EURASIP Journal on Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

Audio bandwidth extension based on temporal smoothing cepstral coefficients

Xin Liu and Chang-Chun Bao*

Abstract

In this paper, we propose a wideband (WB) to super-wideband audio bandwidth extension (BWE) method based on temporal smoothing cepstral coefficients (TSCC). A temporal relationship of audio signals is included into feature extraction in the bandwidth extension frontend to make the temporal evolution of the extended spectra smoother. In the bandwidth extension scheme, a Gammatone auditory filter bank is used to decompose the audio signal, and the energy of each frequency band is long-term smoothed using minima controlled recursive averaging (MCRA) in order to suppress transient components. The resulting 'steady-state' spectrum is processed by frequency weighting, and the temporal smoothing cepstral coefficients are obtained by means of the power-law loudness function and cepstral normalization. The extracted temporal smoothing cepstral coefficients are fed into a Gaussian mixture model (GMM)-based Bayesian estimator to estimate the high-frequency (HF) spectral envelope, while the fine structure is restored by spectral translation. Evaluation results show that the temporal smoothing cepstral coefficients exploit the temporal relationship of audio signals and provide higher mutual information between the low- and high-frequency parameters, without increasing the dimension of input vectors in the frontend of bandwidth extension systems. In addition, the proposed bandwidth extension method is applied into the G.729.1 wideband codec and outperforms the Mel frequency cepstral coefficient (MFCC)-based method in terms of log spectral distortion (LSD), cosh measure, and differential log spectral distortion. Further, the proposed method improves the smoothness of the reconstructed spectrum over time and also gains a good performance in the subjective listening tests.

Keywords: Audio bandwidth extension; Temporal smoothing cepstral coefficients; Minima controlled recursive averaging; Gaussian mixture model

1 Introduction

In current mobile communication systems, the effective bandwidth of the transmitted wideband (WB) audio is limited to the frequency range of 50 ~ 7,000 Hz. Due to the loss of high-frequency (HF) information, the perceived quality of WB audio is significantly degraded compared to super wideband (SWB) audio, which is band-limited to 50 ~ 14,000 Hz in frequency. To improve the auditory quality of WB audio, bandwidth extension (BWE) techniques [1,2] have been developed to artificially restore the missing HF components at receiver only from the decoded WB audio signals and to reproduce the SWB audio signals without any modifications to the existing audio encoding and transmission components.

BWE methods for audio signals can be separated into two tasks: estimation of the spectral envelope and extension of the fine spectrum. The performance of spectral envelope estimation is crucial for the auditory quality of the extended audio [2]. Therefore, research on spectral envelope estimation is attractive in the BWE area. In general, a set of time-domain and frequency-domain features is extracted from WB audio signals, and the HF spectral envelope is estimated by using some statistical learning methods, e.g., codebook mapping [3], neural networks [4], or Bayesian estimation [1], on the basis of a priori knowledge between the WB features and the HF spectral envelope. In early studies, more concern was laid on the bandwidth extension of speech signals. The relationship between the high-frequency and low-frequency (LF) parameters was quantified through mutual information, discrete entropy and separability [5-7], and the upper bounds on

* 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

the quality of bandwidth extension have been presented for both the clean and noisy speech [8,9]. Furthermore, Kleijn et al. suggested that BWE methods need to make use of perceptual properties to upgrade the subjective auditory quality of the extended signals [6]. Based on this view, Mel frequency cepstral coefficient (MFCC) has been applied into BWE methods [10] for parameterizing the WB spectrum. By using Mel-scale filters and cepstrum analysis, MFCC provides more certainty about the HF components, which is quantified as the ratio of mutual information between the HF and LF parameters to the discrete entropy of HF parameters. Then, the joint probability density function of the HF and LF feature vectors is approximated by a Gaussian mixture model (GMM), and the HF spectral envelope is estimated according to the minimum mean square error (MMSE) criterion [11,12]. The method based on MFCC and GMM can effectively reduce the spectral distortion of the extended speech compared to the method based on line spectral frequency parameters [10] and also achieves a good extension performance for audio signals.

The studies on audio perceptual quality [13-15] show that the temporal evolution of audio signals is as perceptually important as the reconstruction of spectral envelope. To some degree, artificially making the reconstructed spectra of audio signals temporally smooth is beneficial for the subjective quality and even can compensate some perceptual distortion caused by the error of the reconstructed spectra. Therefore, hidden Markov models (HMM) have been introduced into the BWE schemes to further capture the temporal relationship between the adjacent frames of audio signals [16-18]. In the HMM-based BWE methods, each state represents the characteristics of one particular sound. The state-specific probability density between the HF and LF features is approximated by GMM, and the envelope change of audio signals is modeled by means of the state transition process. Thus, HMMs contribute to a smoother temporal evolution of the reconstructed spectra which improves the listening quality. In addition, delta-MFCC features can be utilized to directly represent the difference between frames to include some temporal memory in the frontend of BWE schemes. The delta-feature-based BWE methods provide a higher certainty of the HF parameters and achieve better performance in terms of objective quality [19-21]. However, two abovementioned methods also have some drawbacks. HMM brings plenty of model parameters, and the storage space and computational complexity has to increase for model training and Bayesian estimation. Similarly, employing the delta-MFCC objectively adds the dimension of the input vectors for the statistical models and increases the burden on computational complexity.

1.1 Paper overview

In this paper, we attempt to embed the extraction of the temporal relationship of audio signals into cepstral coefficients and propose a novel temporal smoothing cepstral coefficient (TSCC)-based scheme for BWE of audio signals. It can improve the temporal smoothness of the extended HF spectrum, without increasing the dimension of input feature and without burdening the storage space and computational complexity for enhanced/more sophisticated statistical models. Firstly, a Gammatone filter bank is adopted to decompose the audio signals, and the audio energy of each frequency band is long-term smoothed by means of minima controlled recursive averaging (MCRA) to suppress transient signal components. Secondly, the resulting 'steady-state' spectrum is processed by frequency weighting, and TSCCs are extracted by means of the power-law loudness function and cepstral normalization. Finally, the extracted cepstral coefficients are applied into a GMM-based BWE scheme to restore the HF components. The proposed method suppresses the transient components existing in the WB audio signal in the BWE frontend and provides higher mutual information between the LF and HF parameters. Informal listening tests show that, for most audio signals, TSCC plays an important role on the improvement of spectral distortion and the temporal smoothness of the reconstructed audio signals, while for some rock music with the accompaniment of strong percussion music, temporal envelope modification needs to be applied in order to maintain a good extension performance.

In the next section, the TSCC extraction is described in detail, and then the application of the proposed feature to the GMM-based BWE scheme is briefly discussed. Section 3 gives the analysis of mutual information between the HF and LF parameters processed by different parameterizations. In Section 4, the proposed method and the reference BWE methods are evaluated in terms of objective quality measurements and subjective listening tests. Finally, conclusions are drawn in Section 5.

2 Bandwidth extension based on temporal smoothing cepstral coefficients

A block diagram of the proposed BWE method is shown in Figure 1. The input signal is the WB audio signals sampled at 16 kHz with a bandwidth of 7 kHz and is separated into 40-ms frames with 20-ms overlap. Then, the new TSCC features are extracted from the audio frame and are fed into a GMM-based Bayesian estimator to estimate the HF spectral envelope. Here, the HF spectral envelope $M_{\mathrm{HF}}(r)$, r =0, 1, 2, 3 is represented by the root mean square (RMS) values of four sub-bands in the HF band. These sub-bands are distributed on the perceptually correlated equivalent rectangular bandwidth (ERB) scale [22] without overlapping, and their central

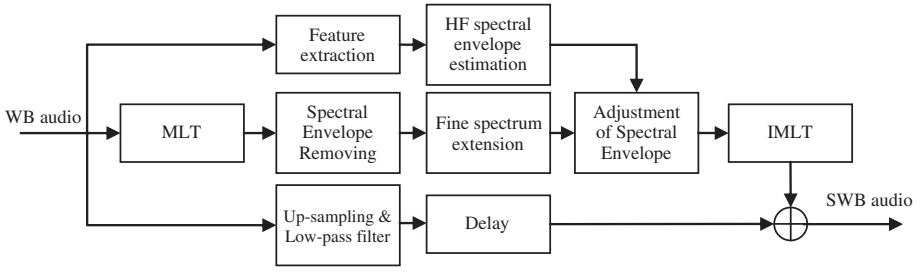

Figure 1 Block diagram of the proposed BWE method.

frequencies are arranged at 8,470, 9,338, 11,653, and 13,657 Hz.

In addition, the fine spectrum is extended by using the traditional spectral translation method [23,24]. Modulated lapped transform (MLT) [25] is performed to convert the input audio signals into the frequency domain, and 320 MLT coefficients $C_{mlt}(i)$, i =0,...,319 are obtained for describing the spectrum below 8 kHz. The fine spectrum of WB audio is obtained through the spectral envelope normalization. In order to keep the normalized spectrum flat, the LF spectral envelope is described as the RMS values of 14 sub-bands which are uniformly spaced in the range of 0 ~ 7 kHz,

$$M_{\mathrm{LF}}(r) = \sqrt{\frac{1}{h(r)-l(r)+1}\sum_{i=l(r)}^{h(r)} {C_{\mathrm{mlt}}}^2(i)}, 0 \le r < 14 \tag{1}$$

where $C_{mlt}(i)$, i =0,...,279 is the MLT coefficients in the range of 0 ~ 7 kHz, and $l(r)$ and $h(r)$ correspond to the lower and upper boundaries of the MLT coefficients in the rth sub-band. Thus, the LF normalized MLT coefficients $C_{norm\_mlt}(i)$, i =0,...,279 are given as,

$$C_{\mathrm{norm\_mlt}}(i) = \frac{C_{\mathrm{mlt}}(i)}{M_{\mathrm{LF}}(r)}, r = \left\lfloor \frac{i}{20} \right\rfloor \tag{2}$$

Next, the fine spectrum in the frequency band of 0 ~ 7 kHz is directly translated into the range of 7 ~ 14 kHz, and the HF spectral envelope is adjusted according to the estimated HF sub-band RMS value $\hat{M}_{HF}(r)$ [26]. Through inverse modulated lapped transform (IMLT), the HF components are transformed into the time domain. Finally, the original WB signals are up-sampled and low-pass filtered. The resulting signals are delayed and combined with the artificially restored HF signals to reconstruct the extended SWB audio. The total algorithmic delay of the proposed method is 40.75 ms, which is caused mainly by the time-frequency transform, low-pass filter, and filter bank used in feature extraction. A more detailed description of the algorithm is provided below.

2.1 Extraction of temporal smoothing cepstral coefficients

For the stationary audio signals, the HF spectrum changes at a relatively slow rate over time in comparison with the LF spectrum. By using the traditional BWE method, the envelope of the extended HF spectrum for the stationary audio signals is affected by the transient components contained in the LF spectrum, and some sudden changes of energy for the HF spectrum may occur between frames and degrade the auditory quality of the stationary audio signals according to the informal listening tests. Inspired by the studies on audio perceptual quality [13-15] that the temporal smoothness of audio spectrum is partially beneficial for the subjective quality, we attempt to eliminate the impact of transients from the features in order to improve the temporal smoothness of the extended HF spectrum. In the proposed method, the 'steady-state' spectrum of audio signals is first obtained by using MCRA for describing the spectrum components which are slowly evolved over time. Then, the resulting 'steady-state' spectrum is further used to remove the transients from the original spectrum of audio signals by the spectral weighting method. Finally, the cepstral features are computed from the weighted spectrum. Though some transients in the HF spectrum may be not well reconstructed by using the proposed method, the quality of the extended audio signals is partially improved on the whole according to the objective and subjective tests shown in Section 4.

Figure 2 shows the extraction principle of the TSCC feature proposed in this paper. Firstly, a Gammatone filter bank is used to decompose the input audio signals into 20 channels, whose central frequencies are non-uniformly distributed in the frequency range of 50 ~ 7,000 Hz [27]. As a standard model of cochlear filtering, the Gammatone filter bank can model the nonlinear response and frequency selectivity of basilar membranes in human ears by means of low-order causal filters with few parameters, and its impulse response function is represented as,

$$g(f,t) = \begin{cases} t^{a-1}e^{-2\pi bt}\cos(2\pi f t), & t \ge 0 \\ 0, & t < 0 \end{cases} \tag{3}$$

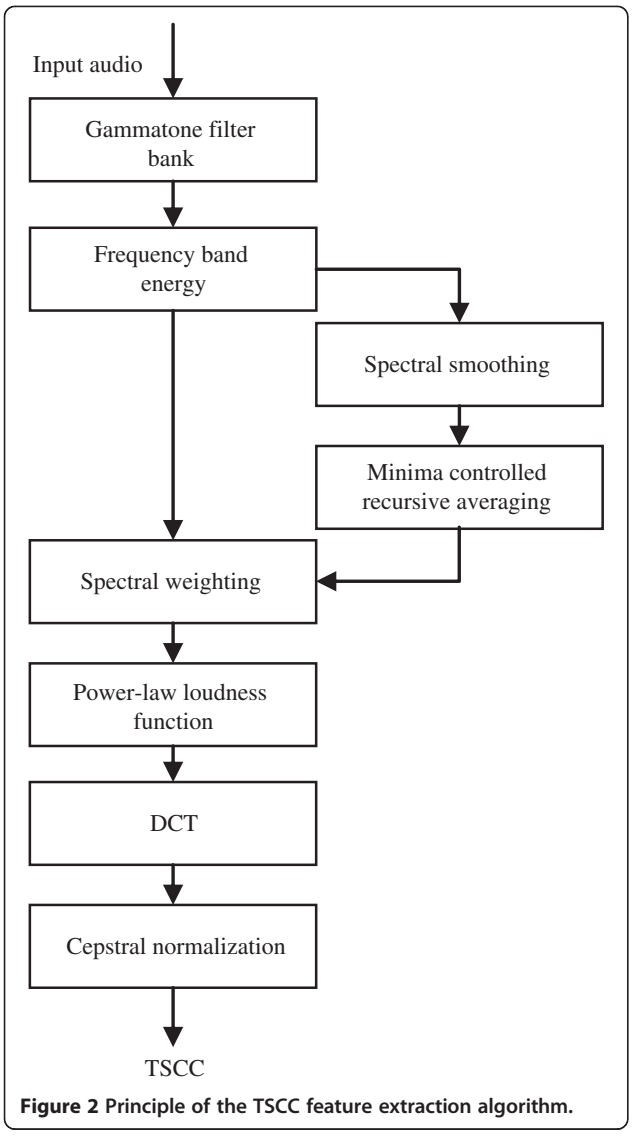

Figure 2 Principle of the TSCC feature extraction algorithm.

where t and f correspond to the time samples and central frequencies of the Gammatone filter in each channel, separately. In order to make the Gammatone filters better-matched to the filtering characteristics of basilar membranes, the filter order a is set to 4, the central frequencies of 20 Gammatone filters are spaced on the ERB scale, and the rectangular bandwidth b of each channel increases with the central frequency f.

The WB audio signal $x(i)$ is fed into the jth Gammatone filter, where i is the sample index. The resulting signal, $x_g(i,j)$, j =0,…,19 is down-sampled to the frame rate of 50 Hz along the time dimension to obtain the energy spectrum $S_f(m,j)$ in the current frame, where m is the frame index.

In order to verify the effects of the proposed cepstral coefficients, a segment of jazz music signal with a length of 11 s is used as an example. The resulting Gammatone spectrogram is shown in Figure 3. It can be clearly found that some transient components appear in the spectrum of the original audio signal and damage the continuity of each frequency band over time. This may degrade the extension performance for the audio signals.

2.1.1 Minima controlled recursive averaging

In the proposed method, MCRA [28] is first employed to extract the 'steady-state' components from the audio energy spectrum, and the adverse effects of BWE caused by the transient components are restrained in the frontend. Dependent on the presence 'probability' of the rapidly evolving components in each channel, the time smoothing factor is adaptively adjusted to estimate the 'steady-state' audio spectrum at the current frame. The detailed procedure of MCRA is as follows.

1) Temporal smoothing

The energy spectrum in the jth channel $S_f(m,j)$ at the current frame is temporally smoothed. The smoothed energy spectrum $S(m,j)$ is given as,

$$S(m,j) = \alpha_s S(m-1,j) + (1-\alpha_s) S_f(m,j) \tag{4}$$

where m is the frame index, $S(m-1,j)$ is the smoothed energy spectrum at the previous frame, and α_s =0.7 is the smoothing factor.

2) Asymmetric filtering

The spectrum of the transient components evolves more rapidly than the 'steady-state' components, so the asymmetric filtering is adopted to obtain the slowly evolving components in the jth channel $S_{slow}(m,j)$ as,

$$S_{\text{slow}}(m,j) = \begin{cases} \gamma S_{\text{slow}}(m-1,j) + \frac{(1-\gamma)}{(1-\beta)}(S(m,j)-\beta S(m-1,j)), & \text{if } S_{\text{slow}}(m-1,j) < S(m,j) \\ S(m,j), & \text{if } S_{\text{slow}}(m-1,j) \geq S(m,j) \end{cases} \tag{5}$$

where m and j are the frame index and the index of frequency bands. γ =0.92 denotes the smoothing factor, and a 'look-ahead' factor β =0.7 is introduced in the asymmetric filtering to adjust the adaptation time of the algorithm [29]. Through asymmetric filtering, the resulting $S_{slow}(m,j)$ approaches the lower envelope of the smoothed energy spectrum $S(m,j)$ to describe the slowly evolving components of the audio spectrum in the jth channel, and the rapidly evolving components are presented by $S(m,j) - S_{slow}(m,j)$. Figure 4 shows the slowly evolving components of the jazz music signal shown in Figure 3.

3) Calculating the presence 'probability' of the rapidly evolving components

The ratio of the smoothed energy spectrum $S(m,j)$ to the slowly evolving components $S_{slow}(m,j)$ can be expressed as,

$$R(m,j) = \frac{S(m,j)}{S_{\text{slow}}(m,j)} \tag{6}$$

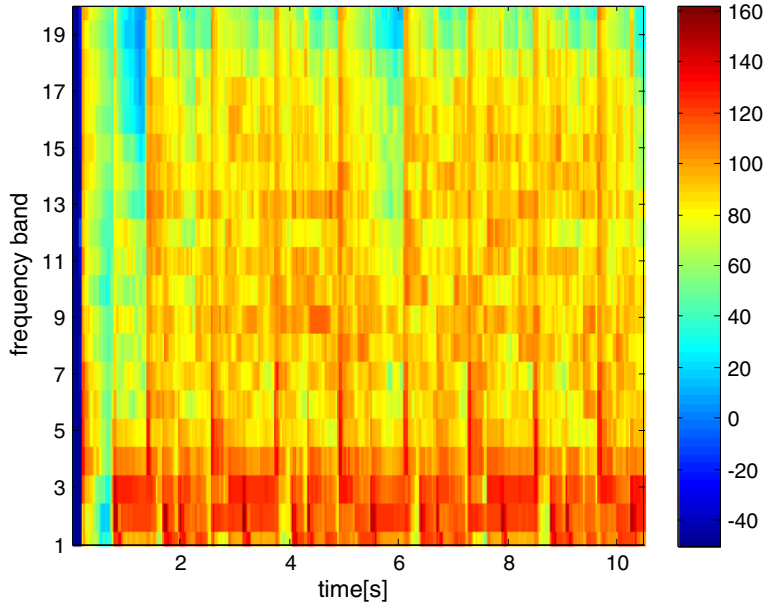

Figure 3 The resulting Gammatone spectrogram for a jazz music signal.

By using the hard threshold method, $R(m,j)$ can be exploited to determine whether the rapidly evolving components are present in the jth channel at the mth frame. The presence 'probability' of the rapidly evolving components is obtained as,

$$p(m,j) = \begin{cases} 1, if & R(m,j) > \delta \\ 0, if & R(m,j) \le \delta \end{cases} \quad (7)$$

where δ =3, and if $R(m,j) > \delta$, then we assume that undesired rapidly evolving components are present, i.e., $p(m,j)$ =1, otherwise we assume that the slowly evolving components are dominant, i.e., $p(m,j)$ =0. Through recursive averaging, the presence 'probability' of the rapidly evolving components $p'(m,j)$ which are continuously evolved over frames can be further given as,

$$p'(m,j) = \alpha_p p'(m-1,j) + (1-\alpha_p)p(m,j) \quad (8)$$

where the smoothing factor is α_p =0.2.

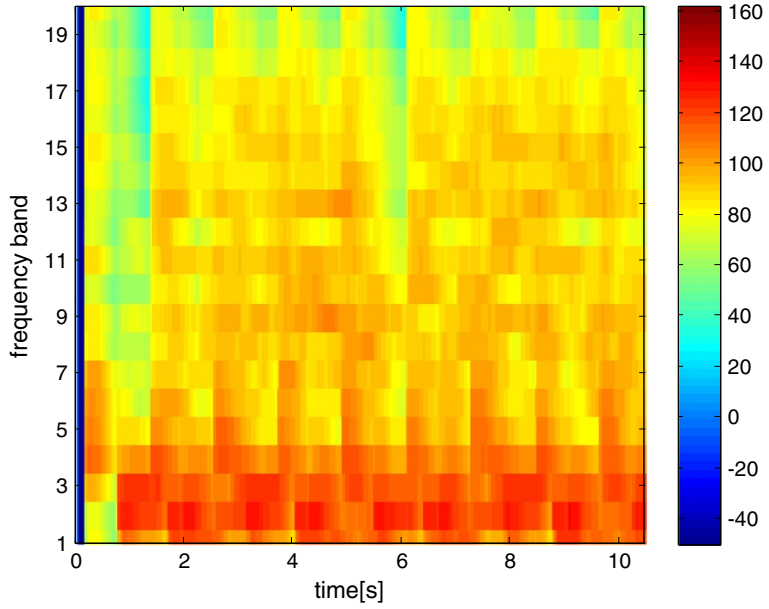

Figure 4 The spectrogram of the slowly evolving components for a jazz music signal.

4) Achieving the 'steady-state' spectrum of audio signals

The so-called 'steady-state' spectrum in the *j*th channel at the current frame $S_{steady}(m,j)$ can be estimated by recursive averaging as,

$$S_{\text{steady}}(m,j) = \alpha_d(m,j)S_{\text{steady}}(m-1,j) + (1-\alpha_d(m,j))S_f(m,j) \tag{9}$$

where $S_{steady}(m\text{-}1,j)$ is the 'steady-state' spectrum at the previous frame. The 'steady-state' smoothing factor $\alpha_d(m,j)$ is dynamically adjusted according to the presence 'probability' of the rapidly evolving components $p'(m,j)$ and is updated as,

$$\alpha_d(m,j) = \alpha + (1-\alpha)p'(m,j) \tag{10}$$

where α =0.85.

Figure 5 illustrates the 'steady-state' spectrum of the jazz music signal shown in Figure 3. It is obvious that, compared with the original spectrogram shown in Figure 3, the transient components are effectively restrained, but the fine structure of the audio spectrum is also blurred owing to long-term smoothing.

2.1.2 Spectral weighting

As shown in Figure 5, the 'steady-state' spectrum of audio signals is able to describe the spectrum components which are slowly evolved over time, but the fine structure of audio spectrum is blurred. If the 'steady-state' spectrum was directly adopted for extracting the cepstral coefficients, the mutual dependencies between the LF and HF parameters were obviously decreased according to the experimental observation, because much information contained in the fine spectrum was missing. In order to make a balance between temporal smoothness of the extended spectrum and mutual dependencies between the LF and HF parameters, the resulting 'steady-state' spectrum is further used to remove the transients from the original spectrum $S_f(m,j)$ of audio signals by the spectral weighting method.

First, the ratio of the 'steady-state' spectrum to the smoothed spectrum of audio signals is determined for each channel at the current frame. Next, the resulting ratio is multi-point averaged over frequency for weighting the audio spectrum. Then, the weighted spectrum can be computed as follows,

$$S_w(m,j) = w(m,j)S_f(m,j) \tag{11}$$

$$w(m,j) = \frac{1}{h(j)-l(j)+1}\sum_{k=l(j)}^{h(j)} \frac{S_{\text{steady}}(m,k)}{S(m,k)} \tag{12}$$

where $h(j) = \max(j+2, 19)$ and $l(j) = \min(j-2, 0)$.

Figure 6 shows an example for the weighted spectrum of the jazz music signal given in Figure 3. It can clearly be seen that spectral weighting decreases the weight value of transient components and retains the 'steady-state' components in the original audio spectrum to upgrade the smoothness of the processed audio spectrum over both time and frequency. Compared with the 'steady-state' spectrum shown in Figure 5, spectral weighting does not only repress the transient components but also preserves more fine structure which is beneficial to describe the time-frequency characteristics of audio signals.

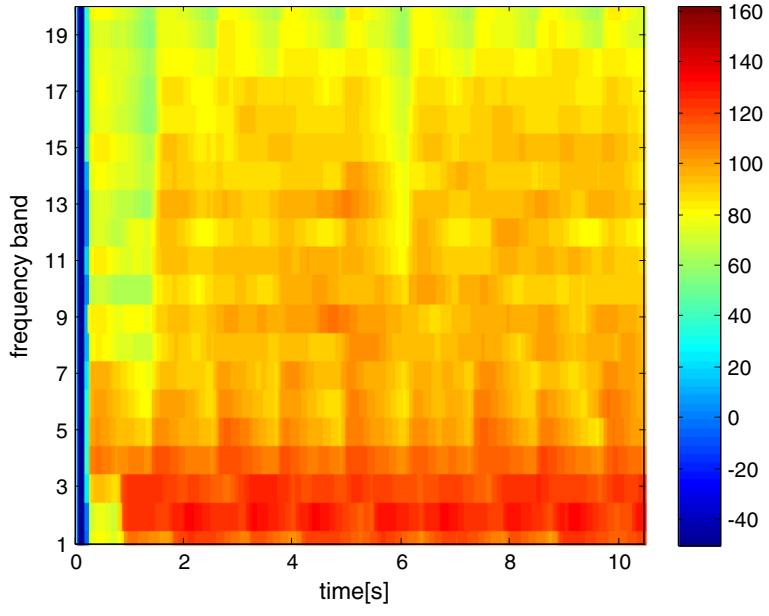

Figure 5 The 'steady-state' spectrum for a jazz music signal.

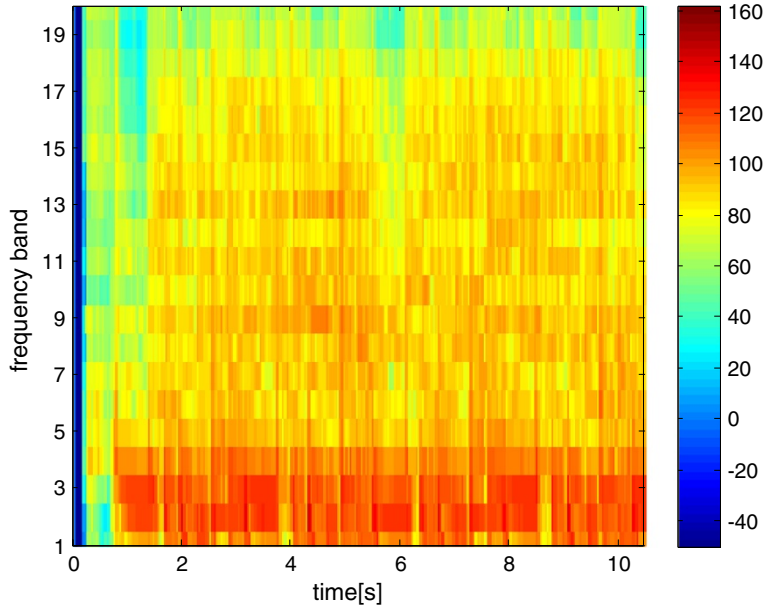

Figure 6 The weighted spectrum for a jazz music signal.

2.1.3 Extraction of cepstral coefficients

In consideration of the BWE performance, the cubic root loudness function [30] is selected to approximate the nonlinear relationship between the intensity of sound and its perceived loudness and can be expressed as,

$$S_c(m,j) = \sqrt[3]{S_w(m,j)} \tag{13}$$

Next, discrete cosine transform (DCT) is applied to de-correlate the coefficients, and the resulting cepstral coefficients are referred to as TSCCs,

$$C(m,j) = \sqrt{\frac{2}{N}}\sum_{i=0}^{19} S_c(m,i)\cos\left(\frac{j\pi}{2N}(2i+1)\right) \tag{14}$$

where j =0,...,19 is the TSCC index.

In order to compress the dynamic range, interval normalization is applied to the TSCCs. Since the silence frames in audio signals affect the value of TSCC, the pre-trained maximum and minimum values of the cepstral coefficients are searched only in the non-silence frames of the training data and are fixed in practical application. The normalized TSCC $C_{norm}(m,j)$ can be obtained as,

$$C_{\mathrm{norm}}(m,j) = \frac{C(m,j) - C_{\min}(j)}{C_{\max}(j) - C_{\min}(j)} \tag{15}$$

where $C_{max}(j)$ and $C_{min}(j)$ correspond to the pre-trained maximum and the minimum of TSCC in the jth channel.

2.2 HF spectral envelope estimator based on Gaussian mixture model

A 20-dimensional TSCC vector is extracted as the WB audio feature vector $F_X = [C_{norm}(m,0),C_{norm}(m,1),...,C_{norm}(m,19)]^{\mathrm{T}}$ and is fed into the GMM-based Bayesian estimator [11] to estimate the RMS values of the HF sub-bands $F_Y = [M_{\mathrm{HF}}(0),M_{\mathrm{HF}}(1),...,M_{\mathrm{HF}}(3)]^{\mathrm{T}}$ under the MMSE criterion. The joint vector of the HF and LF vectors is referred to as $F_Z = [F_X^{\mathrm{T}}F_Y^{\mathrm{T}}]^{\mathrm{T}}$, and a GMM is utilized to model the joint probability density function $p(F_Z)$ as,

$$p(F_{\mathrm{Z}}) = \sum_{i=1}^{N_{\mathrm{G}}} w_i N(F_{\mathrm{Z}};\mu_i,C_i) \tag{16}$$

where N_G is the number of Gaussian mixtures, and $N(.)$ is the multi-variate Gaussian probability density function. w_i, μ_i, and C_i represent the weight, mean vector, and covariance matrix of the ith Gaussian components, respectively. These model parameters were off-line trained from the SWB audio dataset of the live concert recordings with the length of about 1 h. During the training phase, the parallel WB audio signals were generated by low-pass filtering, down-sampling, and time alignment. Then, the level of WB and SWB signals was normalized to −26 dBov, and the WB feature vectors F_X and the HF sub-band RSM values F_Y were extracted from the parallel dataset. Finally, the model parameters of GMM were estimated under the maximum likelihood criterion by using the standard expectation-maximization algorithm.

Given the joint density distribution function $p(F_Z)$, we use the MMSE-based Bayesian estimator and compute the conditional expectation $E[F_Y|F_X]$ of F_Y given the WB

feature vector F_X as the estimated value of HF spectral envelope $\hat{F}_Y$ as follows,

$$\begin{aligned}\hat{F}_Y &= \underset{\hat{F}_Y}{\operatorname{argmin}} \int \|F_{\mathrm{Y}}-\hat{F}_{\mathrm{Y}}\|^2 p(F_{\mathrm{Y}}|F_{\mathrm{X}})dF_{\mathrm{Y}} = E[F_{\mathrm{Y}}|F_{\mathrm{X}}] \\ &= \sum_{i=1}^{N_{\mathrm{G}}} p(i|F_{\mathrm{X}})E[F_{\mathrm{Y}}|i,F_{\mathrm{X}}]\end{aligned} \tag{17}$$

where $p(i|F_X)$ is the posterior probability of the ith Gaussian component given F_X,

$$p(i|F_{\mathrm{X}}) = \frac{w_i p(F_{\mathrm{X}}|i)}{\sum_{j=1}^{N_{\mathrm{G}}} w_j p(F_{\mathrm{X}}|j)} \tag{18}$$

$p(F_X|i)$ is the observed probability of F_X in the ith Gaussian component. It can be computed from the joint probability of F_Z in the ith Gaussian component $p(F_Z|i)$. $E[F_Y|i,F_X]$ is the conditional expectation of F_Y given F_X in the ith Gaussian component and is given as,

$$E[F_{\mathrm{Y}}|i,F_{\mathrm{X}}] = \mu_{\mathrm{Y},i} + C_{\mathrm{YX},i}{C_{\mathrm{XX},i}}^{-1}\left(F_{\mathrm{X}}-\mu_{\mathrm{X},i}\right) \tag{19}$$

where $\mu_{X,i}$ and $\mu_{Y,i}$ are the mean vector of F_X and F_Y in the ith Gaussian component, $C_{XY,i}$ is the cross-correlation matrix, and $C_{XX,i}$ is the correlation matrix of F_X.

2.3 High-frequency component synthesis

In this paper, the fine spectrum of audio signals in the frequency range of 0 ~ 7 kHz is described as the normalized LF MLT coefficients $C_{norm\_mlt}(i)$, i =0,...,279. According to spectral translation [2], the normalized MLT coefficients in the LF band are directly copied to the HF band of audio spectrum to extend the fine spectrum. On the basis of the ERB scale, the HF fine spectrum is also divided into four non-overlapping sub-bands and is multiplied by the estimated RMS values of the HF sub-bands to reproduce the extended HF spectrum. At last, the HF spectrum is converted into the time domain through IMLT. The original WB audio signals are up-sampled, low-pass filtered, and temporally aligned, and then it is combined with the waveform of the HF components to reconstruct the extended SWB audio signals.

3 Mutual information analysis

The BWE performance depends on all mutual dependencies between the HF and LF parameters. On the basis of the information theory, all the linear and nonlinear dependencies between the LF feature vectors F_X and the HF spectral envelope F_Y can be described by their mutual information (MI) $I(F_X;F_Y)$. The larger the MI, the lower is the upper bound on the minimum achievable mean square error of the HF spectral envelope estimated from the LF features [8]. For this reason, the effects of mutual information caused by different WB audio features are studied in this paper.

The marginal probabilities $p_X(F_X)$, $p_Y(F_Y)$ and the joint probability density $p_{XY}(F_X,F_Y)$ of the WB audio features and the HF spectral envelope coefficients are modeled by a GMM with 128 mixtures and with full covariance matrices. The mutual information measure $I(F_X;F_Y)$ can be approximated in the light of the numerical integration method, in order to evaluate the dependencies between the HF and LF parameters as follows,

$$\begin{aligned}I(F_{\mathrm{X}};F_{\mathrm{Y}}) &= E\left[\log_2\left(\frac{p_{\mathrm{XY}}(F_{\mathrm{X}},F_{\mathrm{Y}})}{p_{\mathrm{X}}(F_X)p_Y(F_{\mathrm{Y}})}\right)\right] \\ &\approx \frac{1}{M}\sum_{i=0}^{M=1}\log_2\left(\frac{p_{\mathrm{XY}}(F_{\mathrm{X}}(m),F_{\mathrm{Y}}(m))}{p_{\mathrm{X}}(F_{\mathrm{X}}(m))p_{\mathrm{Y}}(F_{\mathrm{Y}}(m))}\right)\end{aligned} \tag{20}$$

where m is the frame index, M is the frame number. $F_X(m)$ and $F_Y(m)$ are referred to as the LF and HF feature vectors at the mth frame.

In this paper, the traditional MFCC [10] and Gammatone frequency cepstral coefficients (GFCC) [30] are adopted as the reference WB audio feature vectors to analyze the mutual information between the HF and LF parameters processed by different parameterizations, in comparison to the proposed TSCC. For the MFCC parameterization, Fourier transform is adopted to obtain the magnitude spectrum of audio signals, and Mel-scale triangular filters are applied to compute the log energy of 20 non-uniform sub-bands in the frequency range of 50 ~ 7,000 Hz. DCT is applied to the log energies to decorrelate the coefficients and to yield the cepstral coefficients. GFCC bring more auditory properties on the basis of MFCC. A Gammatone auditory filter bank takes the place of the Mel-scale triangular filter bank to mimic the properties of the basilar membrane in cochlear, and the log nonlinearity is replaced by cubic root loudness function for cepstral analysis. The proposed TSCC further adds the minima controlled recursive averaging-based long-term smoothing and spectral weighting to enhance the temporal smoothness of audio spectrum. In order to reduce the dynamic range, all the three versions of cepstral coefficients employ the same cepstral normalization methods. In addition, the dimension of the cepstral coefficients needs to be reduced by removing the trailing dimensions for MI analysis with different feature dimensions, as shown in Table 1.

It is shown that, for all the 20 cepstral coefficients, GFCC gains the MI improvement of 0.1 bit/frame in comparison with MFCC, and TSCC achieves a 1.4 bits/frame higher MI. Exploiting the Gammatone filter bank and loudness function is helpful for improving the MI

Table 1 Comparison of MI between the LF features and the HF spectral envelope

MFCC		GFCC		TSCC	
Dimension	**MI**	**Dimension**	**MI**	**Dimension**	**MI**
20	3.8537	20	3.9717	20	5.2553
18	3.4005	18	3.6244	18	4.6849
16	3.2117	16	3.3796	16	4.2116
14	3.0344	14	3.1797	14	3.7367
12	2.7832	12	2.9214	12	3.4479
10	2.6181	10	2.6686	10	3.1966
5	2.3627	5	2.3977	5	2.5690

between the HF and LF parameters. Long-term smoothing and spectral weighting are beneficial to the smoothness of audio spectrum and retain more MI which is positive about the extension performance without increasing the input feature dimension in the frontend of BWE. In addition, with the decreasing dimension of features, the MI measures of the three versions of cepstral coefficients are gradually reduced. Here, the MI about GFCC gradually approaches to that of traditional MFCC and TSCC still shows the superiority over the reference cepstral coefficients. When the dimension is reduced to 5, the MI value of the three versions of cepstral coefficients is down to 2.6 bits/frame.

Additionally, the selection of the smoothing factors in the TSCC extraction process, such as α_s, β, γ, δ, α_p, and α, may also affect the extension performance of the proposed BWE scheme. So, we experimented with several values of the smoothing factors and empirically optimized them under the MI measure. First, the smoothing factor α_s is analyzed when β, γ, δ, α_p, and α are respectively set to the fixed values, 0.7, 0.92, 3, 0.2, and 0.85, and the MI measure of the TSCC-based BWE method with different α_s is shown in Figure 7. The result demonstrates that the highest MI measure is obtained when α_s =0.7 and the value of the MI measure is rapidly diminished to about 4.9 bits/frame with the decrease of α_s. When α_s >0.7, the MI measure is marginally worse because the excessively smoothed spectrum may blur the fine structure and degrade the differentiation of the extracted cepstral coefficients. Also, we analyze the effect of the parameters for asymmetric filtering on the extension performance in the light of the MI measure with the fixed smoothing factors α_s =0.7, δ =3, α_p =0.2, and α =0.85. Figure 8 shows the MI measure between the WB audio features and HF spectral envelope coefficients as a function of the 'look-ahead' factor β with different γ. For a fixed γ, the value of β which ranges from 0.6 to 0.8 gives a higher MI, and when β >0.8 the MI measure would also decrease because a small value of β makes the audio spectrum excessively smooth and reduces the mutual

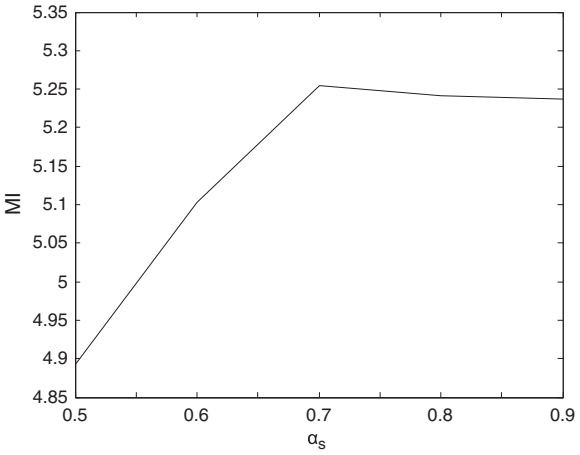

Figure 7 MI measure of the TSCC-based BWE method with different α_s.

dependencies between the LF and HF parameters. Besides, γ =0.92 provides a higher MI value in comparison with other values of γ. The threshold value δ is used to determine whether the rapidly evolving components are present in one channel. Figure 9 shows the MI measure of the TSCC-based BWE method with different threshold values δ. When δ =3, the MI measure reaches 5.2553 bits/frame. When the value of δ is larger or smaller, the MI measure would also decrease because the rapidly evolving components could not be differentiated from the slowly evolving components. Similarly, α_p and α are analyzed when other parameters are set to the fixed values. Figures 10 and 11 show that when α_p =0.2 and α =0.85, the maximum MI measure could be achieved. After all, according to the results of MI analysis, the smoothing factors could be experimentally determined as α_s =0.7, β =0.7, γ =0.92, δ =3, α_p =0.2, and α =0.85.

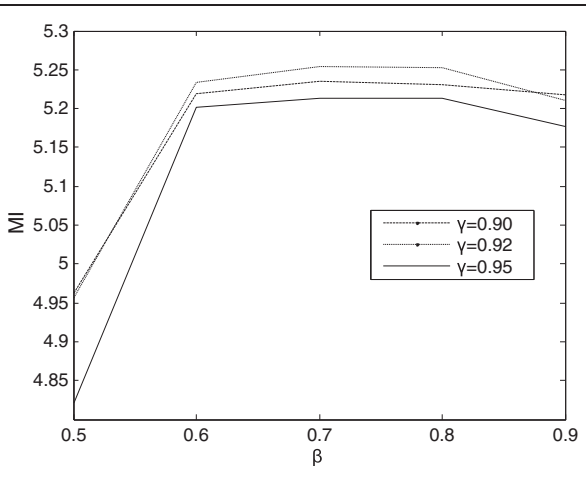

Figure 8 MI measure of the TSCC-based BWE method with different β and γ.

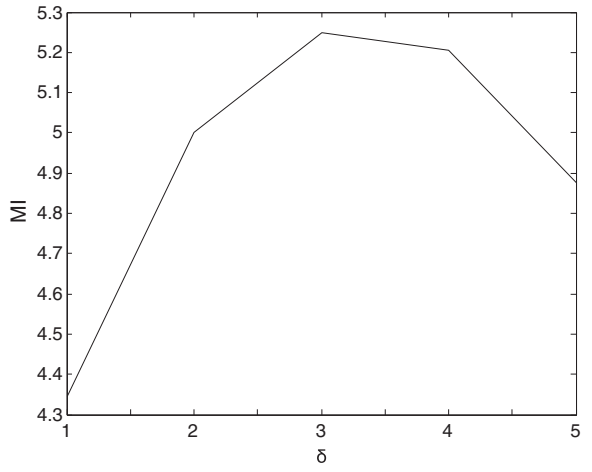

Figure 9 MI measure of the TSCC-based BWE method with different *δ*.

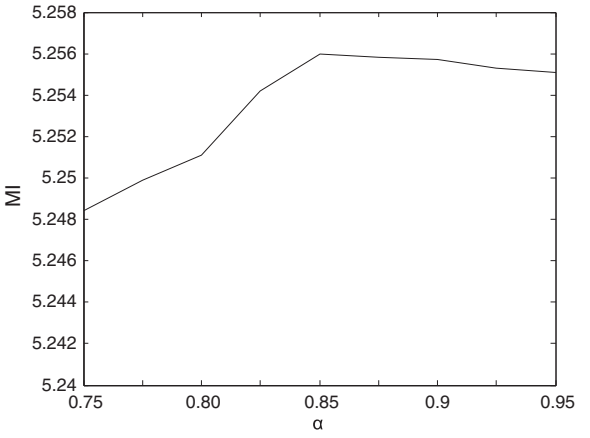

Figure 11 MI measure of the TSCC-based BWE method with different *α*.

4 Performance evaluation and test results

The goal of this work is to enhance the auditory quality of WB audio, without any side information for describing the missing HF components. In this section, we apply the proposed method and the MFCC-based BWE method to extend the bandwidth of the WB audio reproduced by the G.729.1 WB codec [26,31], in order to make a direct comparison with the SWB extension method with side information which is employed in G.729.1 Annex E [32] in terms of the objective and subjective quality tests. In addition, we make a subjective comparison between the original WB audio signals and the audio signals extended by the proposed method for further evaluating the improvement of the proposed method on the auditory quality.

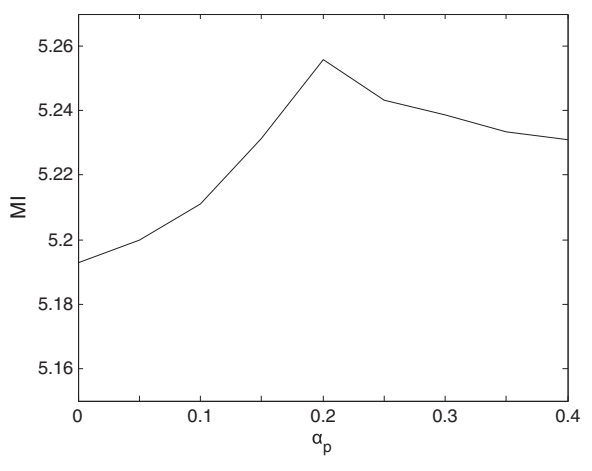

Figure 10 MI measure of the TSCC-based BWE method with different α_p.

4.1 Training

In the proposed method, the joint probability density between the features computed from the WB audio and the HF spectral envelope are modeled as a GMM. The training data comes from the lossless audio data of live concert recordings with the length of about 1 h. It contains different types of dialogues, music, singing, and live background sound. The audio signals were digitally stored by using 16-bit PCM with the sampling frequency of 32 kHz and the bandwidth of 14 kHz. The parallel WB audio signals were generated by low-pass filtering, down-sampling, and time alignment. The cepstral features extracted from the WB signals were used as the 20-dimensional input features F_X. The 4-dimensional HF spectral envelope F_Y was computed from the training samples in the parallel SWB database. The probability density of the joint vector $F_Z = \{F_X, F_Y\}$ were modeled by a GMM. The model parameters were trained using the standard expectation-maximization algorithm. The model has 128 mixtures and full covariance matrices. According to an informal listening test, no evident difference can be perceived with increase of the number of Gaussian mixtures.

4.2 Test data and reference methods

Fifteen SWB audio signals were selected from the MPEG auditory quality test database and were not part of the dataset for GMM training. The test signals contained pop music, instrumental solo music, symphonic pieces, and speech. Each signal had a length between 10 and 20 s. They were low-pass filtered with the cutoff frequency of 7 kHz and down-sampled to WB audio. Before further processing, the level of these WB audio signals [33] was required to be normalized to −26 dBov.

The test signals were processed with the G.729.1 WB codec at 32 kbit/s as WB references. SWB references

were produced with the G.729.1 Annex E at 36 kbit/s. The core layer of G.729.1 Annex E with the bit rate of 32 kbit/s is fully identical to ITU-T G.729.1 to reproduce the WB audio signals, and the SWB extension layer with the additional bit rate of 4 kbit/s adopts a two-mode BWE method [34] in the modified discrete cosine transform (MDCT) domain to extend the bandwidth of the reproduced audio to 14 kHz. The mode selection is done by estimating the tonality of the input audio signals. For non-tonal signals, the 'generic mode' reconstructs the HF components through transposing the LF components with a gain adjustment. For tonal signals, the 'sinusoidal mode' adds a finite set of MDCT peaks to the HF spectrum. In addition, post-processing in time domain is used to improve the synthesized SWB audio at the decoder of G.729.1 Annex E.

Test items for BWE were obtained by applying the TSCC-based method and MFCC-based method to the G.729.1-coded WB signals. TSCC or MFCC is directly extracted from the G.729.1-coded WB signals and is fed into the GMM-based Bayesian estimator for obtaining the HF spectral envelope. Moreover, the reproduced WB signals are transformed into frequency domain via MLT, and the fine structure of the LF-MLT spectrum is translated into the HF band. After the energy adjustment, the regenerated HF components are converted into time domain by inverse MLT, to further form SWB audio signals with the WB audio decoded by G.729.1. In the interest of fairness, the central frequencies of the triangular filters adopted by MFCC should be the same as the central frequencies of the Gammatone filter bank in TSCC, and the resulting cepstral coefficients are also normalized to compress the dynamic range as shown in Equation 15.

4.3 Objective evaluation

Audio signals generated using the proposed TSCC-based method, MFCC-based method, G.729.1 at 32 kbit/s, and G.729.1 Annex E at 36 kbit/s were objectively evaluated in terms of the log spectral distortion (LSD) [35,36], cosh measure [20,35], and differential log spectral distortion (DLSD) [37] in comparison with the original SWB audio.

4.3.1 Log spectral distortion

The LSD [35,36] between the audio signals processed by different methods and the original SWB audio in the HF band of 7 ~ 14 kHz was computed directly from the FFT power spectra as,

$$d_{\mathrm{LSD}}(i) = \sqrt{\frac{1}{N_{\mathrm{high}} - N_{\mathrm{low}} + 1} \sum_{n=N_{\mathrm{low}}}^{N_{\mathrm{high}}} \left[10 \log_{10} \frac{P_i(n)}{\hat{P}_i(n)} \right]^2} \tag{21}$$

where $d_{\mathrm{LSD}}(i)$ is the LSD value at 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 different methods, respectively. N_{high} and N_{low} are the indices corresponding to the upper and lower bound of the HF band with the frequency range of 7 ~ 14 kHz. Before computing the LSD values, all the date needed to be resampled to 32 kHz and were temporally aligned with the original SWB audio. Especially for the up-sampled WB signals decoded by G.729.1, the spectral components above 7 kHz need to be removed by low filtering. Then, the LSD values were computed only for the HF band with 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. For different types of test signals, the LSD measures of the reconstructed audio processed by different methods are shown in Table 2.

As shown in Table 2, G.729.1 only reproduces the WB audio signals and provides an LSD value of 11.3182 dB on average, and the LSD improvement can be achieved through all the BWE methods due to extending the bandwidth to 14 kHz. The reproduced SWB audio signals based on the proposed TSCC achieves an LSD improvement of about 1.2 dB in comparison with the traditional MFCC. Since the additional bit rate of 4 kbit/s is adopted to describe the real HF spectrum, G.729.1 Annex E can reconstruct the HF spectrum more accurately than the BWE methods without side information and further decreases the LSD value down to 7.4005 dB. Additionally, different types of audio signals show different extension

Table 2 LSD measures of the reconstructed audio based on different methods

Data type	TSCC-BWE	MFCC-BWE	G.729.1 Annex E 36 kbit/s	G.729.1 32 kbit/s
Country music	8.6300	10.1043	7.7542	11.8330
Jazz music	8.5145	10.6131	7.9227	12.2260
Rock music	9.4281	11.7701	8.1582	13.7525
Violin solo	6.1353	6.2611	5.9824	6.6624
Symphony	7.7741	8.3748	7.0836	9.0386
Speech	9.7205	10.1199	7.5021	14.3964
Average	8.3671	9.5406	7.4005	11.3182

Table 3 Cosh measures of the reconstructed audio based on different methods

Data type	TSCC-BWE	MFCC-BWE	G.729.1 Annex E 36 kbit/s	G.729.1 32 kbit/s
Country music	19.0821	44.1792	10.7046	183.6361
Jazz music	35.7912	118.3688	28.0020	227.3057
Rock music	57.7031	126.2545	12.9660	333.6178
Violin solo	7.0643	7.5069	6.9839	69.5002
Symphony	14.0920	24.5220	13.7074	51.7750
Speech	103.4225	176.1067	15.5153	377.6968
Average	39.5259	82.8230	14.6465	207.2553

performance in terms of LSD. For jazz music and rock music, the audio signals are accompanied with strong percussion and contain abundant transient component. The TSCC-based method suppresses the transient components of audio signals through MCRA and makes the spectrum of reconstructed audio temporally smooth. So, the TSCC-based method obtains the 2-dB LSD improvement in comparison with the MFCC-based method. In addition, the spectrum of symphony and violin solo presents few transients and maintains the continuity over time, thus the difference between the TSCC-based method and the MFCC-based method is not significant in terms of LSD. For speech, the results of LSD measure show that the TSCC-based method achieves a 0.4-dB improvement and the spectral distortion usually occurs in the starting frames of the unvoiced sound and the transition between the voiced and unvoiced frames.

4.3.2 Cosh measure

In contrast with LSD, the Itakura-Saito distortion provides more heavy weights on the peaks of an audio spectrum and is closely related to the subjective listening quality. It is defined as,

$$d_{\mathrm{IS}}\left(P_i,\hat{P}_i\right)=\frac{1}{N_{\mathrm{high}}-N_{\mathrm{low}}+1}\sum_{n=N_{\mathrm{low}}}^{N_{\mathrm{high}}}\left[\frac{P_i(n)}{\hat{P}_i(n)}-\log_{10}\frac{P_i(n)}{\hat{P}_i(n)}-1\right] \tag{22}$$

Because the Itakura-Saito distortion is not symmetric as distance metric, the cosh measure [35] is used as the modified measure to describe the perceptual distortion of the reconstructed audio signals. Thus, we further introduced the cosh measure to evaluate the performance of the reconstructed audio signals and could obtain more subjectively correlated results [20]. The cosh measure is defined as,

$$d_{\mathrm{COSH}}(i)=\frac{1}{2}\left[d_{\mathrm{IS}}\left(P_i,\hat{P}_i\right)+d_{\mathrm{IS}}\left(\hat{P}_i,P_i\right)\right] \tag{23}$$

The cosh measure was computed only for the HF band and was also averaged over all the frames for test signals to obtain the final cosh measure. For different types of test signals, the cosh measures of the reconstructed audio processed by different methods are shown in Table 3.

As shown in Table 3, since the HF energy is large, the cosh measure of jazz music and rock music is a little higher. The unvoiced speech has abundant HF components, so its cosh measure is also large. For violin solo, the LF components are rich, and the energy of the HF spectrum is faint with the increase of frequency, thus its cosh measure value is lower and the difference between the BWE methods is not significant. Overall, the SWB audio reproduced by G.729.1 Annex E can achieve the best objective quality for all the types of audio. The cosh measure of G.729.1 is the largest because the HF spectrum is truncated. Compared with MFCC, the proposed cepstral coefficients also gain a better extension performance in terms of the subjectively correlated cosh measure, and the mean cosh measure is much closer to that of G.729.1 Annex E.

Table 4 DLSD measures of the reconstructed audio based on different BWE methods

Data type	TSCC-BWE	MFCC-BWE	G.729.1 Annex E 36 kbit/s	G.729.1 32 kbit/s
Country music	5.3129	6.0940	3.8121	7.1232
Jazz music	5.0191	6.5015	5.0955	7.4515
Rock music	5.7910	6.9800	5.1260	8.2231
Violin solo	4.3011	4.7076	3.8624	4.8863
Symphony	4.9078	5.0013	4.5332	5.5112
Speech	5.7345	6.0280	4.7101	8.0015
Average	5.1777	5.8854	4.5232	6.8661

4.3.3 Differential log spectral distortion

The smoothness of the audio spectrum between frames is important for the perceptual quality as well as the accurate reconstruction of the audio spectrum. Thus, DLSD is defined to evaluate how smoothly the spectral envelope of the extended audio signals is temporally evolved [37]. If the DLSD value of the reconstructed audio is low, the spectrum evolves smoothly over time. It is helpful for the overall subjective auditory quality of the reconstructed audio. DLSD is defined as,

$$d_{\mathrm{DLSD}}(i) = \sqrt{\frac{1}{2\left(N_{\mathrm{high}}-N_{\mathrm{low}}+1\right)}\sum_{n=N_{\mathrm{low}}}^{N_{\mathrm{high}}}\left[10\log_{10}\frac{P_i(n)}{P_{i-1}(n)}-10\log_{10}\frac{\hat{P}_i(n)}{\hat{P}_{i-1}(n)}\right]^2} \quad (24)$$

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

The DLSD values of the reconstructed SWB audio based on different BWE methods are shown in Table 4. Similar to the result of LSD and cosh measure, the transient components in country music, jazz music, and rock music increase the DLSD values of the reconstructed audio signals. For the temporally smooth symphony signals, different BWE methods have minor differences in terms of DLSD. For speech, the improvement of the proposed method in DLSD is not remarkable, because spectral distortion occurring in the unvoiced frames has a negative influence on the spectral smoothness of the extended audio. On the average, the proposed TSCC improves the smoothness of temporal evolution of the reconstructed audio spectrum, gains the DLSD improvement of 0.7 dB in comparison with MFCC, and is inferior to G.729.1 Annex E which partly provides the real HF spectrum.

4.4 Subjective listening tests

A CCR listening test which is similar to the subjective assessment method recommended by ITU-T P.800 [38] was used to pair-wise evaluate the differences of audio quality for the proposed TSCC-based method, MFCC-based method, and G.729.1 Annex E. In each test case, two differently processed versions of the same test signals were presented to the listeners. Listeners used the following seven-point comparison mean opinion scores (CMOS) to judge the quality of the second audio sample relative to that of the first: 3, much better; 2, better; 1, slightly better; 0, the same; −1, slightly worse; −2, worse; −3 much worse.

Fifteen male and five female listeners took part in the CCR test, and the age range was from 22 to 30 years old. The test was arranged in the quiet room, and only the test attendee was present in the room during the test. Six test signals including pop music, guitar, sax, drums, and speech were selected at random from the MPEG database, and the level of the original test signals and the processed signals was normalized to −26 dBov. The differently processed types of the six MPEG testing signals were played to both ears through AKG K271

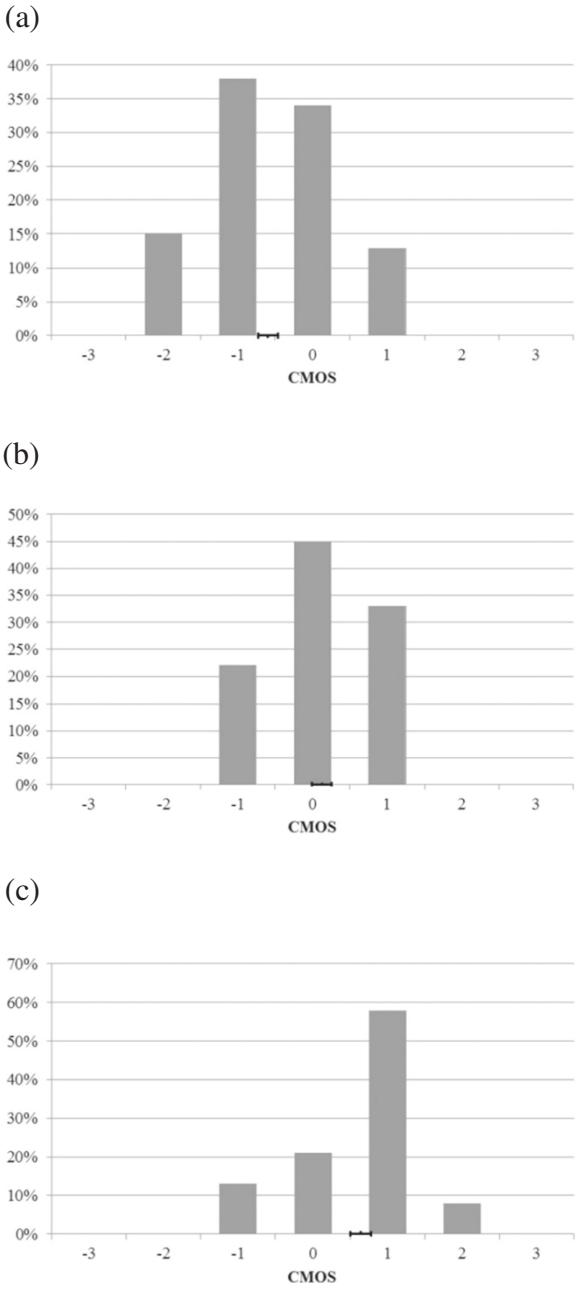

Figure 12 Distributions of listener rating in CCR tests. (a) Comparison between the TSCC-based method and the MFCC-based method. **(b)** Comparison between the TSCC-based method and G.729.1 Annex E. **(c)** Comparison between the MFCC-based method and G.729.1 Annex E.

MKII headphones for listening tests. Each listener had a short practice before actual tests to adjust the volume setting to a suitable level and was allowed to repeat each pair of testing data with no time limitation before giving their answers. After listeners gave their grade, their assessment of the quality for each pair of testing data was recorded in a previously prepared form, and the comments on the test signals for the listeners were collected and analyzed by the experimenters.

Three groups of tests were presented to each listener. They are the comparison between the TSCC-based method and MFCC-based method, comparison between the TSCC-based method and G.729.1 Annex E, and comparison between the MFCC-based method and G.729.1 Annex E. The distributions of listener rating for the test is shown in Figure 12. The bars indicate the relative frequencies of the scores given in the comparisons between the two processing methods. Bars on the positive side show preference for the latter method. The mean score for each group of tests is also shown on the horizontal axis with the 95% confidence interval. As shown in Figure 12, the proposed TSCC-based method and G.729.1 Annex E were considered substantially better than the MFCC-based method in terms of subjective listening tests, and the proposed method showed a little impairment compared to G.729.1 Annex E. According to the comments of the listeners, the proposed method is able to enhance the quality of WB audio signals decoded by G.729.1 and achieved a good performance on the whole, especially for the country music and jazz music. For speech signals, the proposed method effectively extends the bandwidth and improves the naturalness of the G.729.1-coded signals, but some spectral distortion occurred in the unvoiced sound may be perceived and slightly degrade the subjective quality of the audio signals extended by the TSCC-based method. For some rock music with the accompaniment of strong percussion signals, the HF transients of the reconstructed audio are not well restored due to the smoothing operations of TSCC for the WB audio. It might degrade the auditory quality in comparison with the SWB audio signals reproduced by G.729.1 Annex E which could partly restore the HF spectrum by using side information. So, the temporal envelope modification method could be further introduced in the future work in order to maintain a good extension performance for the percussion music signals.

Additionally, we made a comparison between the original WB audio signals and the extended audio signals in terms of the CCR test for further evaluating the performance of the proposed method. Here, the original WB audio signals were referred to as the WB reference, and the proposed TSCC-based method was adopted to extend the bandwidth of the original WB signals. The distribution of listener ratings for the test is shown in Figure 13. The proposed TSCC-based method achieved a better subjective quality than the WB audio signals on the whole. From the comments of the listeners, the quality of audio signals is effectively improved by using the BWE method, though some distortion can be perceived in the frames where the energy of the HF band is relatively high.

4.5 Computational complexity

We made a comparison of computational complexity between the proposed method and the MFCC-based method. The computational complexity was approximately measured in the number of multiplications per frame. Because the main modules of two BWE methods were the same except for feature extraction, only the computational complexity of feature extraction was compared. For computational requirements, the proposed TSCC needs about 53,380 multiplications per frame, and MFCC costs about

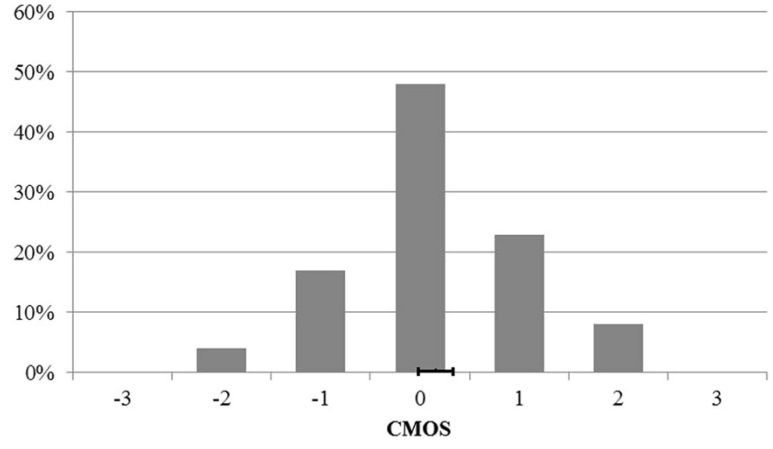

Figure 13 Distributions of listener rating in CCR tests. Comparison between the original WB audio and the audio signals extended by the TSCC-based method.

52,480 multiplications. For TSCC, the major contributions come from the Gammatone filter bank and DCT, and MCRA and spectral weighting only increase a very minor amount of computational complexity in comparison with MFCC.

5 Conclusions

This paper presents a novel BWE method based on TSCC. The audio signals are decomposed by using a Gammatone filter bank, and the audio spectrum is smoothed through MCRA and spectral weighting to suppress the transient components. Combining with the power-law nonlinear loudness function and cepstral normalization, TSCC is computed and then is applied into the GMM-based baseline BWE system for extending the bandwidth of the WB audio signals decoded by G.729.1. The extension performance of the proposed method is evaluated in comparison with the MFCC-based method and SWB extension method with side information in G.729.1 Annex E. Evaluation results show that the TSCC-based BWE method objectively gains the improvement to the MFCC-based method in terms of LSD, cosh measure, and DLSD but is inferior to G.729.1 Annex E. The subjective test results also indicate that the proposed method effectively enhances the auditory quality of the WB audio, makes the temporal evolution of the reconstructed audio signals smoother, and outperforms the MFCC-based reference method.

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, 60872027, and 61471014.

Received: 4 May 2014 Accepted: 8 October 2014
Published online: 25 November 2014

References

1. P Vary, R Martin, *Digital Speech Transmission -Enhancement, Coding and Error Concealment* (John Wiley & Sons Ltd, UK, 2006)
2. E Larsen, RM Aarts, *Audio Bandwidth Extension-Application of Psychoacoustics, Signal Processing and Loudspeaker Design* (John Wiley & Sons Ltd, UK, 2004)
3. Y Qian, P Kabal, *Wideband Speech Recovery from Narrowband Speech using Classified Codebook Mapping, in Proceedings of the 9th Australian International Conference on Speech Science & Technology* (Australian Speech Science & Technology Association Inc., Melbourne, Australia, 2002)
4. H Pulakka, P Alku, Bandwidth extension of telephone speech using a neural network and a filter bank implementation for highband Mel spectrum. IEEE Trans. Audio Speech Lang. Process. **19**(7), 2170–2183 (2011)
5. P Jax, P Vary, *Feature Selection for Improved Bandwidth Extension of Speech Signals, in Proceedings of the International Conference on Acoustics, Speech, and Signal Processing* (IEEE, Montreal, Quebec, 2004)
6. M Nilsson, H Gustafsson, SV Andersen, WB Kleijn, *Gaussian Mixture Model Based Mutual Information Estimation between Frequency Bands in Speech, in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (IEEE, Orlando, FL, USA, 2002)
7. M Nilsson, SV Andersen, WB Kleijn, *On the Mutual Information between Frequency Bands in Speech, in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (IEEE, Istanbul, Turkey, 2000)
8. P Jax, P Vary, *An Upper Bound on the Quality of Artificial Bandwidth Extension of Narrowband Speech Signals, in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (IEEE, Orlando, FL, USA, 2002)
9. T Esch, P Vary, *An Information Theoretic View on Artificial Bandwidth Extension in Noisy Environments, in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (IEEE, Kyoto, Japan, 2012)
10. AH Nour-Eldin, P Kabal, *Mel-frequency Cepstral Ccoefficient-based Bandwidth Extension of Narrowband Speech, in Proceedings of the INTERSPEECH (ISCA* (Brisbane, Australia, 2008)
11. KY Park, HS Kim, *Narrowband to Wideband Conversion of Speech using GMM Based Transformation, in Proceedings of the IEEE International Conference on Acoustics* (Speech and Signal Processing (IEEE, Istanbul, Turkey, 2000)
12. H Pulakka, U Remes, K Palomaki, M Kurimo, P Alku, *Speech Bandwidth Extension using Gaussian Mixture Model-based Estimation of the Highband Mel Spectrum, in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (Prague, Czech Republic, IEEE, 2011)
13. HP Knagenhjelm, WB Kleijn, *Spectral Dynamics is more Important than Spectral Distortion, in Proceedings of the IEEE International Conference on Acoustics* (Speech and Signal Processing (IEEE, Detroit, Michigan, USA, 1995)
14. F Norden, T Eriksson, *A Speech Spectrum Distortion measure with Interframe Memory, in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (IEEE, Salt Lake City, Utah, USA, 2011)
15. B Geiser, P Vary, *Beyond Wideband Telephony - Bandwidth Extension for Super-wideband Speech, in Proceedings of the German Annual Conference on Acoustics (DAGA)* (German Society for Audiology, Dresden, Germany, 2008)
16. P Jax, P Vary, *Wideband Extension of Telephone Speech using a Hidden Markov model, in Proceedings of the IEEE Workshop on Speech Coding* (IEEE, Delavan, WI, USA, 2000)
17. C Yağlı, T TTuran, E Erzin, Artificial Bandwidth Extension of Spectral Envelope along a Viterbi Path. Speech Comm. **55**(1), 111–118 (2013)
18. GB Song, P Martynovich, A study of HMM-based bandwidth extension of speech signals. Signal Process. **89**(10), 2036–2044 (2009)
19. AH Nour-Eldin, P Kabal, *Memory-based Approximation of the Gaussian Mixture Model Framework for Bandwidth Extension of Narrowband Speech, in Proceedings of the Interspeech (ISCA* (Florence, Italy, 2011)
20. AH Nour-Eldin, P Kabal, *Combining Frontend-based Memory with MFCC Features for Bandwidth Extension of Narrowband Speech, in Proceedings of the IEEE International Conference on Acoustics* (Speech and Signal (IEEE), Taipei, Taiwan, 2009)
21. AH Nour-Eldin, TZ Shabestary, P Kabal, *The Effect of Memory Inclusion on Mutual Information between Speech Frequency Bands, in Proceedings of the IEEE International Conference on Acoustics* (Speech and Signal Processing (IEEE, Toulouse, France, 2006)
22. BCJ Moore, BR Glasberg, Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. J. Acoust. Soc. Am. **74**, 750–753 (1983)
23. J Makhoul, M Berouti, *High-frequency Regeneration in Speech Coding Systems, in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (IEEE, Washington, DC, USA, 1979)
24. U Chong, D Magill, The residual-excited linear prediction vocoder with transmission rate below 9.6 kbit/s. IEEE Trans. Comm **23**(12), 1466–1474 (1975)
25. S Shlien, The modulated lapped transform, its time-varying forms, and its applications to audio coding standards. IEEE Trans. Speech Audio Process. **5**(4), 359–366 (1997)
26. International Telecommunication Union, *ITU-T Recommendation G.729.1, G.729-based Embedded Variable Bit-Rate Coder: An 8–32 kbit/s Scalable Wideband Coder Bitstream Interoperable with G.729* (International Telecommunication Union, Geneva, 2006)
27. M Slaney, Auditory toolbox version 2. Interval Research Corporation. Tech. Rep. **10**, 1–52 (1998)
28. I Cohen, B Berdugo, Noise estimation by minima controlled recursive averaging for robust speech enhancement. IEEE Signal Process. Lett. **9**(1), 101–112 (2002)
29. G Doblinger, *Computationally Efficient Speech Enhancement by Spectral , Minima Tracking in Subbands, in Proceedings of EUROSPEECH* (ISCA, Madrid, Spain, 1995)
30. Y Shao, DL Wang, *Robust Speaker Identification using Auditory Features and Computational Auditory Scene Analysis, in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (IEEE, Las Vegas, USA, 2008)

31. S Ragot, B Kovesi, R Trilling, D Virette, N Duc, D Massaloux, S Proust, B Geiser, M Gartner, S Schandl, H Taddei, Y Gao, E Shlomot, H Ehara, K Yoshida, T Vaillancourt, R Salami, MS Lee, DY Kim, *ITU-T G.729.1: an 8–32 kbit/s Scalable Coder Interoperable with G.729 for Wideband Telephony and Voice over IP, in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* ((IEEE), Honolulu, HI, 2007)
32. International Telecommunication Union, *ITU-T Recommendation G.729.1 Amendment 6, New Annex E on Superwideband Scalable Extension* ((International Telecommunication Union, Geneva, 2010)
33. International Telecommunication Union, *ITU-T Recommendation P.56, Objective Measurement of Active Speech Level* ((International Telecommunication Union, Geneva, 2011)
34. M Tammi, L Laaksonen, A Rämö, H Toukomaa, *Scalable Superwideband Extension for Wideband Coding, in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (IEEE, Taiwan, 2009)
35. AH Gray, JD Markel, Distance measures for speech processing. IEEE Trans. Audio Speech Lang. Process. **24**(5), 380–391 (1976)
36. 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)
37. F Norden, T Eriksson, Time evolution in LPC spectrum coding. IEEE Trans. Speech Audio Process. **12**(3), 290–301 (2004)
38. International Telecommunication Union, *ITU-T Recommendation P.800, Methods for Subjective Determination of Transmission Quality* ((International Telecommunication Union, Geneva, 1996)

doi:10.1186/s13636-014-0041-6
Cite this article as: Liu and Bao: **Audio bandwidth extension based on temporal smoothing cepstral coefficients.** *EURASIP Journal on Audio, Speech, and Music Processing* 2014 **2014**:41.