Liu and Bao *EURASIP Journal on Audio, Speech, and Music Processing* (2016) 2016:12
DOI 10.1186/s13636-016-0090-0

EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

Audio bandwidth extension using ensemble of recurrent neural networks

Xin Liu and Chang-Chun Bao*

Abstract

In audio communication systems, the perceptual audio quality of the reproduced audio signals such as the naturalness of the sound is limited by the available audio bandwidth. In this paper, a wideband to super-wideband audio bandwidth extension method is proposed using an ensemble of recurrent neural networks. The feature space of wideband audio is firstly divided into different regions through clustering. For each region in the feature space, a specific recurrent neural network with a sparsely connected hidden layer, referred as the echo state network, is employed to dynamically model the mapping relationship between wideband audio features and high-frequency spectral envelope. In the following step, the outputs of multiple echo state networks are weighted and fused by means of network ensemble, in order to further estimate the high-frequency spectral envelope. Finally, combining the high-frequency fine spectrum extended by spectral translation, the proposed method can effectively extend the bandwidth of wideband audio to super wideband. Objective evaluation results show that the proposed method outperforms the hidden Markov model-based bandwidth extension method on the average in terms of both static and dynamic distortions. In subjective listening tests, the results indicate that the proposed method is able to improve the auditory quality of the wideband audio signals and outperforms the reference method.

Keywords: Audio bandwidth extension, Ensemble of recurrent neural networks, Echo state network, Spectral translation

1 Introduction

Due to the restrictions of network transmission rate and data processing capability, the effective bandwidth of audio signals available for the existing audio communication systems is commonly limited. To improve coding efficiency [1], a practical perceptual audio coding framework would encode and transmit only the low-frequency (LF) components of the audio. However, state-of-the-art audio communication systems are not satisfied with the auditory quality achieved by the wideband (WB) audio reproduction, which are sampled at 16 kHz corresponding to the bandwidth of 50~7000 Hz. Contemporary audio communication systems are looking forward to explore the technique to ensure high-quality audio service with more brightness and expressiveness. Therefore, it is desirable to allow existing WB audio systems to achieve the auditory quality of super-wideband (SWB) audio systems, which are sampled at 32 kHz with the bandwidth of 50~14,000 Hz.

Among the methods to enhance the perceptual quality of the WB audio signals, blind bandwidth extension (BWE) is designed to analyze the statistical relationship between the low-frequency and high-frequency components of WB audio signals and artificially restore the missing high-frequency (HF) components in the frequency range of 7000~14,000 Hz from the decoded WB signals at the decoder end. The advantage of this approach is to avoid any modifications inside the source coding and the network transmission process [2]. In recent decades, many blind BWE solutions have been developed for speech and audio signals, and these typical BWE methods can be summarized to perform two main tasks, namely, the estimation of the spectral envelope and the extension of the fine spectrum [1, 2]. Informal listening test results indicate that the estimation accuracy of the HF spectral envelope is crucial to the improvement of auditory quality for the reproduced signals [3]. Therefore, most BWE approaches have concentrated on modelling the mapping relationship between LF and

* Correspondence: baochch@bjut.edu.cn
Speech and Audio Signal Processing Lab, School of Electronic Information and Control Engineering, Beijing University of Technology, 100124 Beijing, China

HF spectral coefficients based on statistical learning methods and further estimated the HF spectral envelope under some error criterions. In 1994, a statistical recovery method [4] was proposed to predict the HF spectrum and achieved an improved perceptual quality of the reproduced signals. In the same year, Carl and Heute proposed the spectral envelope estimation method based on codebook mapping (CBM) [5]. This method explored the joint codebook between LF features and the HF spectral envelope and modelled their bijective mapping relationship by training. Further, the improved methods with interpolation, soft decision, and split codebook were proposed to decrease the spectral distortion caused by the single codebook [6–8]. In 2000, the Gaussian mixture model-based spectral envelope estimation method was proposed [9]. The Gaussian mixture model (GMM) is adopted to mimic the joint probability density between LF and HF spectral coefficients, and the HF spectral envelope is estimated under the minimum mean square error criterion. The GMM method builds a soft clustering-based statistical model to restrain the spectral discontinuity of the audio signals reproduced by CBM and achieves better performance in both subjective and objective tests. In addition, feed-forward neural networks (FNNs) are also adopted to extend the spectral envelope [10, 11]. Iser and Schmidt made a comparison between FNN and CBM in the context of the BWE application [12]. The results show that there was no significant difference in auditory quality of the extended signals among these two methods but the FNN method required less computational complexity in comparison with the CBM method.

The abovementioned methods lay stress on the relationship between LF and HF spectral coefficients in the current frame and exhibit static characteristic of the audio spectrum. Jax and Vary brought the inter-frame correlation of audio features into extension of the spectral envelope and employed a hidden Markov model (HMM) to represent the temporal evolution of the spectral envelope for audio signals [13, 14]. Some other research results also show that the HMM-based BWE method obtained better performance in terms of both subjective and objective quality [15, 16]. In our previous research work [17], we adopted the HMM-based estimator to extend the bandwidth of WB audio to SWB and obtained good subjective and objective quality of the reproduced audio signals. While improved results are achieved in [17], some dynamic distortion is still perceived. This is caused by the fact that only discrete states are adopted to piecewise approximate the actual evolution of the audio spectrum over time. Thus, we further introduce a continuous state-space model into the spectral envelope extension and propose the blind BWE method based on the ensemble of recurrent neural networks in this paper. The feature space of wideband audio is firstly divided into different regions through clustering. For each region in the feature space, a specific recurrent neural network with a sparsely connected hidden layer, referred as the echo state network, is employed to dynamically model the mapping relationship between wideband audio features and high-frequency spectral envelope. Different from the traditional fully connected recurrent neural networks, an echo state network (ESN) adopts the sparsely connected hidden layer in which the connectivity and weights of hidden neuron nodes are fixed and randomly assigned. The recurrent nature of the connections turns the time varying WB audio features into specific temporal patterns in high dimension. Then, a simple linear transformation is used to map the state of hidden neuron nodes to the desired HF spectral envelope. Given a rich collection of recurrently connected nodes, the ESN is able to effectively fit the nonlinear mapping function between WB audio features and the HF spectral envelope. Next, the outputs of the parallel ESNs for different regions in the feature space are weighted and fused by means of network ensemble techniques, so as to form ensemble echo state networks (EESNs) for further estimating the HF spectral envelope. Finally, combining with the HF fine spectrum extended by spectral translation, the proposed method can effectively extend the bandwidth of the WB audio to SWB and upgrade the subjective and objective quality of the WB audio.

2 Audio bandwidth extension using ensemble echo state network

2.1 Overview

The block diagram of the proposed blind BWE method is shown in Fig. 1. The input signals are the WB audio signals sampled at 16 kHz with a bandwidth of 7 kHz. Through up-sampling and low-pass filtering, the audio signals sampled at 32 kHz with the bandwidth of 7 kHz are separated into 32-ms frames with 16-ms overlap. Then, the audio signals are windowed with the Hamming window and are transformed into the frequency domain by using fast Fourier transform (FFT). The audio amplitude spectrum $A(k)$ in the frequency range of 64~7000 Hz is mapped onto the perceptually correlated equivalent rectangular bandwidth (ERB) scale by 20-channel triangular overlapping windows, in order to extract the 20-dimensional Mel frequency cepstral coefficient (MFCC) vector F_{MFCC}. Next, the extracted MFCCs are normalized to the range between 0 and 1 via min-max normalization. The normalized MFCCs are fed to the pre-trained EESN for estimating the HF spectral envelope. Here, the HF spectral envelope, $F_{\mathrm{RMS}}(i)$, $i = 0,\ldots, 3$, is represented by the root mean square (RMS) values of the audio

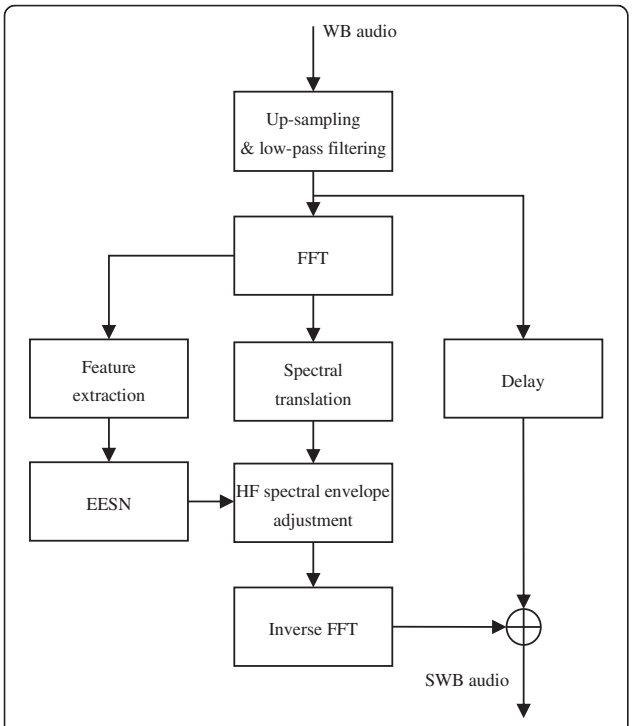

Fig. 1 Block diagram of the proposed blind bandwidth extension method for audio signals

amplitude spectrum in four nonuniform sub-bands in the frequency band of 7~14 kHz as follows:

$$F_{\mathrm{RMS}}(i) = \sqrt{\frac{1}{h(i)-l(i)+1}\sum_{k=l(i)}^{h(i)} A^2(k)}, \quad i = 0, 1, ..., 3 \tag{1}$$

where $A(k)$ is the spectral magnitude of audio signals and $h(i)$ and $l(i)$ correspond to the lower and upper boundaries of frequency bins in the ith sub-band, respectively. Note that there is no overlap between the adjacent sub-bands. The central frequencies of HF sub-bands are arranged at 8470, 9338, 11,653, and 13,657 Hz, according to the perceptually correlated ERB scale.

The extension of the fine spectrum is less critical for the subjective auditory quality of the extended audio signals, in comparison with the extension of the spectral envelope [2]. Thus, in order to reduce algorithm complexity, the HF fine spectrum is extended by using the simple spectral translation method [2]. The LF components are replicated into the HF band in the range of 7~14 kHz. The spectral envelope of the extended HF components is further adjusted by employing the estimated RMS values $\hat{F}_{\mathrm{RMS}}(i)$ of HF sub-bands. Finally, the reconstructed HF components are converted to the time domain by the inverse FFT, windowing, overlapping, and addition operations. By combining the WB audio signals with appropriate delay, the proposed method can reconstruct the SWB audio signals sampled at 32 kHz with the bandwidth of 14 kHz. Details of the proposed BWE method are discussed below.

2.2 Estimation of spectral envelope based on EESN

Let $F_X(m)$ denote the MFCCs of WB audio signals in the mth frame, whose dimension is set to $d_X = 20$. $F_Y(m)$ is the RMS value of HF sub-bands in the mth frame, and its dimension is set as $d_Y = 4$. Through an unknown mapping function $H(\bullet)$, the WB features $F_X(m)$ can be converted into the HF spectral envelope $F_Y(m)$ as follows:

$$F_Y = H(F_X) \tag{2}$$

We assume that the unknown mapping function can be modelled by ensembling several different state-space models. If a kth state-space model is applied, the process of regenerating the HF spectral envelope is approximated by

$$S_{\mathrm{hidden}}(m,k) = H_{\mathrm{state}}{}^{(k)}(S_{\mathrm{hidden}}(m-1,k), F_X(m-1), U(m,k)) \tag{3}$$

$$\hat{F}_Y\left(m|k\right) = H_{\mathrm{observation}}{}^{(k)}(S_{\mathrm{hidden}}(m,k), F_X(m), V(m,k)) \tag{4}$$

where $S_{\mathrm{hidden}}(m, k)$ with the dimension d_S is the hidden state vector of the kth state-space model in the mth frame, and its temporal evolution determines the dynamic characteristics of state space; $\hat{F}_Y(m|k)$ is the HF spectral envelope estimated by the kth state-space model in the mth frame; $H^{(k)}_{\mathrm{state}}(\bullet)$ is referred to as the state updating equation of the kth state-space model and describes the evolution process of hidden states given the input vector $F_X(m-1)$; $H^{(k)}_{\mathrm{observation}}(\bullet)$ is referred to as the observation equation of the kth state-space model and denotes the mapping process from $F_X(m)$ to $\hat{F}_Y(m|k)$ controlled by the hidden state $S_{\mathrm{hidden}}(m, k)$; and $U(m, k)$ and $V(m, k)$ represent the error caused by the state updating and observation processes, respectively.

By integrating different state-space models, we can estimate the HF spectral envelope $F_Y(m)$ from the WB audio feature $F_X(m)$. Therefore, there remain two important issues, namely, how to build an effective state-space model for approximating the true mapping process and how to choose the suitable state-space models among candidate models at each time.

2.2.1 Model structure of ESN

$H^{(k)}_{\mathrm{state}}(\bullet)$ and $H^{(k)}_{\mathrm{observation}}(\bullet)$ can be separately represented by two FNNs to build up a state-space-based recurrent neural network [18] for fitting the true mapping process.

Motivated by this, ESN [19, 20] is employed as the state-space model in this paper. Different from traditional fully connected recurrent neural networks, ESN adopts the sparsely connected hidden layer in which the connectivity and weights of hidden neuron nodes are randomly assigned and unchanged during training. The recurrent nature of the connections turns the time varying WB audio features into specific temporal patterns in high dimension. From the perspective of dynamic system theory, these states of hidden neuron nodes serve as a memory of the input audio feature and are termed as echoes of the input history. This is where ESN draws its name from. Given a rich collection of recurrently connected nodes with nonlinear activation functions, such a hidden layer, being an input-driven dynamical model, could provide a rich and relevant enough feature space, such that the desired HF spectral envelope could be obtained by linear combination from the state of hidden neuron nodes.

In the ESN-based state-space model, the updating process of hidden states for the kth ESN is represented by a leaky-integrated nonlinear activation unit [21] as follows:

$$\tilde{S}_{\text{hidden}}(m,k) = \tanh\left(W^{\text{in}}(k)\begin{bmatrix} 1 \\ F_X(m-1) \end{bmatrix} + W^{res}(k) S_{\text{hidden}}(m-1,k) \right) \quad (5)$$

$$S_{\text{hidden}}(m,k) = (1-\alpha) S_{\text{hidden}}(m-1,k) + \alpha \tilde{S}_{\text{hidden}}(m,k) \quad (6)$$

where $\tilde{S}_{\text{hidden}}(m,\ k)$ is the updating state of $S_{\text{hidden}}(m,\ k)$ in the mth frame. Leaking rate α of the leaky-integrated nonlinear unit represents the dynamic updating speed of hidden states. When the value of α is small, the dynamic change of hidden states may be slow. If α is close to 1, the nonlinear unit approximates a tanh function. $W^{\text{in}}(k)$ is the input weight matrix with the dimension of $d_S \times (d_X + 1)$, and the elements of $W^{\text{in}}(k)$ are uniformly distributed in $[-a_{\text{in}},\ a_{\text{in}}]$. The scaling factor a_{in} of W^{in} decides the nonlinearity of the activation units. When a_{in} is close to 0, the leaky-integrated nonlinear activation units approach linear functions. For a larger a_{in}, the state updating process driven by $F_X(m-1)$ exhibits more nonlinearity. $W^{\text{res}}(k)$ is the recurrent weight matrix with the dimension of $d_S \times d_S$, and its spectral radius a_{res} can be artificially adjusted for controlling network stability in the actual applications. It is noteworthy that a_{res} and a_{in} jointly decide the relative importance between $F_X(m-1)$ and $S_{\text{hidden}}(m-1,\ k)$ in the state updating process. If a_{res} is larger, the previous hidden states own more effects on the current hidden states and the recurrent network would show more long-term correlations. Otherwise, the input vectors become the key factor to derive the hidden states. In addition, some research works indicated that the sparsely connected hidden states could slightly improve the estimation accuracy of ESN and accelerate the speed of model training [19, 20]. So, we can change the percentage of nonzero elements relative to the total amount of elements in W^{res} to adjust the sparsity f_{sparsity} of the recurrent weight matrix.

ESN takes advantage of a simple linear mapping function to represent the observation equation $H^{(k)}_{\text{observation}}(\bullet)$ [19] in the following way:

$$\hat{F}_Y(m|k) = W^{\text{out}}(k)\begin{bmatrix} 1 \\ F_X(m) \\ S_{\text{hidden}}(m,\ k) \end{bmatrix} \quad (7)$$

where $W^{\text{out}}(k)$ is the output weight matrix with the dimension of $d_S \times (1 + d_X + d_S)$ for the kth ESN.

From the perspective of network structure, ESN includes an input layer (F_X), a hidden layer (S_{hidden}), and an output layer (F_Y), as shown in Fig. 2. Here, the hidden layer, also referred to as the reservoir in ESN, consists of a large number of sparse-connected neuron nodes. With the help of the nonlinearity, sparsity, and feedback structure of neurons, the reservoir expands the audio features into a nonlinear high-dimensional feature space and further provides a more temporal context, so as to make the recurrent neural network exhibit rich dynamic characteristics [22]. The conclusions in relevant studies on recurrent learning algorithms [23] show that the weight dynamics favor a functional partition of the recurrent network into a fast output layer and a slower dynamical reservoir, whose rates of weight change are closely coupled. Therefore, ESN randomly generates the

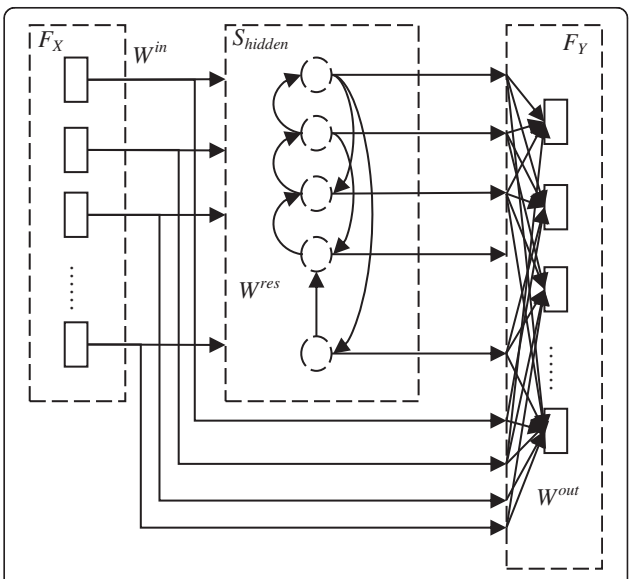

Fig. 2 Network structure of a single echo state network

input and recurrent weight matrices in network initialization and fixes them during network learning. Then, the network is optimized by adjusting the output weight matrix to fit the unknown nonlinear mapping function $H(\bullet)$.

2.3 Model training for ESN

The WB audio feature $F_X(m)$ and the HF spectral envelope $F_Y(m)$ are separately extracted frame by frame from the aligned WB and SWB audio signals, which are available for the training procedure, so as to form the training sample data $[F_X(m)^T, F_Y(m)^T]^T$, where $m = 0,1,\ldots, N_{\text{train}} - 1$ denotes the index of audio frame. In the model training of ESN, the hidden state $S_{\text{hidden}}(-1, k)$ of the kth ESN model is first initialized as 0. Next, $F_X(m)$ is fed into multiple pre-defined ESNs for driving the reservoirs. Then, by constantly updating the hidden states, the reservoirs exhibit different dynamic characteristics. Finally, the output weight matrix of the kth ESN is computed according to the information collected from the sequences of $F_X(m)$, $F_Y(m)$, and $S_{\text{hidden}}(m, k)$. A more detailed description about model training is provided below.

1. Initializing the reservoir parameters
   The initialization of reservoir parameters uses a heuristic method [19] as follows.
   (a) A matrix with the dimension of $d_S \times d_S$ is randomly created, and its element values are uniformly distributed within [−1, 1]. A fraction of the small element values within the random matrix are set to 0, in order to form a sparse matrix W. The sparsity f_{sparsity} is used to adjust the percentage of nonzero elements relative to the total amount of elements in W. Then, W is normalized according to its spectral radius $\lambda_{\max}$, i.e., $W/\lambda_{\max}$.
   (b) The resulting sparse matrix is scaled to obtain the recurrent weight matrix, i.e., $W^{\text{res}} = a_{\text{res}} W/|\lambda_{\max}|$. Here, $a_{\text{res}} < 1$ is the spectral radius of W^{res} and can be experimentally adjusted.
   (c) The input weight matrix W^{in} is randomly created according to a uniform distribution.
2. Driving the reservoir
   The hidden states $S_{\text{hidden}}(m, k)$ is updated in the light of two model parameters $\{W^{\text{res}}(k), W^{\text{in}}(k)\}$. But the update process is impacted by the initial hidden state $S_{\text{hidden}}(-1, k)$ [19]. Therefore, we set a network stability threshold $N_T = 13$ and assume that the whole network is approaching stability after the first N_T state updates.
   Starting with the $N_{T\text{th}}$ updates, the sample data $F_X(m)$ of the WB audio feature, the hidden states $S_{\text{hidden}}(m)$, and the sample data $F_Y(m)$ of the HF spectral envelope are collected for establishing the state collection matrix B and the expectation output matrix Q. The column vectors of the state collection matrix $B(k)$ for the kth ESN are represented as $[1, F_X(m)^T, S_{\text{hidden}}(m, k)^T]^T$, so the dimension of $B(k)$ is $(1 + d_X + d_S) \times (N_{\text{train}} - N_T)$. The column vectors of the expectation output matrix $Q(k)$ for the kth ESN are represented as $[F_Y(m)]$, so the dimension of $Q(k)$ is $(d_Y) \times (N_{\text{train}} - N_T)$. Through the abovementioned method of data collection, the negative effect of $S_{\text{hidden}}(-1, k)$ on ESN is effectively eliminated, and the mapping relationship between the WB audio feature $F_X(m)$ and the HF spectral envelope $F_Y(m)$ can be properly reflected according to the model trained from $B(k)$ and $Q(k)$.
3. Computing the output weight matrix

The output weight matrix $W^{\text{out}}(k)$ for the kth ESN is computed according to $B(k)$ and $Q(k)$ to minimize the error between the network output $\hat{F}_Y(m|k)$ and the true HF spectral envelope $F_Y(m)$. Informal evaluation indicates that, when the element value of $W^{\text{out}}(k)$ is large, $W^{\text{out}}(k)$ may amplify the slight difference of $F_X(m)$ and over-fitting may occur during network training. Therefore, the ridge regression method is employed in order to avoid over-fitting as follows:

$$W^{\text{out}}(k) = \underset{W^{\text{out}}(k)}{\arg\min}\left(\left[\frac{1}{N_{\text{train}} - N_T}\sum_{m=0}^{N_{\text{train}}-N_T-1}\|F_Y(m) - \hat{F}_Y(m|k)\|^2\right] + \beta\|W^{\text{out}}(k)\|^2\right) \tag{8}$$

where $\beta\|W^{\text{out}}(k)\|^2$ is the regular item for penalizing the large elements in $W^{\text{out}}(k)$ and the relative importance between two items in the above error function can be controlled by adjusting β.

Finally, the resulting $W^{\text{out}}(k)$ can be given as

$$W^{\text{out}}(k) = Q(k)B(k)^T\left[B(k)B(k)^T + \beta I\right]^{-1} \tag{9}$$

where I is a unit matrix.

The connection weight matrices $W^{\text{res}}(k)$, $W^{\text{in}}(k)$, and $W^{\text{out}}(k)$ for the kth ESN are obtained by using the training method above. In the actual extension, the frame-wise MFCCs are fed into the ESNs. The hidden states $S_{\text{hidden}}(k)$ are updated with the change of F_{MFCC} in a frame-by-frame basis. The dynamic characteristics of WB audio features are brought into the network model. Further, the HF spectral envelope $F_Y(m)$ is effectively estimated by fusing the outputs $\hat{F}_Y(m|k)$ of multiple ESNs.

2.3.1 Network ensemble fusion

According to the network ensemble theory, frequently, an ensemble of models performs better than any individual model. This is because the various errors of the models average out [24]. In this paper, a network ensemble framework based on the statistical distribution of input feature space is adopted to integrate multiple ESNs for further forming ensemble echo state network.

Figure 3 shows the block diagram of the EESN-based spectral envelope estimator. In the model training procedure of EESN, we first selected the GMM-based clustering method to separate the feature space of the input WB audio feature F_X into N_{ESN} regions. The mean vector of each Gaussian component corresponds to a clustering center of the feature space. Accordingly, the training sample data $[F_X(m)^T, F_Y(m)^T]^T$, $m = 0,1,\ldots, N_{\mathrm{train}} - 1$ could be segmented into N_{ESN} groups on the basis of the separation of feature space. By assuming that the statistical properties of each group of input features were similar, N_{ESN} ESNs with the same network structure were built up to model the local mapping relationship between F_X and F_Y within the corresponding N_{ESN} regions of the feature space. It is worth mentioning that the WB audio features extracted from the training data can be simultaneously fed into all the ESNs in order to continuously drive their state updating equations since their input and recurrent weight matrices are unchanged during model training, while the output weight matrices of different ESNs need to be respectively trained according to their corresponding regions in the feature space. Therefore, the feature space separation-based network ensemble framework does not actually affect temporal modelling of ESNs.

In the actual bandwidth extension procedure, the mth-frame audio feature $F_X(m)$ is fed into all of these ESNs, so as to obtain the corresponding estimated values $\hat{F}_Y(m|k)$ of the HF spectral envelope for the kth ESN. Meanwhile, according to the probability distribution of the WB audio feature modelled by GMM, a posteriori probability $p(k|F_X(m))$ of the kth region in the feature space is computed and adopted as the weights for guiding us on integrating the HF spectral envelope estimated by different ESNs.

Within the GMM framework, $p(k|F_X(m))$ is computed as follows:

$$p(k|F_X(m)) = \frac{p(F_X(m)|k)p(k)}{\sum_{i=0}^{N_{\mathrm{ESN}}-1} p(F_X(m)|i)p(i)} \tag{10}$$

where $p(k)$ is the a priori probability of the kth region in the feature space, and it needs to match the following condition:

$$\sum_{k=0}^{N_{\mathrm{ESN}}-1} p(k) = 1, \quad 0 \le p(k) \le 1 \tag{11}$$

$p(F_X(m)|k)$ is the probability of $F_X(m)$ given the kth region, and it is represented as

$$\begin{aligned} p(F_X(m)|k) &= N(F_X(m); m_k, C_k) \\ &= \frac{1}{(2\pi)^{\frac{D_X}{2}}|C_k|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(F_X(m)-m_k)^T C_k^{-1}(F_X(m)-m_k{}^k)\right] \end{aligned} \tag{12}$$

where $N(\bullet)$ represents the Gaussian probability density function. m_k and C_k are the mean and covariance of the kth Gaussian component, respectively, and can be trained according to the standard EM algorithm. Finally,

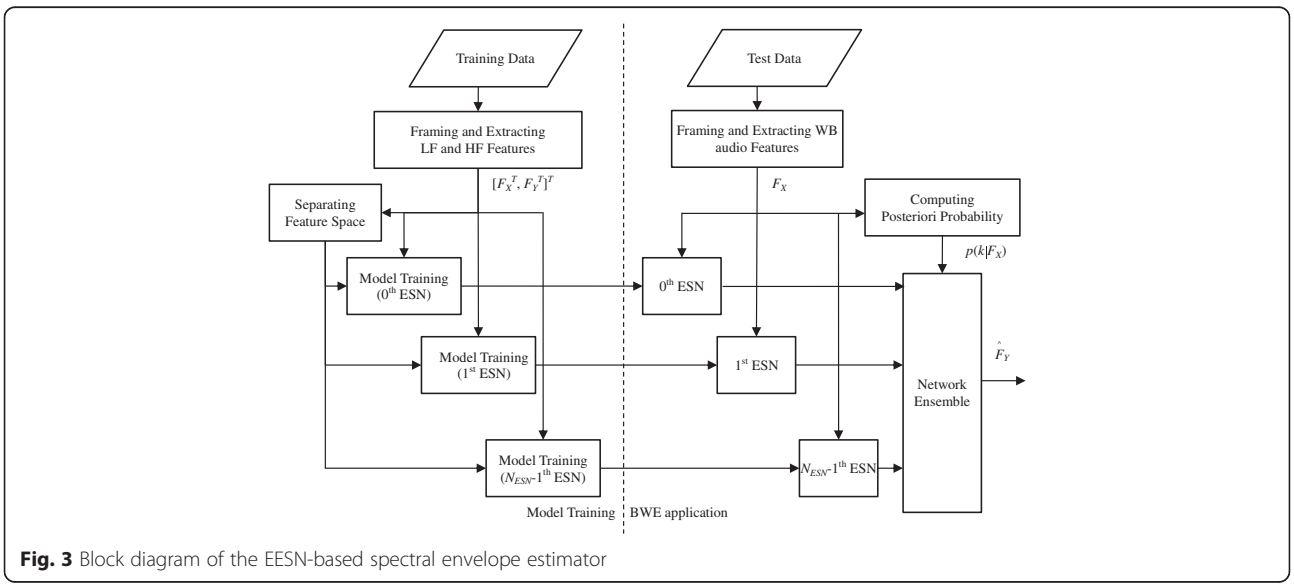

Fig. 3 Block diagram of the EESN-based spectral envelope estimator

in the GMM-based soft-decision ensemble framework, multiple ESNs can be linearly combined by taking advantage of the posterior probabilities $p(k|F_X(m))$, and the estimated HF spectral envelope $\hat{F}_Y(m)$ is given by

$$\begin{aligned}\hat{F}_Y(m) &= \sum_{k=0}^{N_{\mathrm{ESN}}-1} p(k|F_X(m))\hat{F}_Y(m|k) \\ &= \sum_{k=0}^{N_{\mathrm{ESN}}-1} p(k|F_X(m))W^{\mathrm{out}}(k)\begin{bmatrix}1\\ F_X(m)\\ S_{\mathrm{hidden}}(m,\ k)\end{bmatrix}\end{aligned} \tag{13}$$

2.4 High-frequency component reconstruction

In general, the LF and HF components have the similar spectral structure. The slightly different fine structure for the audio spectrum would not significantly affect the auditory quality of the bandwidth-extended audio signals, if the HF spectral envelope is well restored. So, in the proposed blind BWE method, spectral translation is adopted to directly copy the fine spectrum in the frequency range of 0~7 kHz to the HF band of 7~14 kHz. Here, the LF fine spectrum is represented by the normalized spectral magnitude $A_{\mathrm{norm}}(k)$, $k = 0,\ldots, 223$, as follows:

$$A_{\mathrm{norm}}(k) = \frac{A(k)}{F_{\mathrm{RMS_WB}}(i)}, \quad i = \frac{k}{N_{\mathrm{subband}}} \tag{14}$$

where $A(k)$ is the spectral magnitude of audio signals. $F_{\mathrm{RMS\_WB}}(i)$ is the RMS value of the LF sub-bands for describing the LF spectral envelope of audio signals. Here, in order to maintain the spectral flatness of $A_{\mathrm{norm}}(k)$, the LF spectrum is uniformly divided into 14 sub-bands in the frequency scale and each LF sub-band contains $N_{\mathrm{subband}} = 16$ frequency bins. Then, the HF fine spectrum extended by spectral translation is given by

$$A_{\mathrm{norm}}(k) = A_{\mathrm{norm}}(k-224), \quad k = 224, \ldots.447 \tag{15}$$

Since the HF spectral envelope is analyzed in the ERB scale, the HF fine spectrum is also divided into four non-overlapping sub-bands. By combining the estimated HF sub-band RMS values $\hat{F}_{\mathrm{RMS}}(i)$, $i = 0,\ldots, 3$, the extended HF spectrum $A_{\mathrm{swb}}(k)$, $k = 224,\ldots, 447$, can be presented as

$$A_{\mathrm{swb}}(k) = A_{\mathrm{norm}}(k)\hat{F}_{\mathrm{RMS}}(i), \quad i = \mathrm{Subband}(k) \tag{16}$$

in which Subband(k) represents the index of the HF sub-band corresponding to the kth frequency bin.

The phase of HF spectrum $\theta(k)$, $k = 224,\ldots, 447$, is also obtained by spectral translation as follows:

$$\theta(k) = \theta(k-224), \quad k = 224, \ldots.447 \tag{17}$$

Finally, the HF spectrum is converted into the time domain by inverse FFT. The resulting HF signals are combined with the properly delayed WB signals to reconstruct the SWB audio signals.

3 Parameter selection for EESN

In this section, we made some preliminary experiments in order to select the appropriate model parameters according to the objective performance of spectral envelope extension. The SWB audio data used for parameter selection came from live concert recordings with the length of about 6 h and contain dialog, pop music, singing, and live background sounds. These signals were sampled at 32 kHz with the bandwidth of 14 kHz and were stored by 16-bit PCM. After low-pass filtering and time alignment, the parallel WB training dataset was built up. Twenty-dimensional MFCCs and four-dimensional HF sub-band RMS values were extracted from the WB and SWB datasets as the input vectors F_X and the expectation output vectors F_Y for EESN, respectively. Seventy-five percent of these sample data were randomly selected for training the EESN model, and others were used as the validation set for parameter selection.

In this section, the mean square error between the original and estimated spectral envelope in the frequency range of 7~14 kHz was selected as the objective measurement for evaluating the performance of the EESN method. The estimation error of HF spectral envelope can be defined as

$$e_{\mathrm{MS}}(i) = \frac{1}{4}\sum_{n=0}^{3}\left(F_{\mathrm{RMS}}(i,n) - \hat{F}_{\mathrm{RMS}}(i,n)\right)^2 \tag{18}$$

where $e_{\mathrm{MS}}(i)$ is the mean square error of the HF spectral envelope for the ith frame, $F_{\mathrm{RMS}}(i, n)$ and $\hat{F}_{\mathrm{RMS}}(i,n)$ are the true and estimated RMS values of the nth HF sub-band in the ith audio frame, respectively, and $n = 0, 1, 2, 3$ is the index of HF sub-band. Before computing the mean square error, all the data needed to be resampled to 32 kHz and were temporally aligned with the original SWB audio. Then, the resulting error values were averaged over all the frames for each audio signal in the validation set, in order to be used as the distortion measure of the spectral envelope estimation.

The extension performance of the EESN method is mainly affected by seven model parameters: the scaling factor of W^{in} (a_{in}), spectral radius of W^{res} (a_{res}), sparsity of W^{res} (f_{sparsity}), leaking rate (α), regular factor for ridge regression (β), dimension of hidden state (d_S), and number of ESNs (N_{ESN}). In EESN, different ESNs are generally independent of each other, so we can first reduce

the EESN model to a single ESN model by setting $N_{\mathrm{ESN}} = 1$ and optimize the ESN model parameters (a_{in}, a_{res}, f_{sparsity}, α, β, and d_S), by using grid search. Next, several ESNs are constructed with the "optimal" ESN parameters and are integrated for further forming EESN, and then, the optimal value of N_{ESN} can be manually determined, guided by evaluation on the validation set.

3.1 ESN model parameters

In order to improve search efficiency, a two-stage grid search was conducted to optimize the ESN model parameters, while the number of ESN was preset as 1.

In the first-stage search, we manually set the bounds for each model parameter, according to their actual characteristics in the ESN model. Then, a finite set of reasonable values for each parameter were pre-selected, as shown in Table 1, so as to perform the coarse grid search over wider parameter intervals. Next, an ESN was trained with each candidate in the Cartesian product of these six parameter sets. Finally, their performance was evaluated on the validation set in terms of the estimation error of the HF spectral envelope. According to the experiment results, the parameter setting that achieves the lowest estimation error (12.7824) in the validation procedure is obtained as $\{a_{\mathrm{in}} = 1/8,\ a_{\mathrm{res}} = 1/2,\ f_{\mathrm{sparsity}} = 1/4,\ \alpha = 1,\ \beta = 3,\ d_S = 400\}$. The resulting parameter setting can be used to identify the promising regions for each model parameter: $a_{\mathrm{in}} \in (1/16, 1/4)$, $a_{\mathrm{res}} \in (1/4, 3/4)$, $f_{\mathrm{sparsity}} \in (1/8, 1/2)$, $\alpha \in (4/5, 1)$, $\beta \in (2, 4)$, and $d_S \in (200, 600)$.

The second-stage gird search was further adopted with smaller steps, so as to determine the "optimal" ESN model parameters. For each model parameter, five grid points were equally spaced within the corresponding promising region. Then, all the 5^6 possible combinations of the six ESN model parameters were evaluated, and the lowest estimation error (12.4961) of the HF spectral envelope was obtained at $\{a_{\mathrm{in}} = 0.1094,\ a_{\mathrm{res}} = 0.625,\ f_{\mathrm{sparsity}} = 0.2188,\ \alpha = 1,\ \beta = 3,\ d_S = 500\}$.

In addition, we should note that there are no evident differences for different leaking rates of the leaky-integrated nonlinear activation unit in terms of the estimation error values of the HF spectral envelope, and $\alpha = 1$ provides a smaller estimation error than other α values, according to the evaluation results in the two-stage grid search process. This suggests that the long-term memory of hidden states does not clearly improve the performance of the ESN model.

Table 1 Candidates of the ESN model parameters for coarse grid search

a_{in}	{1/16, 1/8, 1/4, 1/2, 1}
a_{res}	{1/4, 1/2, 3/4, 1, 5/4}
f_{sparsity}	{1/16, 1/8, 1/4, 1/2, 1}
α	{1/5, 2/5, 3/5, 4/5, 1}
β	{0, 1, 2, 3, 4}
d_S	{200, 400, 600, 800, 1000}

3.2 Number of ESNs

EESN adopts the a posteriori probabilities of each region in the feature space as weights to integrate the outputs of different ESNs, which model the local mapping functions between LF and HF components within the corresponding regions of the feature space. So, the number of ESNs is an important factor for the proposed EESN-based BWE method. According to the results of the two-stage grid search, the model parameters for each ESN are set as $a_{\mathrm{in}} = 0.1094$, $a_{\mathrm{res}} = 0.625$, $f_{\mathrm{sparsity}} = 0.2188$, $\alpha = 1$, $\beta = 3$, and $d_S = 50$. Then, we experimented with several values of N_{ESN} to form different EESNs and evaluated them under the estimation error of the HF spectral envelope. The mean and standard deviation of the spectral envelope estimation error are shown in Table 2. It is noticeable that the performance of the EESN-based spectral envelope estimator improves quickly with the increase of N_{ESN} and the mean estimation error of HF spectral envelope is reduced down to about 11.3. With the further consideration of computation complexity and memory demands, the number of ESN can be set as 8.

As a consequence of the above, the EESN model parameters are finally selected as $a_{\mathrm{in}} = 0.1094$, $a_{\mathrm{res}} = 0.625$, $f_{\mathrm{sparsity}} = 0.2188$, $\alpha = 1$, $\beta = 3$, $d_S = 50$, and $N_{\mathrm{ESN}} = 8$.

4 Performance evaluation

In this section, log spectral distortion, cosh measure, and differential log spectral distortion (DLSD) were adopted as objective measurements for evaluating the performance of the proposed BWE method in comparison with the HMM-based reference method. In addition, the auditory quality improvement of the extended audio signals was further discussed in terms of subjective preference tests.

Table 2 LSD of EESN model with different N_{ESN}

N_{ESN}	Spectral envelope estimation error	
	Mean	Standard deviation
1	12.4961	3.5859
2	12.0504	3.7867
4	11.6919	3.0442
6	11.4871	4.9626
8	11.3662	3.6441
16	11.3384	4.7234
32	11.3074	3.5859

4.1 Reference method and audio data

Except for the spectral envelope estimation module, the HMM-based reference method follows the similar principle shown in Fig. 1. The MFCCs were extracted from the WB audio and fed into the HMM-based Bayesian estimator of the HF spectral envelope under the minimum mean square error criterion, while the HF fine spectrum was extended by spectral translation. In the HMM-based reference method, the HMM states were defined by vector quantization of the joint space between LF and HF coefficients. For each state of the HMM, the state-dependent conditional expectation of the HF spectral envelope was calculated based on a GMM of the joint probability density of the WB audio feature and HF spectral envelope. The respective a posteriori probabilities of the corresponding states of the HMM were used to weight the conditional expectation, in order to obtain the estimation of the HF spectral envelope. Here, the HMM parameters were learned by using the hybrid approach suggested by Jax and Vary [13], and the state number of HMM and mixture number for each GMM were also selected in terms of the estimation error of the HF spectral envelope. The experiment results indicated that the estimation error was slightly decreased with an increase of the number of hidden states and Gaussian mixtures. Finally, taking into consideration algorithm complexity and auditory quality, we set the number of hidden states as 16 and required each GMM to have 32 mixtures and full covariance matrices in the HMM-based reference method.

The audio data of the training and validation set for the proposed and reference methods were from the lossless audio recorded in a live concert with the length of about 6 h. They contained dialog, music, singing, and background sound. Through resampling and low-pass filtering, the parallel WB and SWB audio datasets were obtained, and the level of all the signals needed to be normalized to −26 dBov before further processing. In addition, 15 SWB audio signals were selected from the MPEG audio quality listening test dataset as the test data and contained pop music, solo instrumental music, symphony, and speech. These signals were sampled as 32 kHz with the bandwidth of 14 kHz and were limited in the length of 10~20 s. Then, they were converted into the WB audio through low-pass filtering and downsampling, and the level of the resulting signals were normalized to −26 dBov as the input of the BWE methods. Since disjoint speakers and disjoint music pieces were used in training, validation, and testing, the accuracy and generalization of the proposed EESN method could be evaluated more objectively and fairly. A detailed analysis of the subjective and objective quality for the extended signals is given below.

4.2 Objective measurement

The objective quality of the SWB signals reproduced by different methods was evaluated in terms of log spectral distortion (LSD), cosh, and DLSD measurements.

4.2.1 LSD

The LSD [25, 26] between the original SWB audio and the extended audio signals in the frequency range of 7~14 kHz was selected as the objective measurement for comparing the extension performance between the EESN method and HMM method. LSD measurement is computed directly from the FFT power spectra as

$$d_{\mathrm{LSD}}(i) = \sqrt{\frac{1}{N_{\mathrm{high}} - N_{\mathrm{low}} + 1} \sum_{n=N_{\mathrm{low}}}^{N_{\mathrm{high}}} \left[10 \log_{10} \frac{P_i(n)}{\hat{P}_i(n)} \right]^2} \tag{19}$$

where $d_{\mathrm{LSD}}(i)$ is the LSD value of the ith frame and P_i and $\hat{P}_i$ are the FFT power spectra of the original SWB audio signals and the audio signals processed with BWE methods, respectively. N_{high} and N_{low} are the indices corresponding to the upper and lower bounds of the HF band with the frequency range of 7~14 kHz. Before computing the LSD values, all the data needed to be resampled to 32 kHz and were temporally aligned with the original SWB audio. Then, the LSD values were computed only for the HF band in the range of 7~14 kHz. The resulting LSD values were averaged over all the frames for each test signal, and the mean LSD was used as the distortion measure.

Table 3 shows the LSD values of the signals extended by the proposed and reference BWE methods. On average, the EESN method outperforms the HMM method in terms of LSD measure. For rock music, violin solo, and speech, the LSD difference between two methods is with the range of ±0.3 dB. But for jazz music, the LSD difference is relatively large, because this type of music has rich HF components and obvious transients over time. In comparison with the HF spectra extended by the HMM method, the spectra extended by the

Table 3 LSD of the signals reproduced by different BWE methods

Data type	HMM	EESN
Country music	6.6788	7.5653
Jazz music	12.6068	8.4862
Rock music	6.3262	6.4720
Violin solo	3.8470	3.5376
Symphony	3.9713	5.2903
Speech	6.1027	5.8650
Average	6.5888	6.2027

proposed EESN method are more similar to the original ones and obtain a lower LSD value. For symphony and country music, the energy of the HF components extended by the EESN method is slightly high, so the LSD value of the EESN method is higher than that of the HMM method, though the audio extended by EESN is brighter.

4.2.2 Cosh measure

In comparison to LSD measure, Itakura-Saito distortion gives more heavy weights on the peaks of the audio spectrum and is more relative to the true subjective auditory quality [26]. Itakura-Saito distortion is defined as

$$d_{\mathrm{IS}}\left(P_i,\hat{P}_i\right)=\frac{1}{N_{\mathrm{high}}-N_{\mathrm{low}}+1}\sum_{n=N_{low}}^{N_{\mathrm{high}}}\left[\frac{P_i(n)}{\hat{P}_i(n)}-\log_{10}\frac{P_i(n)}{\hat{P}_i(n)}-1\right] \tag{20}$$

But, Itakura-Saito distortion is asymmetrical as distance metric, so cosh measure is selected as the modified measurement to describe the perceptual distortion of the reproduced audio. The cosh measure [26] is defined as

$$d_{\mathrm{COSH}}(i)=\frac{1}{2}\left[d_{\mathrm{IS}}\left(P_i,\hat{P}_i\right)+d_{\mathrm{IS}}\left(\hat{P}_i,P_i\right)\right] \tag{21}$$

In this paper, the cosh measure was only computed within the frequency range of 7~14 kHz. The resulting cosh values were averaged over all the frames for each test signal, and the mean cosh value was used as the distortion measure. Table 4 gives the comparison results of the COSH measure of the signals extended by two methods. For jazz music and speech, the EESN method gains a better quality than the HMM method, but for the country music, the EESN method has more distortion since the reconstructed HF spectrum owns higher energy. On average, the SWB audio reproduced by EESN can achieve a better objective quality in terms of the cosh measure.

4.2.3 DLSD

The continuity of the audio spectrum over time is as perceptually important as the accuracy of the spectrum reconstruction. Here, DLSD is utilized as a dynamic distortion measure for evaluating the smoothness of temporal evolution for the spectral envelope of the extended audio signals [27]. If the DLSD value of the reproduced audio is small, it indicates that the spectrum evolves smoothly over time and is helpful for the overall subjective auditory quality of the reconstructed audio. DLSD is defined as

$$d_{\mathrm{DLSD}}(i)=\sqrt{\frac{1}{2\left(N_{\mathrm{high}}-N_{\mathrm{low}}+1\right)}\sum_{n=N_{\mathrm{low}}}^{N_{\mathrm{high}}}\left[10\log_{10}\frac{P_i(n)}{P_{i-1}(n)}-10\log_{10}\frac{\hat{P}_i(n)}{\hat{P}_{i-1}(n)}\right]^2} \tag{22}$$

where P_{i-1} and $\hat{P}_{i-1}$ are the FFT spectra of the original and extended SWB audio signals in the previous frame.

Table 5 shows the DLSD values of the signals extended by two BWE methods. The dynamic distortion of violin solo and symphony is similar, because their energy of the HF components is low and evolves smoothly over time. For country, jazz, and rock music, the EESN method can effectively restore transients and achieve an improvement in DLSD over the HMM method on the average. For speech, the DLSD value of EESN is about 3.89 dB. In some unvoiced frames, the HF energy of the extended signals has some difference from that of the original signals, and it causes some dynamic distortion.

4.3 Subjective preference test

The subjective quality of two BWE methods was assessed by using the pair-wise subjective preference tests. The subjective tests were arranged in a quiet room, and 15 listeners were invited to take part in the tests. Five test signals were selected from the MPEG audio dataset. The listeners were asked to choose which audio they preferred from two presented test items or to indicate no preference in each test case, and they were also allowed to repeat the testing data with no time limitation before giving answers.

Table 4 cosh measure of the signals reproduced by different BWE methods

Data type	HMM	EESN
Country music	29.3467	31.6246
Jazz music	24.4898	18.1792
Rock music	19.4905	19.4480
Violin solo	2.3990	2.3325
Symphony	3.1477	3.0678
Speech	40.2552	34.6777
Average	19.8548	18.2216

Table 5 DLSD of the signals reproduced by different BWE methods

Data type	HMM	EESN
Country music	4.4466	3.9017
Jazz music	6.7463	4.8963
Rock music	3.9707	3.2137
Violin solo	2.1191	2.0595
Symphony	3.3856	3.5123
Speech	3.9606	3.8925
Average	4.1048	3.5793

Table 6 Results of the subjective preference tests

	Prefer the former (%)	Prefer the latter (%)	No preference (%)
WB vs original SWB	6.7	76.0	17.3
EESN vs WB	46.5	25.3	28.2
HMM vs WB	44.6	29.3	26.1
EESN vs HMM	28.7	24.1	47.2
EESN vs original SWB	26.1	31.1	42.8
Original SWB vs HMM	35.1	20.9	44.0

The examples of demmostation are given in the Additional file 1.

Six groups of tests were presented to each listener. They were the comparison between the original WB and SWB signals, the comparison between the extended signals by the EESN method and the WB signals, the comparison between the extended signals by the HMM method and the WB signals, the comparison between the extended signals by the EESN method and the HMM method, the comparison between the extended signals by the EESN method and the original SWB signals, and the comparison between the original SWB signals and the extended signals by the HMM method. The results of the subjective tests are shown in Table 6. It is found that the extended signals by two BWE methods are preferable over the WB signals in terms of subjective auditory quality and the original SWB signals gain the best performance in comparison with the extended signals and the WB signals. In addition, the subjective audio quality of the proposed EESN method is slightly better with a little bit less of artifacts, compared to the HMM reference method, and is subjectively comparable to the original SWB signals. According to the comments of listeners, more dynamic contents in country, jazz, and rock music are restored by using the EESN methods, and it leads to a preference for EESN in terms of subjective listening quality, in comparison with the reference method. For symphony and violin solo, the audio spectrum changes smoothly over time, so no significant difference can be perceived between the audio signals extended by different BWE methods. In addition, some bandwidth-extended audio signals of country and jazz music own higher energy than the original SWB audio signals, so some listeners subjectively prefer the extended audio signals to the original ones.

5 Conclusions

A bandwidth extension method for audio signals based on ensemble echo state network is proposed in this paper. For each region in the feature space, a specific echo state network with recursive structure is utilized to dynamically model the mapping relationship between the LF and HF coefficients in the light of the continuous state updating equation. And, the outputs of multiple ESNs are fused by means of the GMM-based network ensemble techniques, in order to further estimate the HF spectral envelope. Subjective and objective quality test results show that the EESN method achieves improvement in terms of both static and dynamic distortions in comparison with the HMM-based reference method on the average and the auditory quality of the reproduced signals is close to that of the original SWB audio.

Additional file

Additional file 1: Demonstration for subjective listening tests. (PPT 10166 kb)

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grants 61072089 and 61471014.

Received: 29 June 2015 Accepted: 4 May 2016
Published online: 12 May 2016

References

1. P Vary, R Martin, *Digital speech transmission—enhancement, coding and error concealment* (Wiley, UK, 2006)
2. E Larsen, MR Aarts, *Audio bandwidth extension—application of psychoacoustics, signal processing and loudspeaker design* (Wiley, UK, 2004)
3. C Avendaño, H Hermansky, EA Wan, *Beyond Nyquist towards the recovery of broad bandwidth speech from narrow-bandwidth Speech* (Proc European Conference on Speech and Language Processing, Madrid, 1995), pp. 165–168
4. YM Cheng, D O'Shaughnessy, P Mermelstein, Statistical recovery of wideband speech from narrowband speech. IEEE Trans. Speech Audio Process **2**(4), 544–548 (1994)
5. H Carl, U Heute, *Bandwidth enhancement of narrow-band speech signals* (Proc 7th European Signal Processing Conference, Edinburgh, 1994), pp. 1178–1181
6. J Epps, WH Holmes, *A new technique for wideband enhancement of coded narrowband speech* (Proc IEEE Workshop on Speech Coding, Porvoo, 1999), pp. 174–176
7. IY Soon, KY Chai, *Bandwidth extension of narrowband speech using soft-decision vector quantization* (Proc Fifth International Conference on Information, Communications and Signal Processing, Bangkok, 2005), pp. 734–738
8. U Kornagel, Techniques for artificial bandwidth extension of telephone speech. Signal Process. **86**(6), 1296–1306 (2006)
9. KY Park, HS Kim, *Narrowband to wideband conversion of speech using GMM based transformation* (Proc IEEE International Conference on Acoustics Speech and Signal Processing, Istanbul, 2000), pp. 1843–1846
10. CV Botinhao, BS Carlos, LP Caloba, MR Petraglia, *Frequency extension of telephone narrowband speech signal using neural networks* (Proc IMACS Multiconference on Computational Engineering in Systems Applications (CESA), Beijing, 2006), pp. 1576–1579
11. VP Tuan, F Schaefer, G Kubin, *A novel implementation of the spectral shaping approach for artificial bandwidth extension* (Proc 3rd International Conference on Communications and Electronic, Nha Trang, 2010), pp. 262–267
12. B Iser, G Schmidt, *Neural networks versus codebooks in an application for bandwidth extension of speech signals* (Proc European Conference on Speech and Language Processing, Geneva, 2003), pp. 565–568
13. P Jax, P Vary, *Wideband extension of telephone speech using a hidden Markov model* (Proc 7th IEEE Workshop on Speech Coding, Delavan, 2000), pp. 133–135
14. P Jax, P Vary, On artificial bandwidth extension of telephone speech. Signal Process. **83**(8), 1707–1719 (2003)
15. GB Song, P Martynovich, A study of HMM-based bandwidth extension of speech signals. Signal Process. **89**(10), 2036–2044 (2009)
16. C Yagli, MAT Turan, E Erzin, Artificial bandwidth extension of spectral envelope along a Viterbi path. Speech Comm. **55**(1), 111–118 (2013)

17. X Liu, CC Bao, Blind bandwidth extension of audio signals based on non-linear prediction and hidden Markov model. APSIPA Transactions on Signal and Information Processing **3**(e8), 1–16 (2014)
18. G Lear, A state-space-based recurrent neural network for dynamic system identification. J. Syst. Eng. **6**(3), 186–193 (1996)
19. M Lukosevicius, A practical guide to applying echo state networks, in *Neural networks: tricks of the trade*, ed. by G Montavon, GB Orr, KR Müller, 2nd edn. (Springer, Heidelberg, 2012), pp. 659–686
20. M Lukosevicius, H Jaeger, Reservoir computing approaches to recurrent neural network training. Comput. Sci. Rev. **3**(3), 127–149 (2009)
21. H Jaeger, M Lukosevicius, D Popovici, U Siewert, Optimization and applications of echo state networks with leaky-integrator neuron. Neural Netw. **20**(3), 335–352 (2007)
22. H Jaeger, H Haas, Harnessing nonlinearity: predicting chaotic systems and saving energy in wireless communication. Science **304**(5667), 78–80 (2004)
23. UD Schiller, JJ Steil, Analyzing the weight dynamics of recurrent learning algorithm. Neucomputing **63**, 757–779 (2005)
24. LK Hansen, P Salamon, Neural network ensembles. IEEE Trans. Pattern Anal. Mach. Intell. **12**(10), 993–1001 (1990)
25. H Pulakka, L Laaksonen, M Vainio, J Pohjalainen, P Alku, Evaluation of an artificial speech bandwidth extension method in three languages. IEEE Trans. Audio Speech Lang. Process. **16**(6), 1124–1137 (2008)
26. AH Gray, JD Markel, Distance measures for speech processing. IEEE Trans. Audio Speech Lang. Process. **24**(5), 380–391 (1976)
27. F Norden, T Eriksson, Time evolution in LPC spectrum coding. IEEE Trans. Speech Audio Process **12**(3), 290–301 (2004)