A Blind Bandwidth Extension Method of Audio Signals based on Volterra Series

Xingtao Zhang, Changchun Bao, Xin Liu, Liyan Zhang
Speech and Audio Signal Processing Laboratory, School of Electronic Information and Control Engineering,
Beijing University of Technology, Beijing 100124, China
E-mail: taozi@emails.bjut.edu.cn, baochch@bjut.edu.cn, liuxin0930@emails.bjut.edu.cn, zhangliyan1999@emails.bjut.edu.cn

***Abstract*—In this paper, a blind bandwidth extension method of audio signals is proposed in which the fine structure of high-frequency information is recovered based on Volterra series. Combining with Gaussian mixture model and codebook mapping to adjust the spectrum envelope and energy gain of the extended high-frequency components separately, the bandwidth of audio signals is extended to super-wideband from wideband. Furthermore, the proposed method is applied into a real audio codec. The performance of the proposed method is evaluated through objective and subjective tests on the audio signals selected from MPEG items, and it is found that the proposed method outperforms the chaotic prediction method and nearest-neighbor matching method. When the proposed algorithm is applied into ITU-T G.722.1 wideband audio codec, the performance is comparable with that of G.722.1C super-wideband audio codec at 24 kbps.**

I. INTRODUCTION

Due to the limitation of transmission bandwidth and storage capacity, usually the low-frequency (LF) components which is more perceptually important to human ear is transmitted in audio codec, and the less relevant high-frequency (HF) components is truncated during the transmission of audio signals. This inevitably leads to the degradation of audio quality. Obviously, it is necessary to reconstruct the HF components to realize the transmission of high quality components at low bit-rate. For this reason, audio bandwidth extension (BWE) emerges as the times require. Generally, BWE can be divided into non-blind BWE and blind BWE [1-3]. In the blind BWE method, no side information related to the HF components is transmitted at the encoder, and it can be compatible with any type of audio codecs. Therefore, it becomes the key part of audio codec. Similar to speech BWE based on source-filter model, it mainly includes the recovery of spectrum envelope and fine structure of audio signals. But the recovery method of fine structure will not be suitable for audio signals because their characteristics are different.

It has been proved that the nonlinear prediction method keeps a high prediction precision. It can predict the unknown parts by analyzing the known parts for a one-dimension time series. Recently, the nonlinear prediction is gradually introduced into audio signal processing, and experiment results show that audio BWE based on nonlinear prediction outperforms conventional algorithms [2][3]. The chaotic prediction method [2] reconstructs the fine structure of HF components based on joint prediction, and adjusts the spectrum envelope with linear extrapolation (LE) method. Furthermore, the harmonic components of the reconstructed HF information are adjusted. In that method, LE ignores the statistic characteristics between the LF and HF information, thus there will be a large deviation between the estimated value and the actual value. The nearest-neighbor matching method [3] reconstructs the HF components by searching the nearest neighbor in LF phase points, and adjusts the spectrum envelope for the extended HF components with Gaussian mixture model (GMM) which will cause the estimated envelope too smooth. For above reasons, in this paper the nonlinear prediction based on Volterra series is adopted to recover the fine structure of HF components. Then, GMM and codebook mapping [4] are used to adjust the spectrum envelope and energy gain for the extended HF components, respectively. Thus a blind BWE method of audio signals from wideband (WB) to super-wideband (SWB) is realized. The evaluation results show that the proposed method outperforms the chaotic prediction method [2] and the nearest-neighbor matching method [3].

The paper is organized as follows: The principles of BWE based on Volterra series as well as the application in ITU-T G.722.1 audio codec are described in Section □. The quality test results are presented in Section Ⅲ and the conclusions are given in Section Ⅳ.

II. BLIND BANDWIDTH EXTENSION OF AUDIO SIGNALS BASED ON VOLTERRA SERIRS

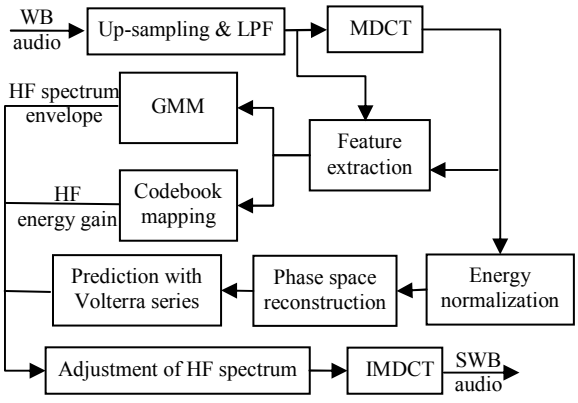

Fig. 1 Block diagram of the proposed method

The block diagram of the proposed BWE is shown in Fig.1.

Firstly, WB audio signal sampled at 16 kHz is up-sampled by the factor of 2 and low-pass filtered. The filtered signal is re-sampled at 32 kHz with bandwidth of 7 kHz. This signal is transformed into frequency domain through Modified Discrete Cosine Transform (MDCT). Secondly, the phase space of LF-MDCT coefficients normalized by the root mean square value of each sub-band is reconstructed, and the HF-MDCT coefficients are obtained based on Volterra series. Thirdly, the wideband features are extracted and the GMM and codebook mapping are adopted to adjust the spectrum envelope and energy gain for the extended HF components, respectively. Finally, the LF and HF components are combined in frequency domain and transformed into time domain through inverse MDCT, and the blind BWE of audio signals is realized. The main contents will be described in the following parts.

A. Phase Space Reconstruction

Phase space reconstruction aims to describe the evolving rules of a one-dimension spectrum series through the nonlinear relationship between the phase points in the reconstructed phase space. Thus, the nonlinear prediction of a one-dimension spectrum series can be made. The main parameters used for phase space reconstruction include time delay τ and embedding dimension m. The time delay τ is to make the adjacent components of each phase point as independent as possible, and embedding dimension m is to make the evolving trajectories of phase points in phase space as unfolded as possible.

Let $x(k)$, k=1,2,…,N as the one-dimension spectrum series and N is the number of LF-MDCT coefficients. Through phase space reconstruction, the N LF-MDCT coefficients are mapped into N-(m-1)τ phase points, and all the phase points compose a phase space $\mathbf{X}^{(m)}$ which is represented as:

$$\mathbf{X}^{(m)} = \begin{bmatrix} \mathbf{X}(1) \\ \mathbf{X}(2) \\ \vdots \\ \mathbf{X}(N-(m-1)\tau) \end{bmatrix} = \begin{pmatrix} x(1) & x(1+\tau) & \cdots & x(1+(m-1)\tau) \\ x(2) & x(2+\tau) & \cdots & x(2+(m-1)\tau) \\ \vdots & \vdots & & \vdots \\ x(N-(m-1)\tau) & x(N-(m-2)\tau) & \cdots & x(N) \end{pmatrix} \tag{1}$$

where $\mathbf{X}(k)$, k=1,2,…,N-(m-1)τ is a m-dimension vector and it represents a phase point in phase space.

In this paper, the de-biasing auto-correlation function (DB-ACF) method [5] which has a low computational complexity is selected to compute time delay τ. The DB-ACF $R_{xx}(\tau)$ of LF-MDCT coefficients is given by:

$$R_{xx}(\tau) = \frac{1}{N-\tau}\sum_{n=1}^{N-\tau}(x(n+\tau)-\bar{x})(x(n)-\bar{x}) \tag{2}$$

where $\bar{x}$ denotes the mean of LF-MDCT coefficients. When $R_{xx}(\tau)$ is reduced to (1–1/e) of $R_{xx}(0)$, the corresponding τ is selected as the proper time delay of phase space.

For the calculation of embedding dimension m, the Cao method [6] is adopted. The ratio of distances between the arbitrary phase point and its nearest neighbor is donated as $a(n,m)$ when the dimension number is changed to m+1 from m, i.e., $a(n,m)$ is defined by:

$$a(n,m) = \frac{\left\|\mathbf{X}_{m+1}(n)-\mathbf{X}_{m+1}(n_{NN'})\right\|_\infty}{\left\|\mathbf{X}_m(n)-\mathbf{X}_m(n_{NN})\right\|_\infty} \quad n=1,2,...,N-m\tau \tag{3}$$

where $\mathbf{X}_m(n_{NN})$ is the nearest neighbor of $\mathrm{X}_m(n)$ in the m-dimension phase space. The mean of $a(n,m)$ is denoted as $E(m)$, and if the variation of $E(m+1)/E(m)$ stops at m, then m is chosen as the proper embedding dimension of phase space.

With the proper time delay τ and embedding dimension m, the phase space of audio spectrum series can be reconstructed. Therefore, the nonlinear prediction of audio spectrum series can be made.

B. Nonlinear Prediction based on Volterra Series

According to Takens theory, the prediction reconstruction of nonlinear series $x(n)$ is an essentially inverse process of the nonlinear dynamics, i.e., the kinetic model $F(\cdot)$ is reconstructed through the state $\mathbf{X}(n)$ of the dynamic system. The relationship can be represented as:

$$x(n+t) = F(\mathbf{X}(n)) \tag{4}$$

where t (t>0) is the prediction step length.

There are many nonlinear functions to approach the kinetic model $F(\cdot)$. Theoretical research and practical experience show that most nonlinear dynamic systems can be represented by the Volterra series [7]. Therefore, we use Volterra series to build nonlinear kinetic model $F(\cdot)$ of MDCT coefficients, that is,

$$\begin{aligned} \hat{x}(n+1) &= F(\mathbf{X}(n)) \\ &= h_0 + \sum_{m_1=0}^{m-1} h_1(m_1)x(n-m_1\tau) \\ &+ \sum_{m_1=0}^{m-1} h_2(0,m_1)x^2(n-m_1\tau) + \sum_{m_1=1}^{m-1} h_2(1,m_1)x(n)x(n-m_1\tau) \\ &+ \sum_{m_1=0}^{m-1} h_3(0,m_1)x^3(n-m_1\tau) + \sum_{m_1=1}^{m-1} h_3(1,m_1)x^2(n)x(n-m_1\tau) \\ &+ \sum_{m_1=1}^{m-1} h_3(2,m_1)x(n)x^2(n-m_1\tau) \end{aligned} \tag{5}$$

where $\mathbf{X}(n)$ is a m-dimension phase point in the reconstructed phase space of LF-MDCT coefficients, $\hat{x}$ (n+1) is the estimated value, m_1 is the index corresponding to the components of phase point, h_p, p=0,1,2,3 represents the model coefficient.

Let $\mathbf{U}$ and H as the input vector and coefficient vector, respectively, which are given by,

$$\begin{aligned} \mathbf{U} = [&1, x(n), \cdots, x(n-(m-1)\tau), x^2(n), \cdots, x^2(n-(m-1)\tau), \\ &x(n)x(n-\tau), \cdots, x(n)x(n-(m-1)\tau), x^3(n), \cdots, \\ &x^3(n-(m-1)\tau), x^2(n)x(n-\tau), \cdots, x^2(n)x(n-(m-1)\tau), \\ &x(n)x^2(n-\tau), \cdots, x(n)x^2(n-(m-1)\tau)]^{\mathrm{T}} \end{aligned} \tag{6}$$

$$\begin{aligned} \mathbf{H} = [&h_0, h_1(0), \cdots, h_1(m-1), h_2(0,0), \cdots, h_2(0,m-1), \\ &h_2(1,1), \cdots, h_2(1,m-1), h_3(0,0), \cdots h_3(0,m-1), \\ &h_3(1,1) \cdots, h_3(1,m-1), h_3(2,1), \cdots, h_3(2,m-1)]^{\mathrm{T}} \end{aligned} \tag{7}$$

Then the equation (5) can be represented as:

$$\hat{x}(n+1) = \mathbf{H}^{\mathrm{T}}\mathbf{U} \tag{8}$$

In the propose method, a group of phase points close to the HF coefficients are chosen as central phase points in order to avoid accumulated error. The specific steps are as follows:

1) Reconstruct the m-dimension phase points of LF-MDCT coefficients through phase space reconstruction;
2) Search the nearest neighbor $\mathbf{X}(n)$ of the last phase point $\mathbf{X}(M)$ at a proper interval, and set $\mathbf{X}(n)$ as the prediction central phase point;
3) Choose the neighbors $\mathbf{X}(n_i)$ of phase point $\mathbf{X}(n)$ based on degree of incidence [8], and calculate the weight w_i, i=1,2,…,L of neighbor $\mathbf{X}(n_i)$;
4) Calculate the model coefficients based on Weighted Recursive Least square [9] by minimizing the following squared error ε:

$$\varepsilon = \sum_{i=1}^{L} \lambda^{L-i} \left\{ w_i \left(x(n_i+1) - \mathbf{H}_i^{\mathrm{T}} \mathbf{U}_i \right) \right\}^2 \tag{9}$$

where n_i is the highest dimension components of neighbor $\mathbf{X}(n_i)$, $x(n_i+1)$ is the actual value of LF-MDCT coefficients, γ=0.9 is a forgetting factor, w_i is a weight used for adjusting each neighbor's effort on the prediction, $\mathbf{U}_i$ and $\mathbf{H}_i$ represent the i^{th} input vector and coefficient vector, respectively, which are given by,

$$\begin{aligned} \mathbf{U}_i = [&1, x(n_i), \cdots, x(n_i-(m-1)\tau), x^2(n_i), \cdots, x^2(n_i-(m-1)\tau), \\ &x(n_i)x(n_i-\tau), \cdots, x(n_i)x(n_i-(m-1)\tau), x^3(n_i), \cdots, \\ &x^3(n_i-(m-1)\tau), x^2(n_i)x(n_i-\tau), \cdots, x^2(n_i)x(n_i-(m-1)\tau), \\ &x(n_i)x^2(n_i-\tau), \cdots, x(n_i)x^2(n_i-(m-1)\tau)]^{\mathrm{T}} \end{aligned} \tag{10}$$

$$\begin{aligned} \mathbf{H}_i = [&h_0^{(i)}, h_1^{(i)}(0), \cdots, h_1^{(i)}(m-1), h_2^{(i)}(0,0), \cdots, h_2^{(i)}(0,m-1), \\ &h_2^{(i)}(1,1), \cdots, h_2^{(i)}(1,m-1), h_3^{(i)}(0,0), \cdots h_3^{(i)}(0,m-1), \\ &h_3^{(i)}(1,1), \cdots, h_3^{(i)}(1,m-1), h_3^{(i)}(2,1), \cdots, h_3^{(i)}(2,m-1)]^{\mathrm{T}} \end{aligned} \tag{11}$$

The recursive equations are given by:

$$\begin{aligned} &\mathbf{H}_i = \mathbf{H}_{i-1} + \mathbf{Q}_i e(i \mid i-1) \\ &e(i \mid i-1) = w_i (x(n_i+1) - \mathbf{H}_{i-1}^{\mathrm{T}} \mathbf{U}_i) \\ &\mathbf{Q}_i = \frac{\mathbf{T}_{i-1}\mathbf{U}_i}{\lambda + \mathbf{U}_i^{\mathrm{T}} \mathbf{T}_{i-1} \mathbf{U}_i} \\ &\mathbf{T}_i = \frac{1}{\lambda}\left[\mathbf{T}_{i-1} - \frac{\mathbf{T}_{i-1}\mathbf{U}_i\mathbf{U}_i^{\mathrm{T}}\mathbf{T}_{i-1}}{\lambda + \mathbf{U}_i^{\mathrm{T}}\mathbf{T}_{i-1}\mathbf{U}_i} \right] \end{aligned} \tag{12}$$

In the proposed method, the coefficient vector is initialized as zero, and the neighbors of the prediction center point are used to adjust the coefficient vector. The last value is selected as the coefficient vector $\mathbf{H}$.

5) Obtain the HF coefficient according to equation (8);
6) Let n=n+1, and continue the nonlinear prediction point by point until all the HF coefficients are obtained.

C. HF Spectrum Envelope Estimation based on GMM

Auditory experiment shows that the HF spectrum envelope plays an important role on the quality of extended audio signals. In this paper, a HF spectrum envelope estimation method based on GMM is proposed.

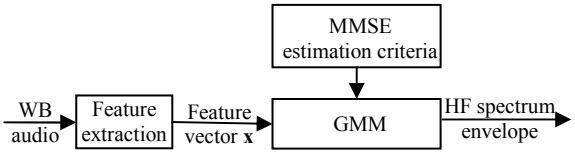

Fig. 2 Block diagram of HF spectrum envelope estimation

The block diagram is depicted in Fig. 2. Firstly, the proper WB feature vector is selected and denoted as $\mathbf{x}$. Combining SWB HF spectrum envelope vector $\mathbf{y}$, the joint vector $\mathbf{z}$ can be made as $\mathbf{z}$=$\{\mathbf{x}, \mathbf{y}\}$. Here, the WB features consist of zero-crossing rate F_{zcr}, gradient index F_{gi}, sub-band envelope F_{rms}, sub-band flux F_f and MPEG-7 timber features including audio spectrum centroid F_{asc}, audio spectrum spread F_{ass} and spectrum flatness F_{sf}. Then the GMM model is adopted to calculate the joint probability, and the Gaussian mixture parameters can be obtained through Expectation-Maximum (EM) algorithm. Finally, the HF spectrum envelope is estimated via the posteriori probability derived from minimum mean-square error (MMSE) criterion.

D. HF Energy Gain Estimation based on Codebook Mapping

Psychoacoustic experiment results show that the HF energy is very sensitive, so it is an important issue to adjust HF energy of SWB audio signals in BWE algorithm.

The same WB feature vector is adopted, and the mapping codebook is the corresponding HF energy gain factor g which is defined by:

$$g = \log_{10}\left(\sum_{k=N_h}^{N_h} x^2(k) / \sum_{k=1}^{N_l-1} x^2(k) \right) \tag{13}$$

Where $x(k)$ is the spectrum coefficients of audio signals, N_l and N_h are the starting frequency and cut-off frequency of HF information, respectively.

After the feature vector in each audio frame is computed, the HF energy gain factor can be obtained based on fuzzy mapping method [4].

E. Application of Bandwidth Extension in Audio Codec

Due to the limitation of transmission bandwidth, the audio quality obviously will degrade if the HF components are discarded. By using the BWE method into audio enhancement module at the decoder, the audio quality can be efficiently improved.

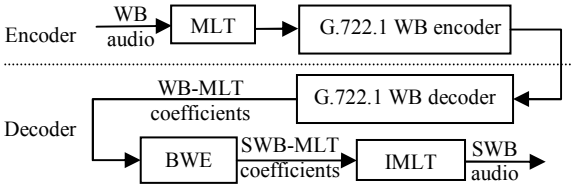

Fig. 3 Block diagram of BWE application in G.722.1 codec

The block diagram of BWE application in G.722.1 codec is shown in Fig.3. At the encoder, the input WB audio signal is firstly transformed into frequency domain through Modulated Lapped Transform (MLT), and then the WB-MLT coefficients are encoded at proper bit-rate. At the decoder, the SWB-MLT coefficients can be obtained with BWE module according to the decoded WB-MLT coefficients, and then SWB audio signals can be reconstructed by an inverse MLT.

III. Evaluation and Test Results

In terms of audio quality measurement, the proposed BWE is compared with two kinds of existing nonlinear BWE methods, which are the chaotic prediction method [2] and the nearest-neighbor matching method [3]. In this paper, the training audio signals are mixed audio signals with the size of 30 minutes, including violin, drum, symphony and so on. The number of Gaussian component in GMM is set to 64 and the size of codebook is 256. The testing audio signals are from

MPEG items and up-sampled to 48 kHz. The objective evaluation used in this paper is PEAQ test according to ITU-R BS.1387-1 [10]. The main parameter in PEAQ test is Objective Difference Grade (ODG) which varies from -4 (very annoying) to 0 (imperceptible difference). The change of ODG in the value of 0.1 is usually perceptually audible.

Table I shows the comparisons of ODG scores of the proposed method and the chaotic prediction method [2] and the nearest-neighbor matching method [3]. The test results indicate that the proposed BWE outperforms the chaotic prediction method [2] and the nearest-neighbor matching method [3].

TABLE I
COMPARISONS OF ODG SCORES
OF THE PROPOSED METHOD AND THE CHAOTIC PREDICTION METHOD
AND THE NEAREST-NEIGHBOR MATCHING METHOD

Audio signals	Proposed method	Chaotic prediction	Nearest-neighbor matching
brahms	-2.660	-2.799	-2.697
music2	-2.502	-2.741	-2.622
music5	-3.338	-3.768	-3.487
phi3	-2.906	-3.074	-3.017
phi5	-3.448	-3.604	-3.630
trilogy	-2.972	-3.509	-3.331
Average	-2.971	-3.249	-3.131

In addition, the extended SWB audio signals in G.722.1 and directly reconstructed SWB audio signals in G.722.1C are compared at the bit-rate of 24 kbps. Table Ⅱ shows the comparisons of ODG scores.

TABLE Ⅱ
COMPARISONS OF ODG SCORES
OF EXTENDED SWB AUDIO SIGNALS IN G.722.1 AND DIRECTLY
RECONSTRUCTED SWB AUDIO SIGNALS IN G.722.1C

Audio signals	Extended SWB audio signals in G.722.1	Directly reconstructed SWB audio signals in G.722.1C
brahms	-3.646	-3.613
music2	-3.382	-3.325
music5	-3.718	-3.753
phi3	-3.554	-3.712
phi5	-3.698	-3.753
trilogy	-3.588	-3.635
Average	-3.598	-3.632

Furthermore, 12 listeners are invited to determine which is more preferable or choose no difference between the G.722.1C audio codec and extended G.722.1 audio codec. Table Ⅲ shows the subjective preference test results of the perceived quality of extended SWB audio signals in G.722.1 and directly reconstructed SWB audio signals in G.722.1C.

TABLE Ⅲ
SUBJECTIVE PREFERENCE TEST RESULTS
OF EXTENDED SWB AUDIO SIGNALS IN G.722.1 AND DIRECTLY
RECONSTRUCTED SWB AUDIO SIGNALS IN G.722.1C

	Extended SWB audio signals in G.722.1	Directly reconstructed SWB audio signals in G.722.1C	No difference
Rate	36.7%	30%	33.3.0%

The objective and subjective test results indicate that the perceived quality of extended SWB audio signals in G.722.1 audio codec is comparable with directly reconstructed SWB audio signals in G.722.1C audio codec at the bit-rate of 24 kbps.

IV. CONCLUSIONS

This paper presents a nonlinear bandwidth extension method of audio signals. The nonlinear prediction method based on Volterra series which is used to build the nonlinear kinetic model of spectrum series is adopted to recover the fine structure of high-frequency components. The timber of extended super-wideband audio signals is maintained with the GMM-based high-frequency spectrum envelope estimation method and high-frequency energy gain estimation method based on codebook mapping. Both the objective and subjective test results indicate that the proposed method outperforms the chaotic prediction method and nearest-neighbor matching method, and the perceived quality of extended super-wideband audio signals in G.722.1 wideband audio codec is comparable with directly reconstructed super-wideband audio signals in G.722.1C super-wideband audio codec at 24 kbps.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (Grant No. 61072089, Grant No. 60872027), the Beijing Municipal Natural Science Foundation (Grant No. 4082006), Scientific Research Key Program of Beijing Municipal Commission of Education (Grant No. KZ201110005005) and the Funding Project for Academic Human Resources Development in Institutions of Higher Learning Under the Jurisdiction of Beijing Municipality.

REFERENCES

[1] Xin Liu, Chang-chun Bao, Mao-shen Jia and Yong-tao Sha, "A Harmonic Bandwidth Extension based on Gaussian Mixture Model," *10th International Conference on Signal Processing (ICSP2010)*, Beijing, CHINA, Oct. 2010, pp. 474-477.

[2] Yong-tao Sha, Chang-chun Bao, Mao-shen Jia and Xin Liu, "High Frequency Reconstruction of Audio Signal based on Chaotic Prediction Theory," *IEEE International Conference on Acoustics, Speech and Signal Processing. (ICASSP)*, Dallas, Texas, USA, Mar. 14-19, 2010, pp. 381-384.

[3] Xin Liu, Chang-chun Bao, Mao-shen Jia and Yong-tao Sha, "Nonlinear Bandwidth Extension based on Nearest-Neighbor Matching," *Processing of the Second APSIPA ASC*, Biopolis, Singapore, Dec. 14-17, 2010, pp. 169-172.

[4] Yong Zhang and Rui-min Hu, "Speech Wideband Extension based on Gaussian Mixture Model," *Chinese Journal of Acoustics*, 2009, Vol. 28, pp. 362-377.

[5] Holger Kantz and Thomas Schreiber, *Nonlinear Time Series Analysis*. Britain: Cambridge University press, 2004.

[6] Xiao-ke Xu, Xiao-ming Liu and Xiao-nan Chen, "The Cao Method for Determing the Minimum Embedding Dimension of Sea Clutter," *2006 CIE International Conference on Radar*, Shanghai, China, 2006, Vol. 1, pp. 77-80.

[7] Ji-rong Gu, Xian-wei Chen and Jie-ming Zhang, "An Algorithm of Prediction for Chaotic Time Series based on Volterra Filter," *2009 Second International Symposium on Electronic Commerce and Security*, 2009, Vol. 2, pp. 205-208.

[8] Zhen-yu Teng and cheng-sheng Pan, "FH Frequency Multi-step Prediction Method Based on Degree of Incidence and Simulation," *Journal of System Simulation*, Nov. 2009, Vol. 21, pp. 7077-9, 7083. (in Chinese)

[9] V. J. Mathews, "Adaptive polynomial filters," *IEEE Signal Processing Magazine*, 1991, Vol. 8, pp: 10-26.

[10] ITU-R Rec. BS.1387-1, *Method for Objective Measurements of Perceived Audio Quality*, 1998-2001.