# SPECTRAL ENVELOPE ESTIMATION USED FOR AUDIO BANDWIDTH EXTENSION BASED ON RBF NEURAL NETWORK

*Hao-jie Liu, Chang-chun Bao, Xin Liu*

Speech and Audio Signal Processing Laboratory, School of Electronic Information and Control Engineering, Beijing University of Technology, Beijing 100124, China
E-mail: liuhaojie@emails.bjut.edu.cn, baochch@bjut.edu.cn, liuxin0930@emails.bjut.edu.cn

## ABSTRACT

Abstract—In this paper a new spectral envelope estimation method based on radial basis function (RBF) neural network is proposed for implementing a blind bandwidth extension method of audio signals. To make the sub-band envelope of high-frequency (HF) components accurately recovered, the RBF neural network is utilized to fit the relationship between low-frequency (LF) features and sub-band envelope of HF components. In addition, the fine structure of HF components which can guarantee the timber of the extended audio signal is reconstructed based on nonlinear dynamics. The objective and subjective test results indicate that the proposed method outperforms the reference methods.

***Index Terms***—Audio signal processing, bandwidth extension, envelope estimation, RBF neural network

## 1. INTRODUCTION

Sound is an essential medium of communication between people and environment. In the audible frequency range of human ear, audio signals can generally be divided into four classes: narrowband (NB) audio, wideband (WB) audio, super-wideband (SWB) audio and full-band (FB) audio. High-frequency (HF) components carry the detail features of audio signals such as brightness and naturalness. Due to the limitation of transmission bandwidth and storage capacity, audio signals are usually compressed by truncating the HF components in audio codec. In this case, the quality of the decoded audio signals will be obviously degraded, and it is necessary to reconstruct the HF components to realize the transmission of high quality at low bit-rate. For this reason, audio bandwidth extension (BWE) emerges as the times require. By this technique, the HF information can be efficiently reconstructed. Thus, the audio quality at the decoder can be improved to a large extent.

Recently, the widely-used BWE techniques are mainly non-blind. These methods need some HF side information to reconstruct the discarded HF information, which can not adapt the requirement of modern mobile audio communication. Therefore, the blind BWE method [1-4] becomes the key part of audio codec. It can reconstruct the HF components without any HF side information and be compatible with any type of audio codec. The traditional blind BWE methods estimate the HF sub-band envelope only by the low-frequency (LF) ones and ignore the inherent features of audio signals. Compared to these methods, RBF neural network takes into account of the LF features which embody the characteristics of different audio signals. In this paper, in order to reconstruct the energy of HF components more accurately, RBF neural network is used to estimate the HF sub-band envelope. In addition, the recovery methods of fine structure based on nonlinear prediction have gotten satisfactory results in recent years, and nonlinear prediction method [3] of RBF neural network has been used to reconstruct the fine structure of HF components in our early works [1-4]. Combining the above technologies, a blind BWE of audio signals based on RBF neural network is implemented.

The paper is organized as follows: The principles of the proposed BWE method are described in section 2. The quality test results are presented in Section 3 and the conclusions are given in Section 4.

## 2. NONLINEAR BANDWIDTH EXTENSION OF AUDIO SIGNALS BASED ON RBF NEURAL NETWORK

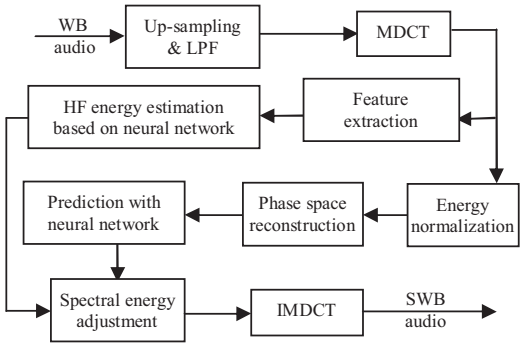


Fig. 1 Block diagram of the proposed method

The block diagram of the proposed BWE method is shown in Fig.1. Firstly, WB audio signal sampled at 16 kHz is up-sampled and low-pass filtered to obtain audio signal sampled at 32 kHz with bandwidth of 7 kHz. The band-limited SWB signal is transformed into frequency domain through Modified Discrete Cosine Transform (MDCT), and the LF MDCT coefficients are normalized by the root mean square (RMS) value of each sub-band to get the fine structure of LF components. Secondly, sub-band energy of HF spectrum is estimated by RBF neural networks which have been trained offline from SWB audio signals. At the same time, the fine structure of HF

components can be predicted by RBF neural networks according to nonlinear dynamics. Finally, the spectral envelope of the extended HF components is adjusted. Combining with the LF MDCT coefficients below 7 kHz, all the MDCT coefficients are transformed into time domain by an inverse MDCT to implement the BWE of audio signals.

In these modules, HF energy estimation module and nonlinear prediction module are the main modules which will be described later in more detail.

### 2.1. Estimation of the HF Spectral Envelope based on RBF Neural Network

*2.1.1. Introduction of RBF Neural Network*

The structure of RBF neural network [5-6] consists of three layers: the input layer, the hidden layer and the output layer. In the network training, different audio would be clustered based on the LF features to estimate the HF sub-band envelope. The function of the hidden layer is Gaussian kernel function with the traits of high nonlinearity, and many hidden layer neurons are combined to implement the function of nonlinear fitting.

*2.1.2. Feature Extraction*

In this paper, taking the perception and MPEG-7 timbre into consideration, 19 features of LF components [2] are extracted. These features have the correlation with the HF sub-band energy and can describe the characteristics of LF components.

The perception-based features are extracted in both time and frequency domain and the MPEG-7 low-level audio descriptions depict the spectrum distribution so as to describe the timber characteristics of audio signals. To produce a better estimation effect, both kinds of the features are combined to reflect the characteristics of different audio signals. The 19 features of LF components are chosen as follows:

a) Zero-crossing rate: $F_{zcr}$ shows the number that audio signal passes the zero level in each frame which is given by:

$$F_{\text{zcr}} = \sum_{n=1}^{N_t} | sign(s(n)) - sign(s(n-1)) | \tag{1}$$

where $N_t$=640 is the number of samples in each frame, and $s(n)$ is the audio signal in time-domain.

b) Gradient index: $F_g$ is the sum of gradients in various directions of signal and is computed as:

$$F_g = \sum_{n=2}^{N_t} \frac{\Psi(n) | s(n) - s(n-1) |}{\sqrt{\frac{1}{N_t} E}} \tag{2}$$

where $E$ represents the energy of current frame, and $\Psi(n)$ is an indicator of signal-changing directions which is defined by:

$$\Psi(n) = \frac{1}{2} | sign(s(n) - s(n-1)) - sign(s(n-1) - s(n-2)) | \tag{3}$$

c) Sub-band envelope: $F_{rms}(i)$, $i$=1,…,7, is the RMS of the $i^{th}$ sub-band and is computed as:

$$\boldsymbol{F_{rms}(i)} = \sqrt{\frac{\sum_{n=(i-1)*N_{sb}+1}^{i*N_{sb}} \boldsymbol{f_{mdct}(n)^2}}{\boldsymbol{N_{sb}}}} \quad i=1, 2, \ldots,7 \tag{4}$$

where $f_{mdct}(n)$, $n$=0,…,279, is the MDCT coefficients of the first 7 sub-bands in the current frame, $N_{sb}$=40 is the number of MDCT coefficients in each sub-band.

d) Flux of sub-bands: $F_f$ represents the amount of local spectral change and is given by:

$$F_f = \sum_{i=2}^{7} | F_{rms}(i) - F_{rms}(i-1) |^2 \tag{5}$$

e) Audio Spectrum Centroid: $F_{asc}$ describes the center of gravity of the log-frequency power spectrum and is computed by:

$$F_{asc} = \frac{\sum_{i=0}^{279} (\log_2(\frac{f_i}{1000}) p_i)}{\sum_{i=0}^{279} p_i} \tag{6}$$

where $f_i$ and $p_i$, $i$=0,…,279, represent the frequency and power value of the $i^{th}$ MDCT coefficient, respectively.

f) Audio Spectrum Spread: $F_{ASS}$ indicates the distribution of the log-frequency power spectrum and is defined by:

$$F_{ASS} = \sqrt{\frac{\sum_{i=0}^{279} (\log_2(\frac{f_i}{1000}) - F_{ASC})^2 p_i}{\sum_{i=0}^{279} p_i}} \tag{7}$$

g) Spectrum Flatness：$F_{sf}(i)$ , $i$=1,…,7, defines the ratio of geometric average and algebraic average of the MDCT coefficients in each sub-band which is given by:

$$F_{sf}(i) = \frac{\sqrt[N_{sb}]{\prod_{n=(i-1)*N_{sb}+1}^{i*N_{sb}} x(n)}}{\frac{1}{N_{sb}} \sum_{n=(i-1)*N_{sb}+1}^{i*N_{sb}} x(n)} \qquad i = 1,..., 7 \tag{8}$$

These features can describe the characteristics of LF information exactly. Meanwhile the energy of HF components to be estimated is represented by RMS value of 7 sub-bands in the frequency range from 7 to 14 kHz. Here, RBF neural network is applied to fit the LF features and the HF energy with audio data of the training sets.

*2.1.3. Energy Estimation of HF Sub-Band Based on RBF*

The RBF neural networks which have been trained by SWB audio data are used to estimate the HF energy and the block diagram of spectral energy estimation is depicted in Fig.2.

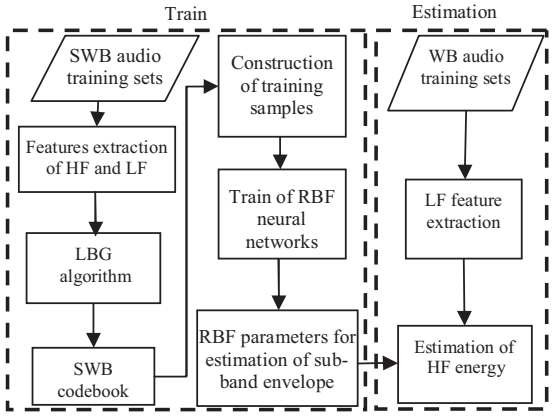


Fig. 2 Block diagram of HF energy estimation

At the training stage, a fifteen-minute-long SWB audio signal is used to extract the 19 LF features and the HF

sub-band envelope $\mathbf{F}_{high}$. All these features are computed to form the SWB vector $\mathbf{Y}=\{F_{zcr}, F_g, F_f, F_{asc}, F_{ass}, \mathbf{F}_{sf}, \mathbf{F}_{low}, \mathbf{F}_{high}\}^{\mathrm{T}}$, where $\mathbf{F}_{sf}=\{F_{sf}(i),\ i=1,\ldots,7\}^{\mathrm{T}}$, $\mathbf{F}_{low}=\{F_{rms}(i),\ i=1,\ldots,7\}^{\mathrm{T}}$ and $\mathbf{F}_{high}=\{F_{rms}(i),\ i=8,\ldots,14\}^{\mathrm{T}}$. Because the training of RBF neural network is limited by the sample number, all these SWB vectors are used to obtain the codebook based on LBG method used for vector quantization [7]. The training samples of RBF neural networks can be constructed from the codebook.

In consideration of the single output structure of RBF neural network, the values of sub-band envelope in different frequency range are taken as the expected output individually to form 7 groups of training samples. In that case, seven RBF neural networks are trained to estimate different sub-band envelope of HF components.

Four structural parameters of RBF neural networks can be calculated by the input vectors of the training samples. They are the number of neurons in input layer and hidden layer, the center and width of each neuron in hidden layer. Besides, in order to estimate different sub-band envelope of HF components, the weights between hidden layer and output layer of these RBF neural networks must be calculated, respectively. The parameters and weights can be obtained with the following steps:

1) Extract the LF features in each code word to construct the 19-dimension input vector of the training sample. The clustering centers of the input vectors are achieved by the K-means algorithm and chosen as the centers of neurons in the hidden layer. The number of neurons in the input layer is determined by the dimensions of input vector and the number of neurons in the hidden layer is determined by the number of clustering centers.
2) Calculate the width of each neuron in hidden layer by equation (9):

$$\sigma = \frac{L_{\max}}{\sqrt{2N_h}} \tag{9}$$

where $L_{\max}$ is the maximum Euclidean distance between two input vectors in the same clustering center，$N_h$ is the number of input vectors in each clustering center.

So far the structural parameters have been calculated.

3) Calculate the weights for estimating $F_{rms}(8)$.

The value of $F_{rms}(8)$ in each code word is taking as the expected output of sample, and combining with the 19-dimension input vector in the same code word to construct the training samples of $F_{rms}(8)$. These samples are utilized to calculate the weights between hidden layer and output layer based on Minimum Mean Squared Error (MMSE). The function of the hidden layer is Gaussian kernel function, so the output of the RBF neural network can be calculated as:

$$y_i = \sum_{j=0}^{N-1} w_j \exp\left(-\frac{1}{\sigma_j^2}\left\|x_i - c_j\right\|^2\right) \tag{10}$$

where $N$ is the number of the hidden layer, $w_j$ is the weight between hidden layer and output layer.

Let $p_{ij}$ as the output of hidden layer, the equation (10) can be simplified as:

$$y_i = \sum_{j=0}^{N-1} w_j p_{ij} \tag{11}$$

The sum of squared error is computed as:

$$E_{RBF} = \sum_{j=0}^{M} (s_i - y_i)^2 \tag{12}$$

where $M$ is the number of training samples, $s_i$ is the expected outputs of the RBF neural network.

In order to minimize the mean squared error, the weight $w_j$ is derived and written in the matrix form which is given by:

$$\mathbf{w} = (p^{\mathrm{T}} p)^{-1} p^{\mathrm{T}} \mathbf{s} \tag{13}$$

where $\mathbf{w}$ and $\mathbf{s}$ represent the column vector composes of the weight $w_j$ and the expected output $s_i$, respectively. $p$ is a $N \times M$ matrix corresponding to the output $p_{ij}$ of the hidden layer.

So far the weights of the RBF neural network for estimating $F_{rms}(8)$ have been calculated

4) Calculate the weights for estimating $F_{rms}(9)$~ $F_{rms}(14)$ in the same way of $F_{rms}(8)$.
5) Record the structural parameters and 7 groups of weights for estimating corresponding sub-band envelope of HF components.

By utilizing the network parameters and 7 groups of weights obtained from training, the sub-band energy of HF components can be estimated by the following steps:

1) Calculate the LF features of the WB audio in each frame, and form the LF features vector in the same order as the input vector of RBF neural networks when they are trained.
2) Construct 7 network models with the structural parameters and 7 groups of weights, respectively.
3) Estimate the sub-band envelope of HF components in the current frame by taking the LF features vector of each frame as input.

So far the sub-band envelope of HF components can be estimated.

### 2.2. Nonlinear Prediction of Fine Structure based on RBF Neural Network

Some studies have proven that the spectral series of audio signal has the obvious characteristics of chaos [8]. In that case, the theory of nonlinear dynamics can be used to process the audio signal in frequency-domain. In order to get more accurate test results of this envelope estimation method, the fine structure of the HF spectrum is predicted by the same method of our early work [3]. The main steps are as follows:

1) Transform the WB audio signal of each frame into frequency domain through MDCT, and the $N_m$ MDCT coefficients are normalized by the RMS value of each sub-band.
2) Process the MDCT coefficients by the theory of nonlinear dynamics. In order to reconstruct the phase space, the de-biasing autocorrelation method and False Nearest Neighbors (FNN) method are adopted to calculate the embedding time delay $\tau$ and the embedding dimension $m$, respectively. Thus, the one-dimension MDCT series $x(k)$， $k=1,2,\ldots, N_m$, can be reconstructed to obtain $m$-dimension phase point $\mathbf{X}(n)=\{x(n),x(n+\tau),\ldots,x(n+(m-1)\tau)\}$, $n=1,2,\ldots, N_m-(m-1)\tau$.

3) Calculate the clustering center with the K-means algorithm according to all the phase points, and get the width of each neuron in hidden layer by equation (9).
4) Let each phase point as the signal of input layer, and take the MDCT coefficients after the last dimension of each phase point as the expected output. Thus, $N_m$-($m$-1)$\tau$-1 training samples can be obtained. Train the RBF neural network by equations (10) ~ (13) to calculate the weights between hidden layer and output layer.
5) Regard the last phase point as the input, and the output value of RBF neural network as the prediction of the $(N_m+1)^{th}$ coefficient. In this way, the new phase point $\mathbf{X}(N_m-(m-1)\tau+1)$ can be constructed by utilizing the new MDCT coefficient.
6) Repeat the 3, 4, 5 steps to estimate the next HF MDCT coefficient until all the HF MDCT coefficients are obtained.

So far the fine structure of HF components can be reconstructed.

## 3. PERFORMANCE EVALUATION

In order to evaluate the performance, the proposed algorithm is compared with two kinds of blind audio BWE methods: the original RBF neural network method [3] which estimates the sub-band envelope of HF components by Linear Extrapolation (LE) and the maximum Lyapunov prediction method (MLP) [4]. The audio signals used for objective and subjective tests are derived from violin, trumpet, drum, guitar and symphony. The objective quality test of PEAQ [9] which is designed according to ITU-R BS.1387 standard was adopted. The scores of the PEAQ test is Objective Difference Grade (ODG) in the range from -4 (very annoying) to 0 (imperceptible difference). The subjective quality test was performed by A/B listening test. 12 listeners participated the A/B listening test. The objective evaluation result is given in Table 1 and the subjective test results are listed in Table 2 and Table 3 respectively.

Table 1 PEAQ comparisons results

| | ODG | | |
|---|---|---|---|
| | Original RBF | MLP | Proposed method |
| Violin | -2.810 | -2.804 | -2.338 |
| Trumpet | -3.273 | -2.986 | -2.300 |
| Drum | -3.034 | -2.878 | -2.526 |
| Guitar | -3.535 | -3.418 | -2.726 |
| Symphony | -2.936 | -2.825 | -2.162 |
| harmonica | -3.140 | -2.731 | -2.455 |
| speech | -3.133 | -2.653 | -2.307 |

Table 2 A/B test comparison between MLP and the proposed method

| Preference | Prefer MLP | Prefer proposed method | No preference |
|---|---|---|---|
| Percentage | 25.0% | 58.3% | 16.7% |

Table 3 Table 2 A/B test comparison between original RBF and the proposed method

| Preference | Prefer original | Prefer proposed method | No preference |
|---|---|---|---|
| Percentage | 16.7% | 50.0% | 33.3% |

The objective quality test and the subjective listening test results indicate that the proposed method outperforms the other two nonlinear BWE algorithms.

## 4. CONCLUSIONS

A blind BWE method from WB-to-SWB audio signals is proposed in the paper. RBF neural network is used to estimate the sub-band envelope of HF components effectively by fitting the relationship between LF features and HF sub-band energy. Moreover, the fine structure of HF spectrum is reconstructed according to the nonlinear characteristics of audio signal. Thus, the blind bandwidth extension is realized. Both the objective and subjective test results demonstrate that the proposed BWE method outperforms the reference BWE algorithms.

## 5. 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.

## 6. REFERENCES

[1] X. Liu, C. C. Bao, M. S. Jia and Y. T. Sha, "Nonlinear bandwidth extension based on nearest-neighbor matching," *Processing of the Second APSIPA ASC*, pp. 169-172, 2010.

[2] X. Liu, C. C. Bao, M. S. Jia and Y. T. Sha, "A harmonic bandwidth extension based on Gaussian Mixture Model," *ICSP2010*, Beijing, CHINA, October, pp. 474-477, 2010.

[3] H. J. Liu, C. C. Bao, X. Liu, X. T. Zhang and L. Y. Zhang "Audio Bandwidth Extension based on RBF Neural Network," *IEEE International Symposium on Signal Processing and Information Technology* (*ISSPIT2011*), Bilbao, Spain, December 14-17, pp. 150-154, 2011.

[4] X. T. Zhang, C. C. Bao, X. Liu, L. Y. Zhang, F. Bao and B. Bu "Audio Bandwidth Extension based on Maximum Lyapunov Prediction," *Processing of the Third APSIPA* (*Asia-Pacific Signal and Information Processing Association*) *Annual Summit and Conference* (*ASC*), Xi'an, China, October, pp. 18-21, 2011.

[5] J. Z. Lu, C. Y. Sun, N. Xu "Prediction model of chloride diffusion coefficients for concrete based on RBF neural network," *International Conference on Electric Technology and Civil Engineering* (*ICETCE*), pp. 2456-2459, 2011.

[6] L. X. Yang, "Based on improved RBF neural network for chaotic time series prediction," *Computational Intelligence and Natural Computing Proceedings* (*CINC*), pp. 124-127, 2010.

[7] C. C. Bao, *Principles of Digital Speech Coding*, Xi'an: Xidian University press, 2007.(in Chinese)

[8] Y. T. Sha, C. C. Bao, M. S. Jia, X. Liu "High frequency reconstruction of audio signal based on chaotic prediction theory," *IEEE International conference on Acoustics speech and signal processing* (*ICASSP*), pp. 381~384, 2010.

[9] ITU-R RECOMMENDATION BS.1387-1, Method for objective measurements of perceived audio quality, 1998-2001.