EFFECTIVE HIGH FREQUENCY REGENERATION BASED ON SINUSOIDAL MODELING FOR MPEG-4 HE-AAC

*Sang-Uk Ryu and Kenneth Rose**

Electrical and Computer Engineering
University of California, Santa Barbara
CA 93106-9560, USA
{sang,rose}@ece.ucsb.edu

Joon-Hyuk Chang

Imaging Media Research Center
Korea Institute of Science and Techonology
Seoul 136-791, Korea
changjh@hi.snu.ac.kr

ABSTRACT

A novel approach is proposed for effective high frequency regeneration in audio coding, which is based on a sinusoids plus noise model. It assumes a standard high efficiency advanced audio coding (HE-AAC) encoder, and modifies the decoder to exploit all available information in estimating the model parameters. From the lower band reconstruction of core AAC, frequency parameters of the high band sinusoids are estimated. Side information about spectral energy and the regenerated high band of standard HE-AAC are employed in estimating the magnitude parameters of the high band sinusoids as well as noise model parameters. The gains achieved by the proposed technique, over conventional HE-AAC, are demonstrated by subjective quality tests that were carried out on audio signals with significant harmonics in the high band.

1. INTRODUCTION

Spectral band replication (SBR) is widely recognized as a useful approach for high frequency regeneration (HFR) in audio and speech coding. Specifically, it reconstructs the high frequency bands of the signal by replication of the lower band components followed by effective normalization of the signal energy and additional correction of sinusoids and noise components. One of the successful applications of the approach is in MPEG's high efficiency advanced audio coding (HE-AAC) [1, 2]. The main advantage of this technique is at low bit rates where it provides higher quality audio by reconstructing the high-band components from their low-band counterparts, at the cost of only a small amount of side information [3].

Despite widespread agreement on HE-AAC's effectiveness with a variety of audio signals, some patterns of quality degradation were reported in [4]. It was described in terms of tonal artifacts at patch boundaries and superfluous replication of uncorrelated lower band components. In order to mitigate such quality degradation observed in HE-AAC's reconstruction, the sinusoidal model-based HFR (SM-HFR) was proposed. At the high frequency bands, a "sinusoids plus noise" model was employed as an efficient mechanism to estimate high frequency components of signals that possess significant harmonic structure. Model parameters were estimated from side information provided by conventional HE-AAC encoder and ancillary data on the fundamental frequency. Substantial performance improvement over conventional HE-AAC (in terms of subjective quality) was observed.

In this paper, an alternative approach for the SM-HFR is proposed, where model parameters for high frequency components are estimated at the decoder without the aid of any encoder modification. First, the fundamental frequency (and hence the frequency parameters of the high band harmonics) is estimated from the decoder reconstruction of core AAC. Second, the model parameters for the noise component are approximated with the noise energy of the high band estimate of HE-AAC which, combined with the spectral energy in the side information of HE-AAC, enable estimation of the magnitudes of the high band harmonics. Through a number of subjective tests, it is found that the proposed approach employing the SM-HFR is effective in improving the reconstruction quality by regenerating the high band signal with the spectral characteristics close to the original.

2. SPECTRAL BAND REPLICATION IN HE-AAC

In this section, we provide a brief overview of SBR in the context of HE-AAC [2, 5]. The time-domain input signal $x[n]$ is applied to an M-channel complex-valued analysis quadrature-mirror filter (QMF) bank. The output from the analysis filter bank is stored in the matrix, $X(k,l)$, $0 \leq k < M$, $0 \leq l < M/2$. Note that the high frequency band for regeneration at the HE-AAC decoder can be specified by the filter bank output $X(k,l)$ with frequency range $k_x \leq k < M$, where the cutoff k_x is the first QMF subband in the high band. A small amount of side information

* This work is supported in part by the NSF under grant no. EIA-0080134, the University of California MICRO Program, Applied Signal Technology, Inc., Dolby Laboratories, Inc., and Qualcomm, Inc.

is extracted to enable HFR at the decoder, which consists of a coarse spectral energy envelope for the high band and parameters for controlling the tonal to noise ratio (TNR) of the reconstructed high band (to match the original signal). The side information is extracted on the basis of the signal-dependent frequency-time grid. If the time axis is divided into T time intervals with boundaries $\{l_t\}_{t=0}^{T}$, and the frequency axis is divided into P frequency bands bounded by $\{k_p\}_{p=0}^{P}$, then the average energy in the (p,t) frequency-time bin is calculated as

$$\bar{E}(p,t) = \frac{1}{\Delta} \sum_{k=k_{p-1}}^{k_p-1} \sum_{l=l_{t-1}}^{l_t-1} |X(k,l)|^2, \quad (1)$$

where $\Delta = (k_p - k_{p-1})(l_t - l_{t-1})$. To minimize the side information bit rate, only coarse spectral energy information is transmitted, namely, the average energy per frequency-time bin.

At the decoder, the output from the core AAC decoder is analyzed with an $M/2$-channel QMF bank, obtaining the lower band signal. The high band is generated by patching consecutive QMF subbands from the lower band with a possible repetition of the cycle, depending on the size of the high band. The regenerated high band signals are subsequently inverse filtered. Then, the gain of the inverse filtered high band signal is adjusted and additional correction in terms of noise or sinusoids is performed.

Although, it has been reported that the technique generally provides a perceptually satisfying estimate of the missing high band [6], reconstruction quality degradation was subsequently observed, especially for audio signals having a strong harmonic structure in the high band. First, when consecutive QMF channels in the lower band are copied into the high band repeatedly, the translated harmonics are often located elsewhere than at the proper integer multiples of the fundamental frequency. Spectral lines and/or holes are observed, particularly at the patch boundaries. In this case, two strong tonal signals that are close in frequency may generate an audible tonal artifact. Second, in the case that the harmonic series in the lower band are superimposed with low frequency background sounds, unwanted lower band sounds are replicated into the high band. Subsequently, the generated high band may suffer from quality degradation.

3. SINUSOIDAL MODEL-BASED HIGH FREQUENCY REGENERATION

In general, the HFR problem can be restated as: estimate the missing high frequency bands of the signal given the lower band component and a small amount of side information, so as to minimize perceptual dissimilarity with the original. In [4], we proposed an HFR technique based on a sinusoidal model for the purpose of eliminating degradation introduced by HE-AAC in the case of audio signals having strong harmonic series in the high band.

In this section, the general derivation of the proposed SM-HFR technique is presented for the class of audio signals that exhibit strong harmonic series in the high band. Let us first assume that, at the $l\text{-}th$ time instance of QMF subsampling, the high band signal $x_H^l[n]$ can be well approximated by a sum of sinusoids,

$$x_H^l[n] \cong \sum_{i=1}^{L} A_i^l \cos(\omega_i^l n + \phi_i^l), \quad (2)$$

where L is the number of sinusoids, and A_i^l, ω_i^l and ϕ_i^l denote their respective magnitudes, frequencies and phases. Since the input signal is modeled as a sum of sinusoids, the output of the QMF bank at the $l\text{-}th$ time instance, $X_H(k,l)$ can be analytically calculated in terms of sinusoidal parameters as follows:

$$X_H(k,l) = \sum_{n=0}^{N-1} h_k[n] x_H^l[Ml-n] = \sum_{i=1}^{L} A_i^l q_{k,i}^l, \quad (3)$$

where $h_k[n]$ is the $k\text{-}th$ channel analysis filter. Note that $q_{k,i}^l$ can be determined by the frequency and phase of the $i\text{-}th$ sinusoid and calculated as:

$$q_{k,i}^l = \frac{1}{2}\Big\{ H(\omega_i^l - \psi_k) \exp\Big(j(\psi_k/2 + \omega_i^l Ml + \phi_i^l)\Big) + H^*(\omega_i^l + \psi_k) \exp\Big(j(\psi_k/2 - \omega_i^l Ml - \phi_i^l)\Big)\Big\}, \quad (4)$$

where $\psi_k = \pi(2k+1)/2M$, and $H(\omega)$ is the Fourier transform of the prototype low pass filter $h[n]$ of order N. Let us define a matrix $\mathbf{Q}^l = \{q_{k,i}^l\}$ for $k_x \le k < M$, $1 \le i \le L$. The filter bank output and sinusoidal magnitudes at time instance l may be rewritten in vector form,

$$\mathbf{X}_H^l = [X_H(k_x,l) \;\; X_H(k_x+1,l) \;\cdots\; X_H(M-1,l)]^T,$$
$$\mathbf{a}^l = [A_1^l \;\; A_2^l \;\cdots\; A_L^l]^T,$$

thereby allowing (3) to be written compactly as $\mathbf{X}_H^l = \mathbf{Q}^l \mathbf{a}^l$.

It follows from (3) and (4) that once we know the sinusoidal model parameters at a given time instance, the QMF domain representation can be analytically calculated. The SM-HFR problem is thus simply equivalent to the parameter estimation problem of the underlying sinusoidal model. In the case of audio signals having strong harmonic series in the high band, the frequency parameters $\{\omega_i^l\}$ can be estimated as integer multiples of the fundamental frequency. In the harmonic speech coding literature, a synthetic phase model has been introduced to provide phase continuity at frame boundaries [7]. We adopt this model to estimate the phase parameters. From the frequency and phase estimates, we can determine the matrix $\mathbf{Q}^l$. Now, the only remaining parameters to estimate are the magnitude parameters. The sinusoidal magnitudes, which are perceptually important, can be estimated from the side information on the spectral energy envelope. For simplicity of presentation, let us assume *for now* that the energy of a subband signal in the high

band, $|X_H(k,l)|^2$ is available for each k and l. Let the k-th row vector in matrix $\mathbf{Q}^l$ be denoted by $\mathbf{q}_k^l = [q_{k,1}^l \; q_{k,2}^l \; \cdots \; q_{k,L}^l]$. Then, the energy of the k^{th} channel is computed as

$$|X_H(k,l)|^2 = |\mathbf{q}_k^l \mathbf{a}^l|^2 = (\mathbf{a}^l)^T \left\{ (\mathbf{q}_k^l)^H \mathbf{q}_k^l \right\} \mathbf{a}^l = (\mathbf{a}^l)^T \mathbf{Q}_k^l \mathbf{a}^l,$$

where we introduced the matrix $\mathbf{Q}_k^l \equiv (\mathbf{q}_k^l)^H \mathbf{q}_k^l$, and where superscript "H" stands for conjugate transposition. The final step in estimating the magnitude parameters of the sinusoidal model employs least squares fitting to identify the optimal magnitude vector:

$$\hat{\mathbf{a}}^l = \arg\min_{\mathbf{a}} \sum_{k=k_x}^{M-1} \left(|X_H(k,l)|^2 - \mathbf{a}^T \mathbf{Q}_k^l \mathbf{a} \right)^2. \quad (5)$$

In practice, however, the spectral energy information delivered to the decoder is $\bar{E}(p,t)$, the energy averaged over the (p,t) frequency-time bin. Furthermore, $\bar{E}(p,t)$ reflects the total energy of all components constituting the original audio signal, while only the energy of tonal components is relevant to accurate estimation of sinusoidal magnitudes. A reasonable approach to achieve this is to estimate the noise energy and subtract it from the total energy. Once the average tonal energy $\bar{E}_T(p,t)$ is appropriately extracted from $\bar{E}(p,t)$, the sinusoid magnitudes averaged over $[l_{t-1}, l_t)$ are estimated by solving the optimization problem:

$$\hat{\mathbf{a}}^t = \arg\min_{\mathbf{a}} \sum_{p=1}^{P} \left(\bar{E}_T(p,t) - \mathbf{a}^T \bar{\mathbf{Q}}_p^t \mathbf{a} \right)^2, \quad (6)$$

where the "average matrix" $\bar{\mathbf{Q}}_p^t$ is defined as

$$\bar{\mathbf{Q}}_p^t = \frac{1}{\Delta} \sum_{k=k_{p-1}}^{k_p-1} \sum_{l=l_{t-1}}^{l_t-1} \mathbf{Q}_k^l. \quad (7)$$

The final SM-HFR step consists of converting the estimated magnitude parameters, combined with frequency and phase parameters, to the QMF domain signal $\{X_H(k,l)\}_{k=k_x}^{M-1}$ by applying $\mathbf{X}_H^l = \mathbf{Q}^l \mathbf{a}^l$.

4. SM-HFR USING DECODER INFORMATION

In the previous section, it was shown that the SM-HFR problem for audio signals with strong harmonics may be equivalently viewed as the parameter estimation problem of the underlying harmonics. For estimation of frequency parameters of the high band harmonics, the fundamental frequency should be presented to the SM-HFR decoder. Moreover, information on the spectral energy of tonal components should be supplied in order to optimally estimate the magnitudes of the harmonics. In [4], the simplest and most convenient way to provide the SM-HFR decoder with such information was implemented, namely, the fundamental frequency was directly delivered to the decoder as additional side information, at a small cost in bit-rate. The spectral energy of tonal components was provided by replacing unneeded control information in the HE-AAC side information with the noise energy, and subtracting such noise energy from the total energy of all components constituting the original signal.

In this paper, an alternative approach for SM-HFR is proposed, where information available at the *standard* HE-AAC decoder is exploited to estimate the model parameters. The block diagram of the proposed SM-HFR in the framework of the HE-AAC decoder is given in Fig.1. The lower band reconstruction of the core AAC is input to the fundamental frequency estimate for the frequency parameters of the high band harmonics. For the spectral energy of tonal components, the regenerated high band by the HE-AAC decoder and the spectral energy in the HE-AAC side information are directly involved. For the (p,t) frequency-time bin defined as $[k_{p-1}, k_p) \times [l_{t-1}, l_t)$, let us denote TNR and energy at the k-th channel of the regenerated high band as $\hat{T}(k)$ and $\hat{E}(k)$, respectively. The noise energy at the k-th channel is computed as $\hat{E}_N(k) = \hat{E}(k)/(1+\hat{T}(k))$. The averaged noise energy of the original signal at the p-th frequency bin $\bar{E}_N(p,t)$ is approximated with an average of $\hat{E}_N(k)$ over $[k_{p-1}, k_p)$. This approximation is justified by the fact that the additional control parameters are determined at the HE-AAC encoder to match the TNR of the regenerated high band with that of the original. Then, the spectral energy of tonal components is obtained by subtracting the approximated noise energy from the total energy specified in the HE-AAC side information, i.e. $\bar{E}_T(p,t) \cong \bar{E}(p,t) - \bar{E}_N(p,t)$. The computed tonal energy is incorporated within the optimization of (6). Also, the computed noise energy $\bar{E}_N(p,t)$ is applied to synthesis of noise component with random phase to complement the SM-HFR and produce more natural regeneration of the high band.

5. EXPERIMENTAL RESULTS

In order to verify the performance of the proposed approach, we incorporated the proposed SM-HFR technique within the conventional HE-AAC decoder. For fundamental frequency estimation, the "two way mismatch" procedure of [8], with modification to ensure consistency with past estimates, was incorporated. The fundamental frequency was estimated at each time bin and interpolated to the time resolution for the QMF subsampling. For the spectral energy of tonal components, the TNR at each channel of the HE-AAC reconstruction was measured with prediction gain (defined in [2]) of subband signals in the QMF domain. The implementation of the proposed decoder based SM-HFR was tested on a number of mono audio signals. Six test sequences having high energy harmonics in the high band were selected for quality measurement. Two saxophone pieces and two harmonica pieces (accompanied by background sounds in the lower band) were sampled from com-

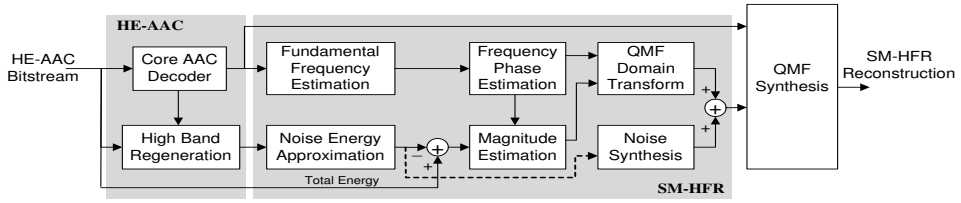

Figure 1: Block diagram of the proposed SM-HFR decoder

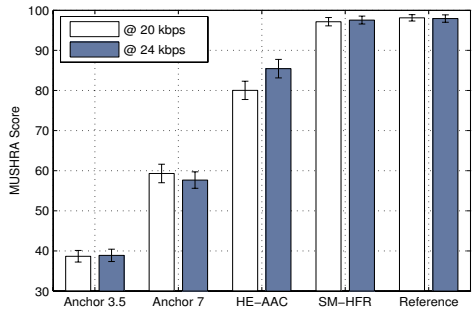

Figure 2: Results of the MUSHRA test with 95 % confidence intervals at 20 and 24 kbps for two anchor signals, hidden reference, and reconstructions by the conventional HE-AAC and the proposed decoder based SM-HFR method.

mercial CDs for demonstrating degradation patterns by tonal artifacts and replication of uncorrelated lower band components, respectively. Also, two test items (bagpipe and pitchpipe) from the verification tests of [9], were selected. For subjective quality assessment under controlled and repeatable conditions, we employed the multi-stimulus test with a hidden reference (MUSHRA), including low-pass filtered anchors with 3.5 and 7 kHz bandwidth [10]. The listening tests were performed by nine listeners, where each listener gave a score for each test sequence.

The SBR bit-stream encoded at several bit-rates was applied to the conventional HE-AAC decoder and the SM-HFR decoder. From the comparison of decoder reconstructions, it was observed that for audio with high energy harmonics, quality degradation due to the unwanted replication of uncorrelated lower band components has severe perceptual impact on the audio quality of the regenerated high band. Also, the quality degradation due to the tonal artifact was noticeably audible in the regenerated high band by HE-AAC decoder, especially at low bit rates, while the proposed SM-HFR achieves improved reconstruction quality by eliminating such patterns of quality degradation. Fig.2 presents the overall subjective quality comparison of the reconstructions by HE-AAC and SM-HFR at 20 and 24 kbps for all test items. The figure confirms that the performance of the proposed SM-HFR was significantly superior to the conventional HE-AAC. Particularly, the performance gains of the proposed approach increase at lower bit rate.

6. CONCLUDING REMARK

We proposed an approach for effective high frequency regeneration based on sinusoidal modeling in the framework of HE-AAC decoder. All information necessary to estimate sinusoidal parameters is extracted at the decoder by exploiting HE-AAC side information and standard HE-AAC decoder information. For the class of audio signals possessing significant high energy harmonics in the high band, simulation results demonstrate that the proposed SM-HFR decoder achieves improved reconstruction quality. Further study areas may include the application to general harmonic signals, with more robust fundamental frequency estimation.

7. REFERENCES

[1] ISO/IEC JTC1/SC29/WG11, "Text of ISO/IEC 14496-3:2001/FDAM1, Bandwidth Extension," ISO/IEC JTC1/SC29/WG11 N5570, Mar. 2003.

[2] 3GPP TS 26.404: "Enhanced aacPlus encoder SBR part". June 2004.

[3] M. Wolters, K. Kjorling, D. Homm and H. Purnhagen, "A closer look into MPEG-4 High Efficiency AAC," *115th AES Convention*, Preprint 5871, Oct. 2003.

[4] S. -U. Ryu, J. -H. Chang, and K. Rose. "Sinusoidal modeling for high frequency regeneration in perceptual audio coding." submitted to *IEEE Trans. Speech and Audio Processing*, 2005.

[5] ISO/IEC JTC1/SC29 WG11 MPEG, "Text of ISO/IEC 14496-3:2001/AMD 1:2003, bandwidth extension,'" Nov. 2003.

[6] P. Ekstrand, "Bandwidth extension of audio signals by spectral band replication," *Proc. 1st IEEE Benelux Workshop on Model based Processing and Coding of Audio*, Nov. 2002

[7] L. B. Almeida and J. M. Tribolet, "Non-stationary spectral modeling of voiced speech," *IEEE Trans. Acoust., Speech and Sig. Process.,* vol. 31, pp. 664-678, June 1983.

[8] R. C. Maher and J. W. Beauchamp, "Fundamental frequency estimation of musical signal using a two-way mismatch procedure," *J. Acoust. Soc. Amer.*, vol. 95, no. 4, pp. 2254-2263, April 1994.

[9] ISO/IEC JTC1/SC29 WG11 MPEG, "Report on the verification tests of MPEG-4 high efficiency AAC," N6009, Oct. 2003.

[10] ITU-R, Method for the subjective assessment of intermediate quality levels of coding systems, 2001.