ICIC Express Letters
Part B: Applications
Volume **7**, Number **10**, October **2016**

ICIC International ©2016 ISSN 2185-2766
pp. **2087–2094**

SPEAKER RECOGNITION WITH MEL SCALE-BASED WAVELET PACKET DECOMPOSITION AND AR-VOLTERRA MODEL

Shuying Yang[1,2], Jun Guo[1,2] and Jiaojiao Jiang[1,2]

[1]School of Computer and Communication Engineering
Tianjin University of Technology
[2]Tianjin Key Laboratory of Intelligence Computing and Novel Software Technology
No. 391, Binshui Xidao, Xiqing District, Tianjin 300384, P. R. China
15222308101@163.com

Received April 2016; accepted July 2016

Abstract. *In order to improve the rate of speaker recognition, this paper proposes a new scheme with Mel scale-based Wavelet Packet Decomposition and AR-Volterra model. First, the speech signal is decomposed by wavelet packet of Mel scale, and calculate energy spectrum of the sub-band signals. The sub-bands of low energy are discarded according to weighting the proportion of energy for all-bands. Second, chaos of sub-band signals are determined by the largest Lyapunov exponent. If sub-band is chaotic, Volterra adaptive model will be used to extract features. Otherwise, the autoregressive model will be used to extract features. Third, linear and nonlinear characteristic parameters will be used to speaker recognition. Hidden Markov model is used to recognize speaker. The experimental results show the extracted features have been obviously improved.*
Keywords: Mel scale, The largest Lyapunov exponent, AR-Volterra model, Speaker recognition

1. **Introduction.** Due to the growing need for secured access and criminalistics investigation, improving speaker recognition systems becomes an attractive challenge. Currently, in order to improve the rate of speaker recognition, the researchers have made a lot of research in the feature extraction. The Linear Prediction Cepstrum Coefficient (LPCC) is widely used in feature extraction [1]. However, the voice signal is chaotic, and the process of voice generation belongs to nonlinear system, so voice feature extracted by linear model is not accurate. Kokkinos proposed nonlinear model and extracted Lyapunov exponent as features. The results show that the features improved the recognition rate of speaker [2]. However, this approach did not have a significant improvement than the linear feature extraction method. The Mel-Frequency Ceptral Coefficients (MFCC) is also widely used in feature extraction [3]. MFCC uses the nonlinear characteristics of the human ear hearing frequency. However, the triangular filter of MFCC is used to divide frequency domain, and the signal attenuation is weak. Daqrouq proposed the Wavelet Transform Linear Prediction Coefficient (WTLPC) method based on combination of Discrete Wavelet Transform (DWT) and LPCC [4], and experimental results show that the LPCC based on wavelet transform or wavelet packet transform is suitable for feature extraction. However, it is still a linear model, and the voice signal is chaotic. Conventional wavelet packet transform mechanisms do not wrap based on the basis of the frequency of human's auditory system.

In this paper, speaker recognition feature extraction scheme with Mel scale-based Wavelet Packet Decomposition (WPD) and AR-Volterra model is proposed [5]. This method conforms to the features of the auditory, and the signal attenuation is smooth. The sub-band signals are close to human beings' perceptive ability. Through the analysis of the chaotic characteristic of sub-band signals, the linear and nonlinear methods

are respectively used for feature extraction. It is more accurate than using the linear or nonlinear model only.

The paper is organized as follows. Section 2 presents the Mel scale-based WPD technique in speech signal. In Section 3, we detail the feature extraction scheme with Mel scale-based WPD and AR-Volterra model. Section 4 presents the experimental setup and results for speaker recognition. In Section 5, we have a summary of this paper.

2. Mel Scale-Based Wavelet Packet Decomposition.

Conventional wavelet packet transform mechanisms do not wrap based on the frequency of the human's auditory system. The speech's MFCC parameters mainly show the signal's low-frequency feature, which is in line with the human's hearing mechanism. However, some high-frequency signals still have lots of valuable feature information. So, we use Mel scale-based WPD to decompose speech signal, which ensures that we can extract speech signal feature with higher integrity on the basis of the human's hearing mechanism.

In this paper, the device with a sampling frequency of 8k HZ is used to obtain the original speech signal. To reduce the signal to noise ratio of the speech signal, we use the following formula for signal pre-emphasis.

$$H(z) = 1 - 0.97z^{-1} \tag{1}$$

Then, the speech signal is converted from time domain to the frequency by fast Fourier transform (FFT). In the process of solving the MFCC, the Mel filter is replaced by the haar wavelet packet function and the frequency domain speech signal is decomposed based on the bandwidth of Mel-scale.

Then the sub-band coefficient signals are reconstructed and the signal changed from the frequency domain to the time domain by the inverse Fourier transform. Calculate the energy of each sub-band signal, and select high-energy signal as a valid signal to prepare for extracting characteristic parameters. The calculation formula for the sub-band energy of each frame signal is as follows:

$$E_j = \frac{\sum_{j=1}^{N_j} \left[W_j^p x(n)\right]^2}{N_j}, \quad j = 1, 2, \ldots, 24 \tag{2}$$

In Formula (2), N_j is the total number of coefficients presented in that particular band, $W_j^p x(n)$ is the p-th wavelet cofficients of the j-th sub-band signal, and $x(n)$ is sub-band signal.

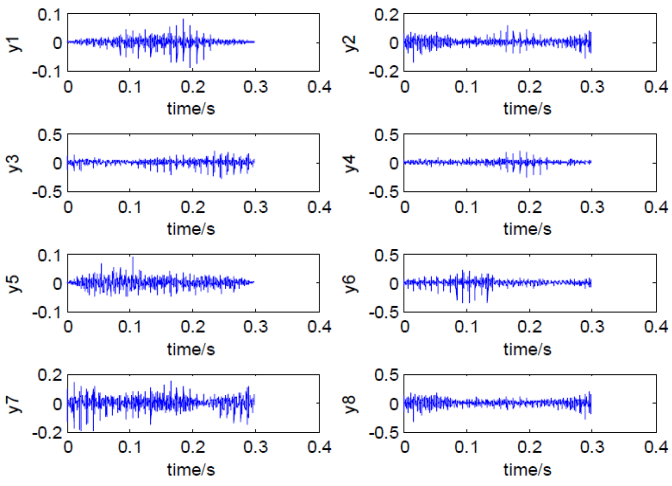

FIGURE 1. Three-layer Haar wavelet packet decomposition based on Mel scale

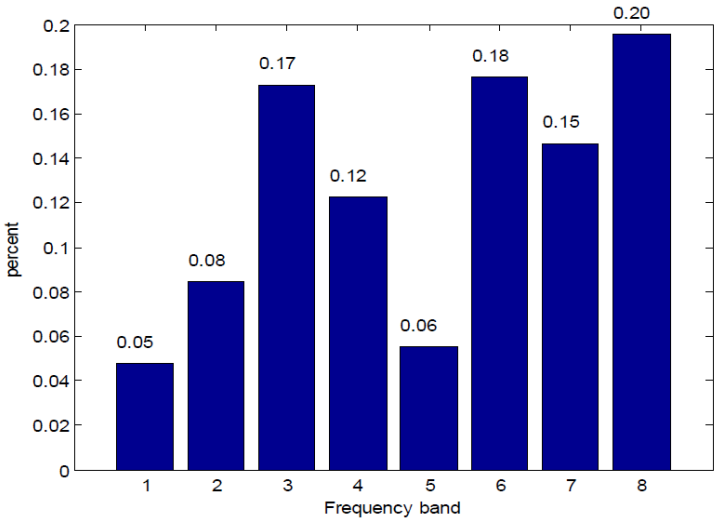

FIGURE 2. The energy percentage of the decomposed sub-band signals

Voice signal "blue sky" is decomposed based on Mel scale wavelet packet. Here we decompose the speech signal into 8 sub-band signals on the basis of Mel scale, and the result is shown in Figure 1; calculate energy spectrum of the sub-band signals and the result is shown in Figure 2.

3. **Extracted Features of AR-Volterra Mixture Model.** The proposed speech feature extraction scheme basically consists of three parts. At first, the input signal will be processed by the endpoint detection technology or some other pretreatment technology and the active components of speech signal are obtained. The signal is transformed into frequency domain by fast Fourier transform (FFT). The signal is divided by the Mel scale-based WPD. Sub-band signals of extremely low energy are discarded by weighting the proportion of energy for all-bands. Second, chaos of sub-bands signals is determined by the largest Lyapunov exponent. Third, if sub-band is chaotic, Volterra adaptive model will be used to extract features, and weight vectors of Volterra filter are obtained for speaker recognition. Otherwise, the autoregressive (AR) model will be used to extract features, and the cepstral coefficients and the coefficients of AR models and acceleration coefficients are obtained for the speaker recognition.

3.1. **Chaos validation of speech signal.** By calculating characteristic parameters of strange attractor factor, chaos can be determined in the speech signal. We can use Lyapunov exponent to determine and describe whether the nonlinear time series belongs to the chaotic system. For the Lyapunov exponent, if $\lambda > 0$, the system becomes unstable and will lose sensitiveness to initial values. Therefore, $\lambda > 0$ can be the decision of whether a speech time series is chaotic.

In the phase space of the speech time series, $x(t_0)+\delta x(t_0)$ is assumed close to the initial point $x(t_0)$, t_0 is the initial time. From the recurrence of n times, the following equation can be derived:

$$x\left(t_n+1\right)+\delta x\left(t_n+1\right)=f\left[x\left(t_n\right)+\delta x\left(t_n\right)\right]\approx f\left[x\left(t_n\right)\right]+\delta x\left(t_n\right) f'\left[x\left(t_n\right)\right] \tag{3}$$

Equation (4) can be obtained from above:

$$\delta x\left(t_n+1\right)=\delta x\left(t_n\right) f'\left[x\left(t_n\right)\right] \tag{4}$$

In Formula (4), t_n is the current time. After n times' recurrence, the distance between the two points on the phase track is as follows:

$$|\delta x(t_n)| = \left|\delta x(t_0)\right| \left|\prod_{i=0}^{n-1} f'[x(t_i)]\right| = |\delta x(t_0)| e^{\lambda t_n} \tag{5}$$

In Formula (5), $\delta x(t_0)$ is the distance between the two points on the phase track in the initial time, λ is the Lyapunov exponent of the system, and the expression is as follows:

$$\lambda = \lim_{t_0 \to \infty} \frac{1}{t_n} \sum_{i=0}^{n-1} \ln |f'[x(t_i)]| \tag{6}$$

Here, Wolf method is used to identify chaos of eight sub-band signals. The result is shown in Table 1.

TABLE 1. Sub-band signals chaos of validation

Sub-band signal	The largest Lyapunov exponents
y1	0.5639
y2	0.2450
y3	0.2060
y4	0.5313
y5	−0.1608
y6	0.3196
y7	−0.1276
y8	−0.1329

3.2. **Extracted feature of AR model.** The linear prediction model is defined as the form $\sum_{j=1}^{p} a_j x(n-j)$, which approximates the input signal $x(n)$ for $n = 0, \ldots, N-1$ in a least square sense, where p denotes the model order, and a_j is the coefficients for $j = 1, \ldots, p$. The calculation method of LPCC is on the basis of full point model for linear prediction coefficient (LPC) parameters recursive, forming the LPC cepstrum. Its recursive form is as follows:

$$\begin{cases} c_1 = a_1 \\ c_n = a_n + \sum_{k=1}^{n-1} \frac{k}{n} c_k a_{n-k}, \ 1 < n \le p \\ c_n = \sum_{k=1}^{n-1} \frac{k}{n} c_k a_{n-k}, \ n > p \end{cases} \tag{7}$$

where $a_1, a_2, \ldots, a_p$ are the feature vectors for p-order linear prediction coefficient (LPC), and $c_n, n = 1, \ldots, p$ are the first p values cepstrum. When the LPCC order number is smaller than LPC order number, the second type will be used for calculation. On the contrary, the third type will be used.

In this paper, the order number of the LPC is 12, and the AR model coefficients of power spectrum will be obtained by the auto-correlation sequence. The order number of the LPCC is 12 too, and 13 dimensional cepstral coefficients are derived by the standard cepstral recursion. Then the 39 dimensional feature parameters combine with cepstral coefficients and the coefficients of AR models and acceleration coefficients are obtained for the speaker recognition.

Here y5, y7, and y8 are respectively used to the experiments by AR model and Volterra model. Experiments results show that using the model of linear is better for sub-band signals which are not chaotic. The result of y7 is shown in Figure 3 and Figure 4. (In order to show clearly, these figures only show 72 predictive values parts of 2000).The average error of Volterra adaptive prediction is 0.0955, and AR model prediction is 0.0168.

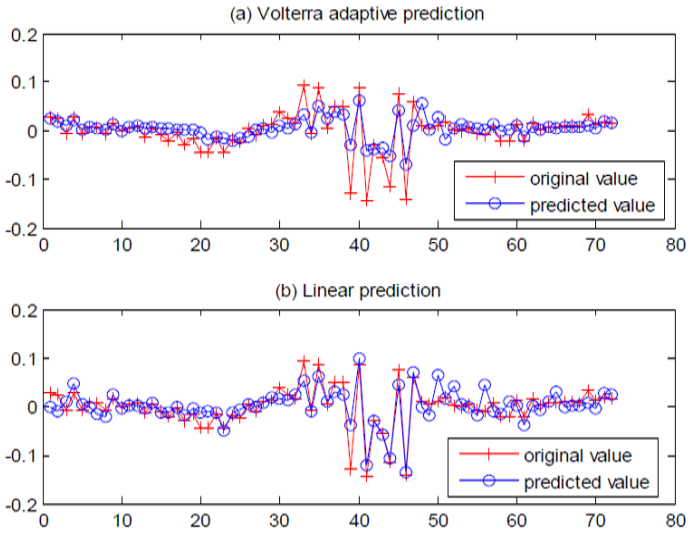

FIGURE 3. Volterra adaptive prediction and linear prediction of speech signal

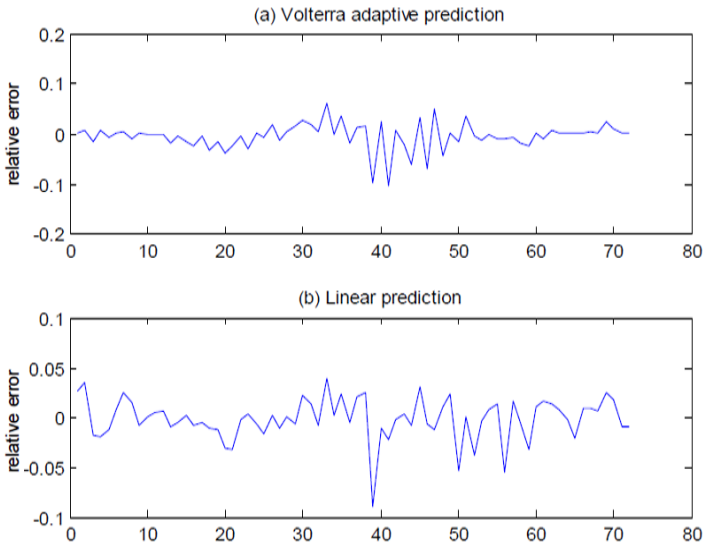

FIGURE 4. Volterra adaptive prediction and linear prediction error of speech signal

3.3. **Extracted feature of Volterra adaptive model.** To establish a mathematical model that expresses the process of the voice time sequence generation, Volterra functional model will be introduced in modeling of nonlinear time series [6-9]. Volterra series filter of chaotic time series is as follows:

$$\begin{aligned} & x(n+1) \\ &= F\left(X(n)\right) \\ &= h_0 + \sum_{m=0}^{+\infty} h_1(m)x(n-m) + \sum_{m_1=0}^{+\infty}\sum_{m_2=0}^{+\infty} h_2(m_1,m_2)x(n-m_1)x(n-m_2) + \cdots \\ &\quad + \sum_{m_1=0}^{+\infty}\sum_{m_2=0}^{+\infty}\cdots\sum_{m_p=0}^{+\infty} h_p(m_1,m_2,\ldots,m_p)x(n-m_1)x(n-m_2)\cdots x(n-m_p) + \cdots \end{aligned} \tag{8}$$

where $h_1, h_2, \ldots, h_n$ is the kernel function in Volterra series, it is implicit function of the system, and reflects the macroscopic of the speech signal. According to the characteristics

of voice time series, to reduce the amount of computation, second order Volterra adaptive prediction model usually is selected to truncated forms of expression as follows:

$$\hat{x}(n+1) = h_0 + \sum_{i_1=0}^{N-1} h_1(i_1)x(n-i_1\tau) + \sum_{i_1,i_2=0}^{N-1} h_2(i_1,i_2)x(n-i_1\tau)x(n-i_2\tau) + e(n) \quad (9)$$

Enter the amount of a linear adaptive FIR filter defined as $U(n)$, and coefficient vector is defined as $H(n)$ in the following expression:

$$U(n) = \left[1, x(n), x(n-1), \cdots, x(n-m+1), x^2(n), x(n)x(n-1), \cdots, x^2(n-m+1)\right]^T \quad (10)$$

$$H(n) = \left[h_0, h(0), h(1), \cdots, h(m-1), h_2(0,0), h_2(0,1), \cdots, h_2(n-m+1)\right]^T \quad (11)$$

Since the Volterra adaptive filter coefficients can be directly determined by linear adaptive FIR filter algorithm, the Formula (9) can be expressed as:

$$\hat{x}\,(n+1) = H^T\,(n)\,U\,(n) \quad (12)$$

Of Formula (12), second-order adaptive filter using orthogonal time adaptive algorithm, the algorithm input vector $U(n)$, and the coefficient vector $H(n)$ can be described as:

$$\begin{aligned} \hat{x}\,(n) &= H^T\,(n-1)\,U\,(n-1) \\ H\,(n) &= H\,(n-1) + c \times \frac{e\,(n-1)}{U^T\,(n)\,U\,(n)} U\,(n-1) \\ e\,(n) &= x\,(n) - \hat{x}\,(n) \end{aligned} \quad (13)$$

In Formula (13), c is the parameters of the control convergence. Calculate the prediction error $e(n)$ by Equation (13), and calculate the mean square error. When the mean square error value reaches the set threshold (in this paper, threshold is 0.08), extract Volterra filter weight vector or predict the error vector value as a feature in Speaker Recognition System.

Here y1, y2, y3, y4, and y6 signals are respectively used to the experiments by AR model and Volterra model. Experiments results show that the sub-band signals are chaotic, and Volterra model is better than the AR model. The result of y1 is shown in Figure 5 and Figure 6. (In order to show clearly, these figures only show 70 predictive values parts of 2000). The average error of Volterra adaptive prediction is 0.067, and AR model prediction is 0.103.

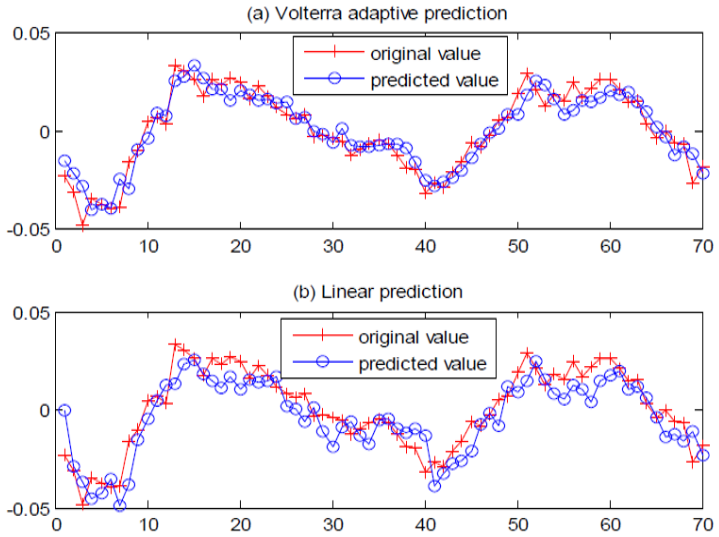

FIGURE 5. Volterra adaptive prediction and linear prediction of speech signal

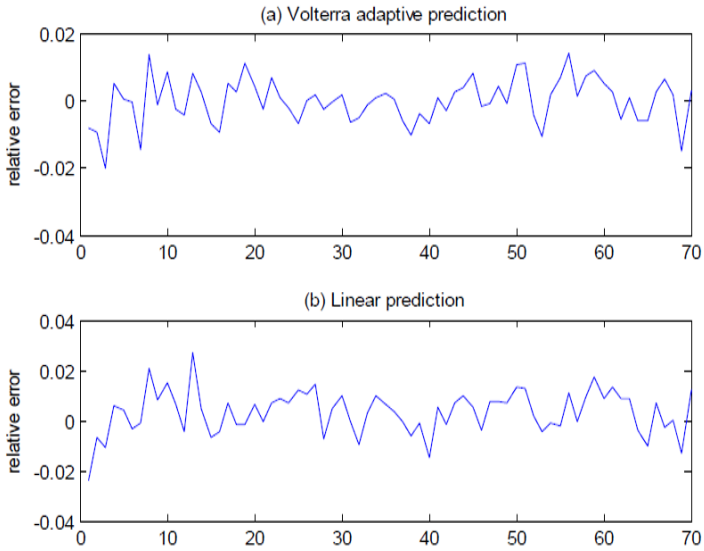

FIGURE 6. Volterra adaptive prediction and linear prediction error of speech signal

TABLE 2. Speaker recognition rates (%) based on the HMM

Extract feature	MFCC	Conventional WPD	WPD based on Mel scale
LPCC features (%)	80.0	84.0	88.0
Volterra features (%)	82.0	88.0	90.0
AR+Volterra features (%)	86.0	92.0	98.0

4. **Experimental Setup and Results.** In this paper, Timit speech database will be used to experiment. This experiment included 50 speakers. MFCC, conventional wavelet packet and Mel-based WPD are respectively used at the same time to test the recognition effect of LPCC, Volterra, and AR-Volterra. Feature parameters are obtained for each speech sample by the AR-Volterra mixture model. The output speech features are trained based on the Hidden Markov Model (HMM). The speech data in test files will be used to evaluate the matching degree by the trained HMM. The values of matching degree obtained by several times test beyond a certain value will be identified as the training and test samples deriving from the same people. The new proposed features are used to evaluate the speaker recognition system based on HMM. The performance metric of speaker recognition system is the matching degree which is a probability obtained by Viterbi algorithm of HMM.

The result of speaker recognition experiment is shown in Table 2. Speaker recognition rates show that the speech signal is decomposed by wavelet packet of Mel scale which is superior to conventional wavelet packet decomposition and MFCC. Speaker recognition accuracy of Mel scale-based WPD is generally higher than conventional wavelet packet and MFCC. Whether it is allocated bands by the conventional wavelet packet or Mel scale-based WPD speaker recognition accuracy rate of AR-Volterra model is the highest, and the system of stability is the best.

5. **Conclusions.** In this paper, we have proposed Mel scale-based wavelet packet decomposition and AR-Volterra model for speaker recognition. The speech signal will be decomposed by wavelet packet based on Mel scale. This method conforms to the features of the auditory, and the signal attenuation is smooth. Using the AR-Volterra mixture models to extract feature has higher accuracy. The results of the experiment show that the feature extraction scheme has a significant improvement. Also, this method is more

accurate and it reduces the distortion degree of features. In this paper, the experiment was conducted under ideal conditions and the influence of the noise was not considered. In order to validate the universal applicability of the method, the speech signal of noise environment will be experimented in the future work.

Acknowledgment. This work is partially supported by Key Laborator of Computer Vision and System, Ministry of Education, Tianjin, China. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] U. Bhattacharjee, A comparative study of LPCC and MFCC features for the recognition of Assamese phonemes, *International Journal of Engineering Research and Technology*, ESRSA Publications, vol.2, no.1, 2013.

[2] I. Kokkinos and P. Maragos, Nonlinear speech analysis using models for chaotic systems, *IEEE Trans. Speech and Audio Processing*, vol.13, no.6, pp.1098-1109, 2005.

[3] K. S. Ahmad, A. S. Thosar, J. H. Nirmal et al., A unique approach in text independent speaker recognition using MFCC feature sets and probabilistic neural network, *The 8th International Conference on Advances in Pattern Recognition (ICAPR)*, pp.1-6, 2015.

[4] K. Daqrouq and K. Y. Al Azzawi, Average framing linear prediction coding with wavelet transform for text-independent speaker identification system, *Computers & Electrical Engineering*, vol.38, no.6, pp.1467-1479, 2012.

[5] S. Srivastava, S. Bhardwaj, A. Bhandari et al., Wavelet packet based Mel frequency cepstral features for text independent speaker identification, *Intelligent Informatics*, Springer Berlin Heidelberg, pp.237-247, 2013.

[6] B. Shoaib, I. M. Qureshi et al., Adaptive step-size modified fractional least mean square algorithm for chaotic time series prediction, *Chinese Physics B*, pp.129-137, 2014.

[7] N. Chaitra, D. M. Mohan and D. N. Dutt, Nonlinear dynamical analysis of speech signals, *Proc. of International Conference on VLSI, Communication, Advanced Devices, Signals & Systems and Networking*, Springer India, pp.343-351, 2013.

[8] W. Ji and W. S. Gan, Identification of a parametric loudspeaker system using an adaptive Volterra filter, *Applied Acoustics*, vol.73, no.12, pp.1251-1262, 2012.

[9] G. Favier, A. Y. Kibangou and T. Bouilloc, Nonlinear system modeling and identification using Volterra PARAFAC models, *International Journal of Adaptive Control and Signal Processing*, vol.26, no.1, pp.30-53, 2012.