-independent recognition methods.

Current research on speaker verification promote the use of statistical and probabilistic modeling of the speaker-specific characteristics [3, 4] as well as classical machine learning and pattern analysis-based speaker models [5, 6]. Although many technological advances and implementation successes in speaker recognition have been achieved recently, there are still major problems impeding the effective deployment of voice-based person recognition systems [7]. Most of these problems can be attributed to speaker variability and noise interferences. Speaker-induced variability (e.g. speaking rate, acoustic variability due to colds or disguise) and the variability in recording conditions as well as channel distortions affect the quality of the measured voice biometrics data. To resolve these problems, a set of speaker-differentiating features that is robust against speaker-variability as well as a recognition system that is able to efficiently cope with such variability and distortions are necessary.

This paper proposes the use of the newly developed cerebellar-based computational model named PSECMAC to perform text-dependent speaker verification. The research is motivated by everyday observations that aptly demonstrate how a human

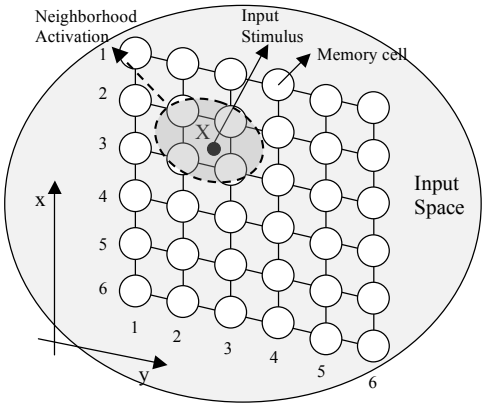
effortlessly employs his/her natural innate ability that facilitates the accurate perception of different auditory stimulants to perform speaker recognition proficiently. The human ability to differentiate sounds stems from the human auditory system's capacity to distinguish the different frequency components of acoustic signals [8]. Biological research has established that the human auditory nerves process sound *tonotopically* [9], where different sets of auditory nerve fibers respond selectively to the different frequencies of the acoustic signal. The auditory areas in the human brain are organized in a distinctive neural *sound map* [8] formation where the nerve fibers are topologically arranged with respect to their respective frequency stimulus. This pattern of organization is highly similar to the one observed in the human cerebellum, where different regions of the cerebellar cortex process the information from different sensory inputs [10]. This similarity subsequently motivates the use of the PSECMAC cerebellar model to perform automatic speaker verification.

The rest of this paper is organized as follows. Section 2 briefly describe the architecture of the PSECMAC network and highlights the cerebellar-inspired memory formation and knowledge acquisition process of the network. Section 3 presents the mechanisms of the proposed cerebellar-based speaker recognition system. The experimental results and analysis of the performances of the text-dependent PSECMAC-based speaker verification system are presented in Section 4. Section 5 concludes this paper.

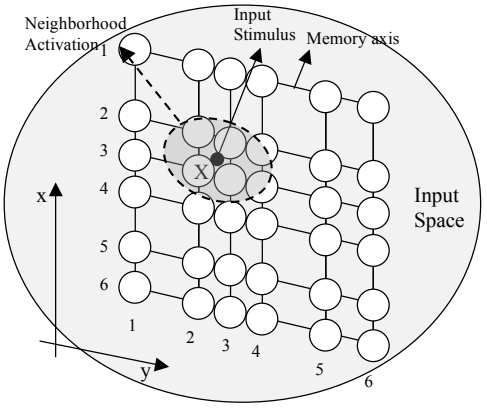
2 The PSECMAC Network

The cerebellum constitutes a part of the human brain that is important for motor control and cognitive functions [11], including motor learning and memory. The human cerebellum is postulated to function as a movement calibrator [10], which is involved in the detection of movement error and the subsequent coordination of the appropriate skeletal responses to reduce the error. It functions by performing *associative mappings* between the input sensory information and the cerebellar output required for the production of temporal-dependent precise behaviors [8]. The human cerebellum has been classically modelled by the Cerebellar Model Articulation Controller (CMAC) [12]. As a computational model of the human cerebellum, CMAC manifests as an associative memory network, where the memory cells are uniformly quantized to cover the entire input space. The CMAC network operation is characterized by the table lookup access of its memory cells. This allows for advantages such as localized generalization and rapid algorithmic computation.

This paper proposes the use of a brain-inspired cerebellar-based learning memory model named Pseudo Self-Evolving Cerebellar Model Arithmetic Computer (PSECMAC) as a generic functional model of the human cerebellum for solving approximation,



(a) CMAC



(b) PSECMAC

Figure 1: Comparison of CMAC and PSECMAC memory quantization for 2D input problem

modeling, control and classification problems. This architecture differs from the CMAC network in *two* aspects. Firstly, the PSECMAC network employs *one* layer of network cells, but maintained the computational principles of the layered-based CMAC network by adopting a neighborhood activation of its computing cells to facilitate: (1) smoothing of the computed output; (2) distributed learning paradigm; and (3) activation of highly correlated computing cells in the input space. Secondly, instead of uniform partitioning of the memory cells, the PSECMAC network employs the PSEC clustering technique [13] to form an experience-driven adaptive memory quantization mechanism of its network cells. Figure 1 illustrates this fundamental architectural distinction.

The adaptive quantization process of the PSECMAC network is performed in per dimension basis. The non-uniform quantization of the PSECMAC memory structure is inspired by the neurophysiological properties of the brain development, where the precise wiring in the adult brain is a result of

experience-dependent refinement of initial architecture through repeated exposures to external stimuli. This experience-dependent plasticity is also observed in the human cerebellum [14], and is incorporated to the PSECMAC network through the PSEC clustering algorithm. Each training data point is a learning episode to the network. In each input dimension, the PSEC clustering algorithm is used to compute clusters of data density, and the memory axes in each dimension are allocated based on the observed density profile of the training data. Thus, more memory cells are allocated to the densely populated regions of the input space. The details on the adaptive quantization algorithm is reported in [15].

The PSECMAC network employs a *Weighted Gaussian Neighborhood Output* (WGNO) computational process, where a set of neighborhood-bounded computing cells is activated to derive an output response to the input stimulus. For each input stimulus $\mathbf{X}$, the computed output is derived as follows:

Step 1: Determine the region of activation
Each input stimulus $\mathbf{X}$ activates a neighborhood of PSECMAC computing cells. The neighborhood size is governed by the neighborhood constant parameter N, and the activated neighborhood is centered at the input stimulus (see Fig 1(b)).

Step 2: Compute the Gaussian weighting factors
Each activated cell has a varied degree of activation that is inversely proportional to its distance from the input stimulus. These degrees of activation functioned as weighting factors to the memory contents of the active cells.

Step 3: Retrieve the PSECMAC output
The output is the weighted sum of the memory contents of the active cells.

Following this, the PSECMAC network adopts a modified *Widrow-Hoff learning rule* [16] to implement a *Weighted Gaussian Neighborhood Update* (WGNU) learning process. The network update process is briefly described as follows:

Step 1: Computation of the network output
The output of the network corresponding to the input stimulus $\mathbf{X}$ is computed based on the WGNO process.

Step 2: Computation of learning error
The learning error is defined as the difference between the expected output and the current output of the network.

Step 3: Update of active cells
The learning error is subsequently distributed to all of the activated cells based on their respective weighting factors.

3 The PSECMAC-based Speaker Verification System

Figure 2 depicts the block diagram of the proposed PSECMAC-based speaker verification system. The speaker verification system in Figure 2 consists of two main modules: a feature extraction block and the PSECMAC network. The feature extraction module computes the Mel-Frequency Cepstral Coefficients (MFCCs) to characterize the speech signals from the different speakers. During the training process, the PSECMAC network is used to learn and model the speaker-specific characteristics derived from the MFCCs values. In the testing phase, the PSECMAC-based speaker models are employed to perform the frame by frame verification of the incoming speaker voice. The computational mechanisms of the PSECMAC-based speaker verification system are described in the following sub-sections.

3.1 Preprocessing of Speech Signal

In this paper, the front-end preprocessing techniques that are applied to the incoming speech signal consist of: (1) speech signal segmentation; (2) windowing; and (3) voicing detection. The voice samples from each speaker are segmented into frames, where each voice frame consists of $M = 256$ data samples (this is approximately 30 ms long for a sampling frequency f_s of 8192 samples/sec). A 20 ms overlap between successive frames is employed to avoid information loss due to the use of an improper starting point in the segmentation process. Subsequently, the Hamming windowing technique is applied to attenuate the effect of signal discontinuities at the beginning and the end of each frame and to minimize the spectral distortion introduced by the segmentation process.

In the final step of the preprocessing stage, voicing detection is applied to the windowed speech segments. The purpose of the voicing detector is to identify and extract the voiced speech frames and to remove the unvoiced and silence frames. Unvoiced speech is produced when the air flow from the lungs is not modulated by the vocal cords and this results in a white-noise-like excitation signal to the vocal tract. The unvoiced speech frames therefore do not possess the quasi-stationary property that can be exploited for speaker characterization. In this paper, a Modified Zero Crossing Rate (MZCR) algorithm [17] is employed to effectively discern between the voiced, unvoiced and silence speech frames.

3.2 Feature Extraction

Mel-Frequency Cepstral Coefficients (MFCCs) [9] refer to the set of cepstral coefficients that are computed using the Mel-frequency scale that closely approximates the frequency responses of the human auditory system. The Mel-frequency scale is a perceptual scale of pitches derived from empirical studies on human listeners. It follows a linear

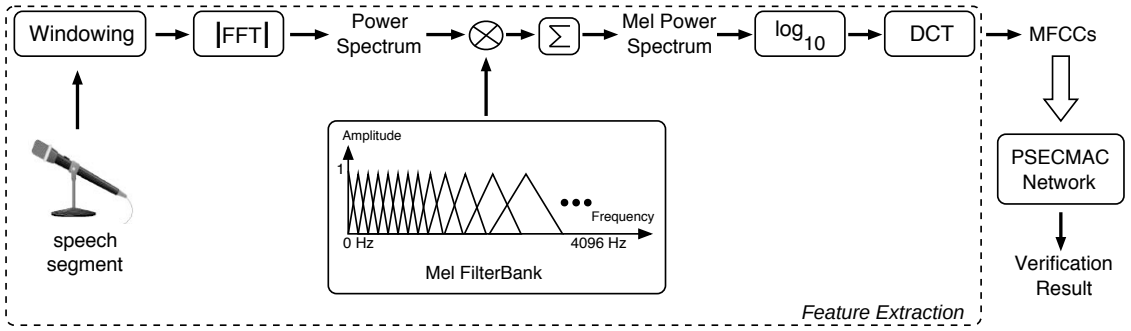


Figure 2: The proposed PSECMAC-based speaker verification system

frequency spacing for the frequency range below 1000 Hz and a logarithmic spacing for the frequency range above 1000 Hz (see Figure 2). For the purpose of the study, a total of 21 MFCCs are extracted per speech frame. However, the first MFCC component (i.e. MFCC_0) is excluded from the set of features employed for the speaker verification task. This is because MFCC_0 represents the mean value of the speech segment and thus contains little speaker-specific information [17].

3.3 PSECMAC Modeling of Speaker-Specific Characteristics

In this study, each speaker is characterized by the first six MFCCs (i.e. $\text{MFCC}_1 - \text{MFCC}_6$) out of the 20 valid MFCCs computed from his/her digitized voiced speech samples. This is motivated by the fact that the characteristics of the vocal tract that discern between the individual speakers are concentrated in the low quefrency domain of the voiced speech signal [9]. Thus, PSECMAC performs the associative mapping between the speaker-specific characteristics of the vocal tract to the identity of the respective speaker.

4 Experiments and Results

4.1 Dataset

The voice samples used in this study were collected from 10 randomly selected adult speakers consisting of six males and four females. The collected speech samples are first converted to wave files (*.wav) with a sampling frequency (f_s) of 8192 samples/second to obtain a speech quality that is compatible to that of typical day-to-day telephony applications. Based on the computed MFCCs of the voiced speech frames, speaker verifications are performed. Each speaker is characterized by the first six MFCCs of his/her voiced speech segment. The MFCCs of each speech segment form a data tuple. Therefore, there is a total of 100 such data tuples for each speaker in the experiments. The presentation order of the data is randomized and a three-fold cross validation (CV) approach was adopted throughout the evaluation process. Each CV group consists of a training and a testing set. In the experiments, the training set comprises of 40% of the entire MFCCs dataset of each speaker. The remaining 60% of the dataset constitutes the testing set such that the training and testing sets of each CV are mutually exclusive. On the other hand, there is a 25% overlap between the training sets of successive CV groups.

For each speaker, the final training and testing sets of a CV group consist of the training and testing samples of the corresponding CV groups of all the speakers. A single output is subsequently used to differentiate between the MFCCs data samples belonging to the actual speaker and those belonging to the impostors. The data samples that belong to the actual speaker are denoted with an output "1" while those that belong to the impostors are marked with an output "0". Since the number of impostor input samples far exceeds that of the actual speaker in the resultant training sets, the training of the PSECMAC-based speaker verification system using the CV groups described above is termed as the "unbalanced" training scenario.

4.2 Experimental Results and Analysis

The simulation is performed for all the three CV groups of the ten speakers; that is, a total of 30 experiments. The classification threshold (to discern between the actual speaker and the impostors) is varied to derive the receiver-operating-characteristics (ROC) curves for each evaluated CV group. The *Equal Error Rate* (EER) readings extracted from the ROC curves are subsequently employed as the performance measure of the speaker verification system. Type I error is defined as the error of falsely rejecting the voice input of the actual speaker whereas Type II error is the error of accepting the impostor's voice input as that of the actual speaker. EER denotes the point where Type I error equals Type II error.

A PSECMAC network with a memory size of 6 cells per dimension is constructed for the speaker verification task. As benchmarks, the set of experiments is repeated by using various well-established computational architectures. The benchmarking systems evaluated in this study are: (1) the basic CMAC network [12]; (2) the Multi-Layered Perceptron (MLP); (3) the Radial Basis Function

Table 1: The performances of the benchmarked speaker verification systems – Unbalanced training scenario

	Average Equal Error Rate [%]										**Average EER**	**EER**
Network	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**	**P9**	**P10**	**Total** [%]	**Std Dev**
CMAC	21.75	15.98	10.43	7.60	14.32	9.46	18.92	9.46	15.38	8.47	13.18	5.18
PSECMAC	**15.28**	**13.05**	**8.50**	**8.58**	**16.13**	**12.29**	**14.95**	**10.53**	**12.24**	**9.22**	**12.08**	**2.72**
MLP (6-13-1)	36.93	39.59	24.92	21.38	10.15	24.54	27.60	40.17	36.09	19.44	28.08	17.17
RBF	22.30	17.95	16.44	6.42	20.10	11.70	22.51	10.77	17.13	12.37	15.77	5.29
GenSoFNN-CRI	22.20	16.02	18.76	11.69	20.26	17.67	13.60	11.60	19.18	13.92	16.49	3.94

Table 2: The performances of the benchmarked speaker verification systems – Balanced training scenario

	Average Equal Error Rate [%]										**Average EER**	**EER**
Network	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**	**P9**	**P10**	**Total** [%]	**Std Dev**
CMAC	21.78	16.56	10.66	7.58	16.55	9.47	18.63	9.44	15.16	8.48	13.43	5.22
PSECMAC	**11.84**	**12.48**	**9.96**	**7.81**	**12.93**	**10.53**	**14.03**	**9.51**	**14.98**	**9.54**	**11.36**	**3.48**
MLP (6-13-1)	28.72	33.62	9.86	5.63	13.68	12.74	23.98	15.19	8.77	17.27	16.95	9.16
RBF	19.19	17.20	14.01	8.85	18.10	16.68	16.79	11.72	13.47	11.15	14.72	2.76
GenSoFNN-CRI	22.38	15.66	16.68	12.53	21.28	19.48	15.50	12.68	17.95	16.04	17.02	3.78

network (RBF); and (4) the Generic Self-Organizing Fuzzy Neural Network with the Compositional Rule of Inference scheme (GenSoFNN-CRI) [18]. The network size of CMAC has been defined as 6 cells per dimension to maintain a fair performance comparison with PSECMAC. The MLP has a predefined structure that consists of six input, thirteen hidden and one output nodes respectively while the RBF network is initialized to contain 100 hidden layer nodes. Meanwhile, the parameters of the GenSoFNN-CRI model has been empirically optimized.

Table 1 summarizes the average frame-by-frame EER values achieved by the various systems. From Table 1, one can observe that the PSECMAC network achieved the best verification performances among all the benchmarked systems. PSECMAC reports the lowest average EER of approximately 12% and a EER standard deviation of only 2.72% across all the evaluated speakers, thereby demonstrating the accuracy and consistency of its speaker models. The PSECMAC speaker model also outperformed that of the benchmarked CMAC network. The degraded speaker verification performances of CMAC are largely due to the uniform allocation of its memory cells. The rigid uniform partitioning of the input space limits the modeling accuracy of the CMAC network and thus leads to a suboptimal performance. From the results tabulated in Table 1, one can also observe that the MLP network reports the poorest speaker verification performances with an average EER of approximately 28%. The inferior performances of the MLP-based speaker verification system may be attributed to the use of the "unbalanced" training scenario, where the number of impostor samples in the training set far exceeds the number of authentic ones. Machine learning research has long established that the MLP is a connectionist network that employs global learning, whereby each presentation of a training sample adapts the weights of its entire network structure. Thus, training the MLP with the "unbalanced" training scenario will result in speaker models that are heavily biased towards the rejection of the impostors' speech samples. Consequently, the performances of the resultant MLP-based speaker verification system are severely impaired by the large Type I errors.

To investigate how the data composition of the training set affects the performances of the speaker verification systems, the set of simulations on the benchmarked architectures is subsequently repeated with a "balanced" training scenario. In the "balanced" training scenario, the training sets of the three CV groups of each speaker are modified by duplicating the positive training samples (i.e. samples belonging to the authentic class) until the number of authentic samples equals the number of impostor samples. The testing sets of the CV groups remain unchanged. Table 2 tabulates the speaker verification performances of the benchmarked architectures for the "balanced" training scenario. From Table 2, it is again evident that the PSECMAC-based speaker verification system comprehensively outperformed all the benchmarked systems based on the average EER values across all the evaluated speakers. The results tabulated in Tables 1 and 2 show that the "balanced" training scenario improves the average EER performances of the PSECMAC-based speaker verification system by approximately $5.9\%((12.08 - 11.36)/12.08)$. The overall improvement in the verification accuracy of the PSECMAC-based speaker verification system can be attributed to the balanced data distribution of the positive (actual) and negative (impostor) samples in the training set, which allows for better memory cell allocations in the resultant trained PSECMAC speaker models. This generally results in the higher modeling accuracies of the speaker models that subsequently translate to improved verification performances.

The simulation results in Tables 1 and 2 also showed that the average EER values of the MLP network improved by $39.6\%((28.08 - 16.95)/28.08)$ under the

"balanced" training scenario, thereby demonstrating the sensitivity of the MLP-based speaker models towards the structure of the training sets. Slight improvements were also noted for the RBF-based speaker verification systems and these can be attributed to the balanced data distribution of the two classes in the training sets. On the other hand, minor degradations in the performances of the CMAC and GenSoFNN-CRI-based speaker verification systems suggest that the duplicated positive samples introduced to the training sets have a slight detrimental effect on the performances of the corresponding speaker models. This may be due to the inherent learning and computational process of these two networks.

5 Conclusion

This paper presented a cerebellar-based approach to the text-dependent speaker verification problem. The proposed speaker verification system employs the novel PSECMAC network, which is a neurologically-inspired computational model of the human cerebellum, to model the speaker-specific characteristics of the human voice via the MFCC values extracted from the sampled voice segments. This study was motivated by the physiology of the human auditory system and the psychology of the human perception to acoustic sounds, which facilitate the human innate ability to accurately perform the speaker recognition process in everyday life.

The proposed PSECMAC-based speaker verification system was subsequently employed to verify the voice inputs of ten adult speakers. The verification performances of the PSECMAC speaker models were evaluated against those of the basic CMAC and GenSoFNN networks as well as the classical machine learning models of MLP and RBF networks. The experimental results had sufficiently demonstrated the superior accuracy of the PSECMAC-based speaker verification system to the benchmarked models.

6 References

[1] W. Tetschner, *Voice Processing*, Artech House, Boston (1993).

[2] J. P. Campbell Jr., "Speaker Recognition: A Tutorial", *Proceedings of the IEEE*, 85(9), pp 1437–1462 (1997).

[3] M. Lund, "A robust sequential test for text-independent speaker verification", *Journal of the Acoustical Society of America*, 99(1), pp 609–621 (1996).

[4] N. Z. Tishby, "On the application of mixture AR hidden Markov models to text-independent speaker recognition", *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-39, pp 563–569 (1991).

[5] J. Oglesby and J. S. Mason, "Radial basis function networks for speaker recognition", *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pp 393–396 (1991).

[6] J. Oglesby and J. S. Mason, "Optimization of neural models for speaker identification", *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pp 261–264 (1990).

[7] S. Furui, "50 years of progress in speech and speaker recognition", *Proceedings of the International Conference on Speech and Computer*, pp 1–9 (2005).

[8] E. R. Kandel, J. H. Schwartz and T. M. Jessell, *Principles of Neural Science, 4th Edition*, McGraw-Hill, Health Professions Division. (2000).

[9] B. Gold and N. Morgan, *Speech and Audio Signal Processing*, John Wiley & Sons, Inc. (2000).

[10] J. S. Albus, "Marr and Albus Theories of the Cerebellum: Two Early Models of Associative Memory", *Proceedings of IEEE COMPCON*, pp 577–582 (1989).

[11] F. A. Middleton and P. L. Strick, "The Cerebellum: An Overview", *Trends in Cognitive Sciences*, 27(9), pp 305–306 (1998).

[12] J. S. Albus, "A new approach to manipulator control: The cerebellar model articulation controller (CMAC)", *Journal of Dynamic Systems, Measurement, and Control. Transactions ASME*, pp 220–227 (1975).

[13] K. Ang and C. Quek, "Stock Trading using PSEC and RSPOP: A novel evolving rough set-based neuro-fuzzy approach", *Proc. IEEE Congress on Evolutionary Computation* (2005).

[14] K. D. Federmeier, J. A. Kleim and W. T. Greenough, "Learning-induces multiple synapse formation in rat cerebellar cortex", *Neuroscience Letters*, 332, pp 180–184 (2002).

[15] S. D. Teddy, C. Quek and E. M.-K. Lai, "PSECMAC: A Novel Self-Organizing Multi-Resolution Associative Memory Architecture", *IEEE Transactions on Neural Network, under 2nd review* (2007).

[16] B. Widrow and S. D. Stearns, *Adaptive Signal Processing*, New Jersey: Prentice-Hall inc. (1985).

[17] A. Wahab, G. S. Ng and R. Dickiyanto, "Speaker authentication system using soft computing approaches", *Neurocomputing*, 68, pp 13–37 (2005).

[18] W. L. Tung and C. Quek, "GenSoFNN: A Generic Self-Organizing Fuzzy Neural Network", *IEEE Transactions on Neural Networks*, 13(5), pp 1075–1086 (2002).