

The Construction of Wavelet Network for Speech Signal Processing

D. Shi^{1*} F. Chen¹ G. S. Ng¹ and J. Gao²

¹School of Computer Engineering, Nanyang Technological University, Singapore 639798

²School of Information Technology, Charles Sturt University, NSW 2795 Australia

(* Corresponding author, Email: asdmschi@ntu.edu.sg)

Abstract

Wavelet decomposition reconstructs a signal by a series of scaled and translated wavelets. Incorporating discrete wavelet decomposition theory with neural network techniques, wavelet networks have recently emerged as a powerful tool for many applications in the field of signal processing, such as data compression and function approximation. In this paper, four contributions are claimed: (1) From the point of view of machine learning, we analyze and construct wavelet network to achieve the compact representation of a signal. (2) A new algorithm of constructing wavelet network is proposed. The Orthogonal Least Square (OLS) is employed to prune the wavelet network. (3) Our experiments on speech signal processing results show that the wavelet network pruned by OLS achieves the best approximation and prediction capabilities among the representative speech processing techniques. (4) Our proposed methodology has been successfully applied to speech synthesis for a talking head to read web texts.

Keywords: Wavelet neural network, pruning, orthogonal least square, speech signal processing

1 INTRODUCTION

Speech signal processing is a well-studied field with a wide range of potential applications. In order to process a speech signal, it is typically represented by some features. One of these representations is signal decomposition by some basis functions. However, a signal can be decomposed by different sets of basis functions, but the performance at run time highly depends on the choice of these basis functions. By their observational nature, training data are usually finite and non-uniformly sampled, so the problem is consequently ill-posed. Conversion to a well-posed problem is typically achieved with some form of capacity control, which aims to balance the fitting of the data with constraints on the model flexibility, producing a robust model that generalizes successfully [1]. In practice, such an optimization is accomplished by searching for the minimum number of the basis functions under the well-recognized Occam's Razor hypothesis: "*plurality should not be posited without necessity*", or in other words, the simpler a solution is, the more reasonable it is. This research aims to the optimal compact representation to achieve the highest generalization capability and the lowest system complexity.

Recently, Gorriz, Puntonet, Salmeron and de la Rosa proposed a new model for time-series forecasting using radial basis functions and exogenous data [2]. They improved Platt's resource-allocation network (RAN) [3] by matrix decomposition and genetic algorithms. The RAN is a network that can find a compact representation with a reasonable amount of computation. The original RAN learns by allocating new units and adjusting the parameters of existing units, so it is particularly useful in on-line learning. Salmeron, Ortega, Puntonet and Prieto use orthogonal techniques such as QR factorization and singular value decomposition to improve such that an

optimum optimal set of prediction lags is determined and irrelevant radial basis functions (RBFs) are pruned [4].

In conventional speech signal processing, pitch synchronous overlap add (PSOLA) [5] and sinusoidal model [6] play important roles in controlling and adjusting the waveform of signals. In PSOLA, speech signal are represented by a sequence of separate and overlapping short-term signals. The idea of sinusoidal model is to decompose a signal into a set of sinusoids with time-varying amplitudes and frequencies. However, speech signals are non-stationary signal with many abrupt changes, window or Fourier transform based techniques are not always effective.

As an advanced alternative over classical Fourier analysis, wavelets have been successfully used in many aspects of signal processing, such as de-noising, compressing, edge detection, etc. [7-9]. The fundamental idea behind wavelets is to process data at different scales or resolutions. In such a way, wavelets provide a time-scale representation of a sequence of input signal. Additionally, since the wavelet is a local function, it is good at approximating any signals in finite domains. The main drawbacks of wavelet analysis are: (1) usually limited to problems of small dimensions; (2) the wavelets and their coefficients may not be optimal. As explained earlier, a signal can be decomposed into different sets of basis functions, but the set with the minimum number of basis functions is supposed to have the highest generalization ability.

To address the above-mentioned problems with wavelet transform, Zhang and Benveniste [11] proposed wavelet network in 1992. They constructed a special feed-forward neural network supported by the wavelet theory. Taking advantages of both the scaling property of wavelets and the effective learning mechanism of neural networks, wavelet networks are becoming a powerful

regression estimator for many applications. However, the optimization of wavelet network structure still needs to be investigated.

In this paper, we will discuss how to construct a neural network with the optimal set of wavelets. The rest of the paper is organized as follows. A brief introduction to wavelet network is given in Section 2. In Section 3, orthogonal least square is applied to find the optimal wavelets for a given regression application. Experiments and application are described in Section 4, followed by our conclusions and future work in Section 5.

2 WAVELET TRANSFORM VERSUS WAVELET NETWORK

Wavelet decomposition reconstructs a signal by a series of scaled and translated wavelets. A function $f \in L_2(\mathbf{R})$ can be represented as an infinite sum of weighted wavelet expansions:

$$f = \sum_{j,k \in \mathbf{Z}} \beta_{jk} \psi_{jk}, \quad (1)$$

where a typical $\psi_{j,k}(x) = \left\{ 2^{\frac{j}{2}} \psi(2^j x - k) : j, k \in \mathbf{Z} \right\}$, j is the scale index, and k is the translation

index. We can see that the elements of wavelet basis are translated and dilated versions of the mother wavelet. However, to satisfy the constraints of basis, the wavelet set, $\{\psi_{jk}\}$, is very rigid, and irrelevant to the input data. Thus, it can be expected that the wavelet reconstructor will be more efficient if the wavelet “basis” is constructed with respect to the input data.

Therefore, based on the theory of discrete wavelet transform, Zhang and Benveniste [11] tried to link the network coefficients with this appropriate transform. In 1992, they proposed wavelet

network for approximation, which had an explicit link between the network coefficients and the discrete inverse wavelet transform. In this way, they took advantage of the concept of discretization of inverse wavelet transform to initialize the coefficients. The wavelet network structure is of the following form:

$$f(\mathbf{x}) = \sum_{i=1}^N w_i \psi(\text{diag}(\mathbf{s}_i)(\mathbf{x} - \mathbf{t}_i)) + \mathbf{c}^T \mathbf{x} + \bar{g} \quad (2)$$

where N is the number of wavelets,

w_i is the weight of the i th wavelet,

$\psi(\cdot)$ is the mother wavelet,

\mathbf{t}_i is the translation factor of the i th wavelet,

\mathbf{s}_i is the dilation factor of the i th wavelet,

diag denotes diagonal matrix,

\mathbf{c} is a weight matrix from the input to the output,

\bar{g} is a bias to the output.

From Equation (2), one can see that the architecture of a wavelet network is exactly specified by the number of wavelets required for a given classification or regression application. The optimal wavelet network structure will achieve the best approximation and prediction capability. In this research, a library of wavelets will be selected to be the candidate hidden neurons (wavelets), and then the optimal architecture is constructed by pruning hidden neurons.

3 WAVELET NETWORK PRUNING

As mentioned above, the optimal wavelet network can be constructed in this way: First, select some scaled and translated wavelets (regressors), and then, some network pruning algorithms were employed to select the optimal regressors [12].

3.1 Wavelet Library Selection

A wavelet library consists of discretely scaled and translated versions of a given mother wavelet ψ . Thus, our job is to determine the scaling and translation values of the mother wavelet. First, we should specify the discretization step sizes for both factors, s_0 for scaling discretization, t_0 for translation discretization. Typically, a dyadic lattice is employed. After that, we must decide the maximum scale level, j , of the wavelet network. Then, for each scale level (from 1 to j), we will calculate all possible translation values regarding the input data, \mathbf{x} . In this way, we finally construct the wavelet library as follows:

$$\{\psi(\text{diag}(s_0^j)(\mathbf{x} - \mathbf{k}t_0)) : j \in \mathbf{Z}, \mathbf{k} \in \mathbf{Z}^d\}, \quad (3)$$

where d is the input dimension. These initial values of scaling and translation factors will be updated in the training procedure.

3.2 Network Pruning by Orthogonal Least Square Algorithm

The number of wavelets, N_s , may be decided manually or automatically. Therefore, in an N -candidate wavelet library, the goal is to find the optimal N_s wavelet regressors, which minimize the error between the real output and the expected output. This is a typical problem with RBF networks. There are two different ways to specify the number of hidden neurons. One way is to cluster the training examples and then assign a neuron to each cluster, the other way is to create a neuron for each training example and then to prune the hidden neurons [10].

Orthogonal least square (OLS) algorithm [13] is to find the regressors, which provide the most significant contribution to approximation error reduction. It has been widely accepted as an effective method for regressor selection in RBF network [14-16]. The advantage of employing OLS is that the responses of the hidden layer neurons are decorrelated so that the contribution of individual candidate neurons to the approximation error reduction can be evaluated independently.

Let $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_L)^T$ be the output matrix corresponding to all the number L of training examples. $\mathbf{y}_i (i = 1, 2, \dots, L)$ is an M -dimensional vector, denoting a total of M output units. We have $\mathbf{Y} = \mathbf{H}\mathbf{W} = (\mathbf{Q}\mathbf{A})\mathbf{W}$, where \mathbf{Y} , \mathbf{H} , \mathbf{W} are $L \times M$, $L \times N$, $N \times M$ matrices, respectively. The selection of wavelets is equivalent to the selection of the most significant columns of \mathbf{H} :

$$\mathbf{H} = \begin{bmatrix} h_{11} & h_{12} & \cdots & h_{1N} \\ h_{21} & h_{22} & & \vdots \\ \vdots & & & \vdots \\ \vdots & & & \vdots \\ h_{L1} & \cdots & \cdots & h_{LN} \end{bmatrix} \quad (4)$$

The matrix \mathbf{H} can be decomposed into \mathbf{QA} . The \mathbf{Q} is an $L \times N$ matrix with columns $[\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_N]$, and \mathbf{A} is an $N \times N$ upper triangular matrix as follows:

$$\mathbf{A} = \begin{bmatrix} 1 & a_{12} & \cdots & \cdots & a_{1N} \\ 0 & 1 & a_{23} & & \vdots \\ \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & & 0 & 1 & a_{(N-1)N} \\ 0 & \cdots & \cdots & 0 & 1 \end{bmatrix} \quad (5)$$

Only one column of \mathbf{H} is orthogonalized at a time. At the K th iteration, the K th column is made orthogonal to each of the $K-1$ previously orthogonalized columns and the operation for $K=2, \dots, N$ is repeated. The iteration can be stopped by introducing a criterion -- Akaike's final prediction error (FPE) [17]. It is actually the mean square error (MSE) of the function $f(x)$, multiplied by a factor of penalty in terms of the number of parameters. The definition of FPE is:

$$FPE(f) = \frac{1 + n_p/N_t}{1 - n_p/N_t} \frac{1}{2N_t} \sum_{n=1}^{N_t} (f(x_n) - y_n)^2 \quad (6)$$

where n_p is the number of regressors in the estimator, y_n is the expected output of the training data x_n , and N_t is the length of training data. If the FPE value begins to increase monotonically, and the change of MSE between two consecutive iterations is less than a threshold value, ε , the pruning procedure will be stopped. The stopping criteria are given by

$$\begin{cases} FPE^{(k)} - FPE^{(k-1)} \geq 0, & \text{and} \\ \Delta MSE < \varepsilon \end{cases} \quad (7)$$

A formal statement of the network pruning algorithm for the selection of wavelets is given as follows:

STEP 1 Initialization. Form the matrix \mathbf{H} in equation (4) by the wavelet function responses of all the training samples.

STEP 2 First wavelet selection. Set $I=\{1,2, \dots, N\}$, then for each $i \in I$, calculate the approximation error with the \mathbf{h}_i . The column that provides the maximum error is selected as the first column of $\mathbf{Q}^{(1)}$.

STEP 3 Orthogonalization and wavelet selection. Let $K=2$.

STEP 4 Orthogonalize all remaining columns of \mathbf{H} with all the columns of $\mathbf{Q}^{(K-1)}$ by OLS.

STEP 5 Calculate the FPE and MSE values at each K th iteration. Go to STEP 7 if the stopping criteria are satisfied.

STEP 6 $K:=K+1$, go to STEP 4.

STEP 7 End.

Figure 1 The pruning algorithm of wavelet network.

The wavelets corresponding to the first K columns in $\mathbf{Q}^{(K-1)}$ will be selected to build the wavelet network. The objective of the training procedure is to minimize the expectation of the MSE, thus a stochastic gradient algorithm is used to recursively adjust these parameters.

4 EXPERIMENTAL RESULTS

In this section, we will verify our OLS-based wave network with speech signal processing, and apply our proposed methodology to a talking head, which involves speech synthesis and speech-lip synchronization.

4.1 Performance Comparison

Two experiments have been done to show the performance of our proposed methodology. The first experiment is to show the effectiveness of the OLS-based pruning. The second experiment is to compare the approximation and prediction capabilities between wavelet network and other models on speech signal processing. The experiments were conducted on segment of speech signal. The sampling rate of the speech signal is 22.50 kHz, and the bits per sample are 16. The test segment has 1000 samples. The mother wavelet function is Mexican Hat wavelet.

There are three different network pruning algorithms used in [12], namely, Residual based selection, stepwise selection by orthogonalization and Backward elimination. The idea of residual based selection is to select the wavelet that best fits the training data, and then repeatedly select the wavelet that best fits the residual of the fitting of previous stage. Stepwise selection by orthogonalization first selects the wavelet that best fits the training data, and then repeatedly selects the wavelet that best fit the training data while working together with the previously selected wavelets. Backward elimination first builds the regression with all the wavelets, and then eliminates one wavelet in each step while trying to increase as less as possible the residual.

By considering MSE and FPE in the selection procedure, we can get a tradeoff between the minimum set of wavelets and the minimum error caused. Figure 2 shows the performance of the wavelet networks pruned by the above-mentioned three algorithms in [12] and our OLS-based algorithm. Since the construction of wavelet library is based on a dyadic lattice (i.e. the number of candidate wavelet is equal to $2^n - 1$, where n is the maximum scale level of wavelet network), to the 1000-sample speech signal, we selected the levels from level 4 to level 9. Thus, the number of candidate wavelets was from $(2^4 - 1)$ to $(2^9 - 1)$.

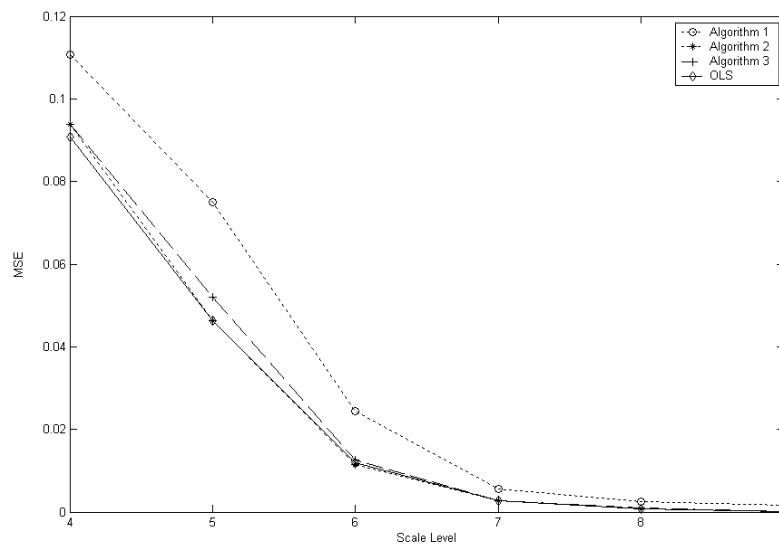


Figure 2 MSEs during the different pruning procedures.

It can be seen that the OLS performs better than other network pruning algorithms. Those methods, directly adding or eliminating regressors, do not take account of the correlation between regressors. In contrast, the OLS algorithm decouples the correlations among the responses of candidate wavelets. With correlation analysis, the individual contribution of each wavelet to the approximation error reduction can be calculated. That is why the OLS algorithm can achieve better results. Stepwise selection by orthogonalization is quite similar to OLS, however, our OLS-based

algorithm has a less computational complexity, thanks to its normalized diagonal and Gram-Schmidt method to compute columns.

We are now in a position to compare the approximation and prediction capabilities amongst our proposed wavelet network and some other representative methods on speech signal processing. The performance comparison among wavelet network, PSOLA and sinusoidal model is shown in Table 1. Here, the MSE is chosen to assess the approximation capabilities, whereas FPE is used to evaluate the prediction capabilities.

Table 1 Performance comparison between wavelet network and other existing speech models.

	MSE	FPE
Wavelet Network	0.00349	0.00257
PSOLA	0.00728	0.00369
Sinusoidal Model	0.01375	0.00817

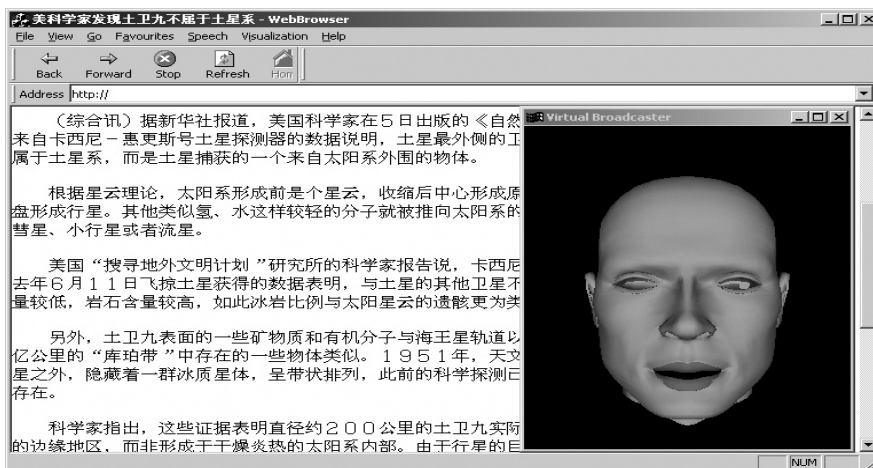
From Table 1, we can see that wavelet network outperforms both PSOLA and sinusoidal model. The advantages of wavelet network include signal analysis at different scales, as well as high learning ability. Furthermore, with its flexible structure, it can achieve different accurate results by simply adjusting the number of wavelets. Since sinusoidal model extracts tracks in frequency domain, its waveform approximation result in time domain is lower than both PSOLA and wavelet network. It must be noted that, only the optimal sparse representation is our concern in this research, and wavelet networks are time-consuming in training. So currently, wavelet networks are suitable for off-line signal processing, such as speech synthesis, time series data forecasting, etc.

4.2 Application to Speech Synthesis for Talking Head

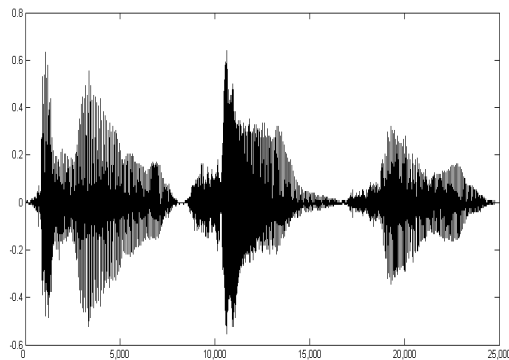
Our proposed methodology has been applied to a talking head system, which can read web texts [18]. In this research, a web browser system with talking head has been developed to read web texts in Madarine Chinese, as shown in Figure 3(a).

In comparison with English phonetics, Mandarin Chinese phonetics have some special features: (1) Every Chinese character is a monosyllable character; (2) A typical Chinese syllable consists of a consonant and a vowel; (3) Every syllable has a tone. Generally, there are four different tones in Mandarin pronunciation system; (4) One character may have more than one syllables, and vice versa. (5) The domain of the tone is the syllable, but tones of nearby syllables often affect each other. Figure 3(b) shows the synthesized speech signal of *Zao Shang Hao*, which means *Good morning*.

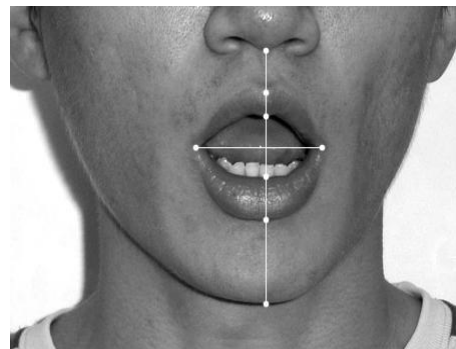
For each Chinese input character, its corresponding phoneme combination will be found in a Chinese character dictionary first. Then the phoneme combination is decomposed into an optional consonant part and a vowel part according to the table of character-phoneme conversion. Consequentially, each phoneme is converted to a viseme (or visemes) by the pre-defined mapping rules. Typically, a Chinese character will be converted into three corresponding visemes: a consonant viseme, a vowel viseme, and an end-viseme. After the conversion of text to viseme, corresponding speech signals will be analyzed to locate the frame position for each viseme. In lip modeling, it is not necessary to control all the lattice points on the 3D model. These lattice points on the lips can be adjusted according to seven specific feature points, or control points, as shown in Figure 3(c). Wavelet network is employed to control these control points to achieve audio-visual synchronization.



(a)



(b)



(c)

Figure 3Wavelet network applied to talking head. (a) A talking head for reading web texts in Chinese. (b) The synthesized speech signal of *Zao Shang Hao*, which means *Good morning*. (c)

The seven control points for lip movement modeling

5 CONCLUSIONS AND FUTURE WORK

In this paper, wavelet network has been analyzed from the point of view of machine learning, and successfully applied to speech signal processing. The construction of the optimal wavelet network is fulfilled by the selection of the optimal set of hidden neurons from a wavelet library. Our

experimental results have shown that the OLS algorithm can obtain wavelet networks with higher approximation and prediction abilities than the originally proposed pruning algorithms. Our experiments have also shown that the optimized wavelet network outperforms the two most popular speech processing techniques: PSOLA and sinusoidal model. Based on our proposed methodology, a talking head system has been successfully developed to read web texts.

As we know, support vector machines [19] enjoy higher generalization capability, thanks to their structural risk minimization. To further optimize the architecture of wavelet networks, our future work includes the construction of the optimal wavelet network based on support vector learning, borrowing some ideas from [20]. Our future work also includes the investigation of a faster training algorithm and/or specialized hardware to make wavelet network suitable for real time applications.

REFERENCES

- [1] J. B. Gao, C. J. Harris and S. R. Gunn, On a Class of Support Vector Kernels Based on Frames in Function Hilber Spaces, *Neural Computation*, **13**:1975-1994, 2001.
- [2] J. M. Gorriz, C. G. Puntonet, M. Salmeron and J. J. G. de la Rosa, A New Model for Time-Series Forecasting Using Radial Basis Functions and Exogenous Data, *Neural Computing and Applications*, **13**:101-111, 2004.
- [3] J. Platt, A Resource-Allocating Network for Function Interpolation, *Neural Computation*, **3(2)**:213-225, 1991.
- [4] M. Salmeron, J. Ortega, C. G. Puntonet and A. Prieto, Improved RAN Sequential Prediction Using Orthogonal Techniques, *Neurocomputing* **41**:153-172, 2001.
- [5] E. Moulines, F. Charpentier, Pitch Synchronous Waveform Processing Techniques For Text-To-Speech Synthesis Using Diphones. *Speech Communication*, **9**:453-467, 1990.
- [6] R. J. McAulay and T. F. Quatieri, Speech Analysis / Synthesis Based On A Sinusoidal Representation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **ASSP-34**:744-754, 1986.
- [7] S. G. Mallat, A Theory of Multiresolution Signal Decomposition: The wavelet representation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **11**: 674-693, 1989.
- [8] I. Daubechies, The Wavelet Transform, Time-Frequency Localization and Signal Analysis, *IEEE Transactions on Information Theory*, **36**: 961-1005, 1990.

- [9] S. G. Mallat and S. Zhong, S, Characterization of Signals from Multiscale Edges, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **14(7)**: 710-732, 1992.
- [10] C. M. Bishop, Improving the Generalization Properties of Radial Basis Function Neural Networks. *Neural Computation*, **3(4)**:579-588, 1991.
- [11] Q. Zhang and A. Benveniste, Wavelet Network. *IEEE Transactions on Neural Networks*, **3(6)**:889-898, 1992.
- [12] Q. Zhang, Using Wavelet Network in Nonparametric Estimation. *IEEE Transactions on Neural Networks*, **8(2)**:227-236, 1997.
- [13] S. Chen, C. F. Cowan and P. M. Grant, Orthogonal Least Squares Learning Algorithms for Radial Basis Function Networks, *IEEE Transactions on Neural Networks*, **2(2)**:302-309, 1991.
- [14] S. Chen, E. S. Chng and K. Alkadhim, Regularized Orthogonal Least Squares Algorithm for Constructing Radial Basis Function Networks, *International Journal of Control*, **64(5)**:829-837, 1996.
- [15] S. Chen, Y. Wu, and B. L. Luk, Combined Genetic Algorithm Optimisation and Regularised Orthogonal Least Squares Learning for Radial Basis Function Networks, *IEEE Transactions on Neural Networks*, **10(5)**:1239-1243, 1999.
- [16] J. B. Gomm and D. L. Yu, Selecting Radial Basis Function Network Centers with Recursive Orthogonal Least Squares Training, *IEEE Transactions on Neural Networks*, **11**:306-314, 2000.
- [17] H. Akaike, Fitting Autoregressive Models for Prediction. *Annals of the Institute of Statistical Mathematics*, **21**:243-347, 1969.
- [18] F. Chen, V. Spinko and D. Shi, Real-Time Lip Synchronization Using Wavelet Network, In: Proceedings of International Conference on Cyberworlds, Singapore, 2005.
- [19] V. N. Vapnik, An Overview of Statistical Learning Theory, *IEEE Transactions on Neural Networks*, **10**:988-999 1999.
- [20] B. Scholkopf, K. K. Sung, C. J. C. Burges, F. Girosi, P. Niyogi, T. Poggio and V. Vapnik, Comparing Support Vector Machines with Gaussian Kernels to Radial Basis Function Classifiers. *IEEE Transactions on Signal Processing*, **45(11)**: 2758-2765, 1997.