The 23rd Convention of Electrical and Electronics Engineers in Israel, Tel-Aviv, Israel, Sept 7-8, 2004.

IMPROVEMENT OF A PARAMETRIC MODEL FOR AUDIO SIGNAL COMPRESSION AT LOW BIT RATES

Michael Moskovitz, Dan Chazan and David Malah

Department of Electrical Engineering
Technion- IIT, Haifa 32000, Israel
moscomic@techunix.technion.ac.il, chazan@il.ibm.com, malah@ee.technion.ac.il

ABSTRACT

The HILN (harmonic, individual lines, and noise) audio coder is included in the MPEG-4 audio standard for coding audio signals at very low bit rates (at and below 16 kbps). It uses a parametric model to efficiently represent audio signals under low bit rate constraints.
In this paper we propose several improvements to the estimation and coding of the HILN model parameters. These include: estimation of the frequencies of closely spaced tones, estimation of multi-pitch periods, improved amplitude representation of harmonics, and better use of the underlying perceptual model. The proposed improvements result indeed in better audio quality, manifested in a 0.4 points improvement in EAQUAL score, used for evaluating the audio quality, as compared to HILN, at both 16 and 12 kbps.

1. INTRODUCTION

In the context of evolving multimedia applications new demands for very low bit rate audio coding arise. Coping with limited resources such as the bandwidth of transmission channels and memory for storage applications requires high coding efficiency.

In the last decade, there has been a widespread use of MPEG standards for audio compression, such as MP3 (MPEG-1 layer 3), AAC (Advanced Audio Coding), Twin-VQ (Transform domain Weighted Interleaved Vector Quantization), and HILN (Harmonic Individual Lines and Noise) [1,2,3]. The later standards have produced coding techniques for audio signal compression at very low bit rates (16 kbps and below), although at reduced audio quality. All standards are designed to extensively exploit the properties of signal perception by the human auditory system, and therefore prevent redundant coding of information which will not be heard, anyway, by the human ear. The reason that high compression is feasible is the limited sensitivity of the human ear [4]. This is reflected, for example, in the fact that some sounds are masked by certain louder sounds. This means that masked sounds need not to be coded, reducing the amount of information needed to represent the audio signal. Consequently, the masking property is one of the most important factors in attaining good audio compression.

This paper focuses on improving the HILN parametric model for audio signals (speech and music) sampled at 16 KHz and coded at a low bit rate of 16 kbps (one bit per sample) and below, to obtain better subjective audio quality.

The organization of the paper is as follows: Section 2 gives an overview of the HILN parametric model. The improved HILN is described in Section 3. Results of tests evaluating the subjective quality of the improved HILN are presented in Section 4 and conclusions are drawn in Section 5.

2. THE HILN MODEL

The HILN coder is included in the MPEG-4 Audio standard targeted for coding audio signals at very low bit rates. This model is based on the decomposition of the input signal into audio objects, which are described by appropriate source models and are represented by model parameters [5, 6]. The audio objects are individual sinusoids, harmonic tones and noise.
The model represents the audio signal as a finite sum of sinusoids. Each of the L sinusoids is described by its frequency f_i, amplitude a_i and phase φ_i:

$$\hat{x}(t) = \sum_{i=1}^{L} a_i(t) \cdot \sin(\varphi_i + 2\pi f_i t) \qquad (1)$$

Because of the low phase sensitivity of the human ear, sinusoids phase information is not transmitted; on the other hand it is essential to ensure phase continuity of sinusoidal tracks [5-9].
A harmonic tone object is characterized by its fundamental frequency (pitch) and the amplitudes of all harmonic partials. A noise object is described by its power spectral density and therefore is represented by parameters relating to intensity and spectral shape.

Due to the very low target bit rate, only the parameters of a small number of components can be transmitted. Therefore an auditory perception model is employed to select those components which are most important for the perceptual quality of the signal.

The parametric model analysis is done on a frame by frame basis, because most audio signals are quasi-stationary, i.e., their properties change slowly with time. For each time frame (typically of 32 msec in duration) a set of model parameters is computed which describe the input signal in this frame. The frames are transformed to the frequency domain, where the decomposition into audio objects is done.
The sinusoid components are extracted iteratively, using an analysis by synthesis loop, which exploits the properties of sound signals perception by the human auditory system. In each iteration, the most prominent sinusoid above the masking

threshold is found. Hence, the most important components for sound perception are extracted first. This allows a measure of control over the total number of sinusoids which will be extracted, according to the desired bit rate.

The extraction of the sinusoids is followed by fundamental frequency estimation, which describes the frequencies of many harmonics as multiples of the fundamental frequency. The remaining sinusoids, which do not match an integer multiple of the fundamental frequency, create a set of individual sinusoids. The residual signal, obtain after removing all the extracted sinusoids from the input signal, is considered a noise-like signal.

3. IMPROVED MODEL

The HILN model has several disadvantages that are addressed below.

3.1 Estimation of closely spaced frequencies

In HILN, Two closely spaced sinusoids may be detected as a single sinusoid, because of frequency resolution limitation. In the improved model, a new technique for identifying closely spaced components is applied. The technique is based on maximizing the correlation between the sinusoidal representation of the estimated components and the input signal x(n). This approach attempts to minimize the model fitting error E,

$$E = \sum_{n=0}^{N-1} \left(\left(x(n) - \sum_{i=1}^{L} a_i \cdot \cos(2\pi f_i \cdot \frac{n}{fs} + \varphi_i) \right) \cdot w(n) \right)^2, \quad (2)$$

where *fs* is the sampling rate frequency, *w(n)* is a window function and the frame length is *N* samples.
For efficient analysis this equation is written in the frequency domain, where $\underline{X}$ is the FFT of the input signal, $\underline{A}$ is vector of spectral coefficients and Q is a matrix having L columns with each column representing the window function frequency response shifted by f_i:

$$E = (\underline{X} - Q \cdot \underline{A})^T \cdot (\underline{X} - Q \cdot \underline{A}) \quad (3)$$

The unknown parameters are the spectral coefficients, the frequencies, and their number - L. Given the frequencies it is possible to solve the linear equations for the spectral coefficients in $\underline{A}$ which minimize E so their explicit appearance is eliminated. This reduces the size of the parameter space. The (complex) spectral coefficients are given by

$$A = (Q^T Q)^{-1} Q^T X \quad (4)$$

Substituting this expression for $\underline{A}$ in the above error expression, we note that the error E can be rewritten in terms of the matrix Q as follows:

$$E = \underline{X}^T \left(I - Q(Q^T Q)^{-1} Q^T \right) \underline{X} = \underline{X}^T \cdot (I - P_T) \underline{X} \quad (5)$$

E is minimized by maximizing $\underline{X}^T P_T \underline{X}$, where P_T is dependent only on the sinusoid frequencies (independent of the amplitudes).

The calculations in the frequency domain enables a calculation of the minimum of this equation over a reduced frequency band (focusing on a relevant frequency region) so that $\underline{X}$ and Q includes only the relevant frequency bins.

For the case of two sinusoids having closely spaced frequencies, we set L=2. Fig. 1 shows the three-dimensional plot of the expression $\underline{X}^T P_T \underline{X}$ as a function of two frequencies. The global peak location in the three-dimensional plot gives the estimates of the two frequencies, one along each axis.

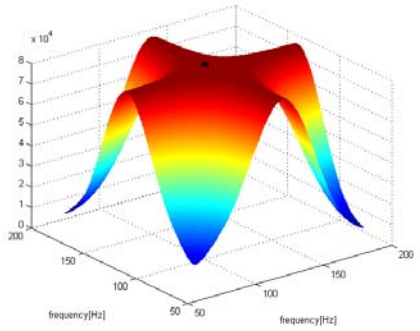

Fig. 1: Plot of the error surface as a function of two frequencies. In this example the frequencies of the two sinusoids are 130Hz and 140Hz.

3.2 Multi-Pitch Estimation

The HILN uses a single pitch, which may yield a poor representation of complex audio signals that typically have more then one pitch. Usually, there are very few harmonic components which are represented by a single pitch, leaving out many individual sinusoids. Therefore, many sinusoids won't be coded, due to the lack of transmission bits. The improved model uses a new technique for multi-pitch estimation, based on searching of fundamental frequencies that *maximally cover* a given set of frequencies.

It is an iterative process that determines dominant pitch values (fundamental frequencies). Each iteration works on the set of frequencies that are not represented by any fundamental frequencies detected earlier. In the first iteration the set contains all existing individual sinusoids frequencies. Each frequency f_i contributes to a decision function by applying a comb function on the integer divisors of f_i, as shown in Fig. 2. The frequency divisors lie within the search band, which is between 50Hz to 2000Hz.

Summing up all the comb functions gives a decision function, whose maximum value represents the dominant fundamental frequency. Fig. 3 shows an example for this decision function (for the set of frequencies: 100, 300, 600, 900, 1200 and 3000 Hz). The y-axis shows the number of harmonics which can be represented by a specific frequency. As it can be seen from Fig. 3 the maximal value results in 100Hz. This means that the frequency 100Hz gives maximal cover for all 6 frequencies, so it will be chosen to be the fundamental frequency. In practice, two-fundamental frequencies were found sufficient for efficient representation.

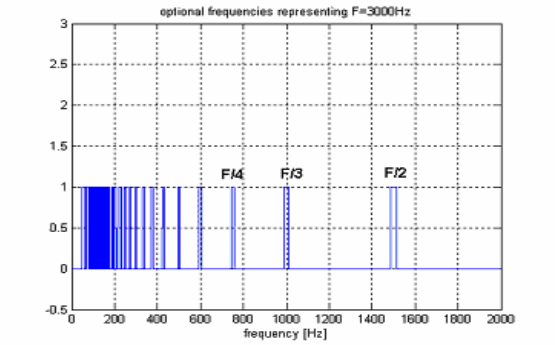

Fig. 2: Comb function showing the optional frequencies.

In the example, we look for a fundamental frequency that represents the frequency 3000Hz. The frequency divisors appear at 1500Hz (division by 2), 1000Hz (division by 3), 750Hz (division by 4), and so on.

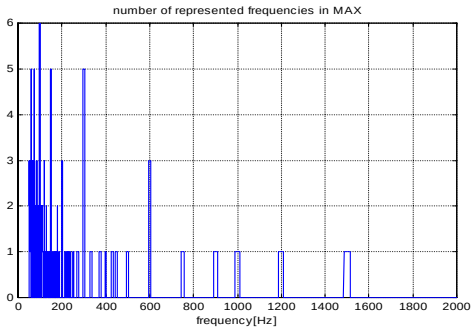

Fig. 3: Searching for fundamental frequencies by maximal cover. In this example the input set of frequencies are: 100, 300, 600, 900, 1200 and 3000 Hz. The frequency of 100Hz provides maximal cover.

3.3 Amplitudes Representation

The HILN represents the harmonic amplitudes by a coarse spectral envelope. The spectral envelope is represented by a set of LPC coefficients. Usually, the envelope gives a large deviation from the true amplitudes, which usually causes a significant degradation in sound quality.

The improved model better represents the amplitudes of the harmonic partials by a modified spectral envelope, using an iterative method for calculating the LPC coefficients to adjusts the harmonic amplitudes to the model amplitudes (obtained by sampling the envelope), in addition to reducing the amplitudes dynamic range and the inclusion of perceptual properties of the human auditory system in the calculations [11]. While conventional LPC modeling accuracy depends on the spectral shape, it may be more appropriate to increase the accuracy for perceptually more important frequencies. This is achieved by warping the frequency scale to devote a larger portion of the total spectrum modeling accuracy to these frequencies [11]. Furthermore, we added the option of matching two spectral envelopes to even and odd harmonics separately, whenever the bit rate permits.

In the iterative method the LPC coefficients are calculated on a synthesized spectrum [11]. The LPC coefficients define the spectral envelope and thus the model amplitudes. The first iteration is a regular LPC computation that typically results in amplitudes that are either too low or too high, as compared to target amplitudes, as shown in Fig. 4(a). The plot shows the spectrum, where each peak represents the amplitude and frequency components.

Next, the distortion is measured by averaging the logarithmic differences between the source amplitudes and the model amplitudes. In next iteration the synthesized spectrum is modified according to the last iteration results. When the model amplitude is higher than the source amplitude at a specific frequency, the synthesized spectrum is reduced at that frequency and when the model amplitude is smaller than the source amplitude the synthesized spectrum is amplified. The iterative process stops when the distortion is sufficiently reduced. Fig. 4(b) shows the model amplitude after the described iterative process is applied.

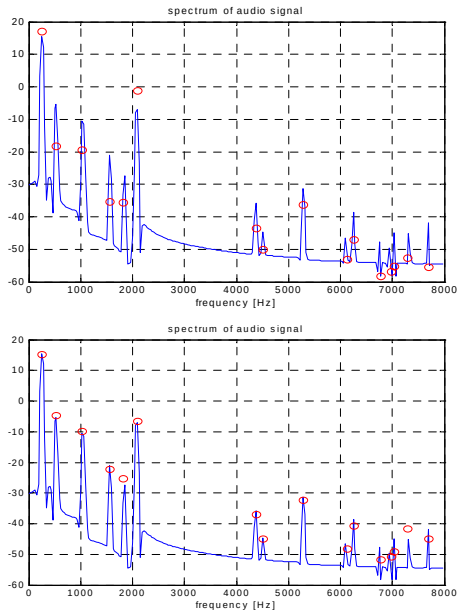

Fig. 4: Amplitudes representation. The synthesized spectrum is shown by a solid line, while the model amplitudes, which are derived from the spectral envelope, are marked by circles. (a) Coarse spectral envelope, calculated in the first iteration (regular LPC). (b) Shaping the spectral envelope by the iterative algorithm for improved amplitude matching.

3.4 Sinusoidal components extraction

In HILN, a limited number of sinusoids are extracted in the analysis by synthesis loop, due to the lack of transmission bits. The proposed coder is more efficient and can afford the representation of all the sinusoidal components in the input signal. While HILN calculates in each iteration a masking threshold evoked by the sinusoids extracted in the previous iterations. The proposed improved model makes better use of the masking characteristics by calculating the masking threshold evoked by all signal components at once. The sinusoids whose amplitudes are below the masking threshold are removed, since they won't be heard by the human ear.

3.5 Coding

The model parameters are finally quantized and multiplexed to form a bit-stream, which is transmitted to the decoder. This work is mainly concerned with model improvements and less with the quantization process, thus a commonly used quantization scheme is employed at the final coding stage.

The spectral shape of the noise object and the harmonic amplitudes are represented by spectral envelopes, via the LPC coefficients. The coefficients are transformed to LSF (Line Spectral Frequencies) parameters, which are quantized by vector quantization. The frequency and amplitude parameters of individual sinusoid objects and the fundamental frequencies of harmonic objects are quantized using a logarithmic law. For a sinusoid that continues from the previous frame, only the frequency and amplitude changes are transmitted, since this requires fewer bits. Each harmonic object requires an additional parameter that indicates the harmonic location. This parameter quantifies the difference from the previous harmonic in terms of an integer multiples of the fundamental frequency and is coded using a Huffman table. The proposed system operates at both fixed and variable rates in the range of 12 to 16 kbps. Further details can be found in [12].

4. SUBJECTIVE QUALITY EVALUATION

The improved model was tested for perceptual quality using the EAQUAL software [13], which provides an objective quality measure for reconstructed audio files as compared to the original, and is claimed to match subjective quality ranking. The most interesting output score given by EAQUAL is the ODG (Objective Difference Grade). An ODG of -4 means a very annoying disturbance, while ODG of 0 means that there is no perceptible difference. The test results are shown in Table 1. It can be seen that there is an improvement of 0.4 points (from -3.3 to -2.9) in comparison to HILN and an improvement of about 0.5 points in comparison to TWIN-VQ at 16kbps. The proposed model, without quantization of the parameters, achieves a grade of about -2.8.

Grade	Bit Rate (kbps)	Coder
-3.28	16	HILN
-3.43	12	HILN
-3.36	16	TWIN-VQ
-3.02	12-variable	Proposed
-2.78	-	Proposed model
-2.88	16	Proposed

Table 1: EAQUAL ODG results for test coders.

5. CONCLUSIONS

In this paper the HILN model for coding audio signals was reviewed and several improvements were proposed. The use of an algorithm for resolving sinusoids with closely spaced frequencies increased the richness of the model and improved the representation of the frequency components. Using multi-pitch for harmonic representation measurably increased the number of coded sinusoids without increasing the bit-rate. The improved amplitude representation yields a better fit of the spectral envelope to the sinusoids amplitudes. The improved model was used for (mono) audio coding at 16 and 12 kbps and was compared to HILN coder at the same bit rate. A quality evaluation showed an improvement of 0.4 points in EAQUAL score at the cost of about twice the run-time.

6. ACKNOWLEDGMENTS

The authors are thankful to Dr. Heiko Purnhagen for running his standard HILN coder on our audio data set used in the comparison tests. We are also thankful to Alex Kobzanchev for the fruitful discussions about the frequency domain approach for resolving close frequencies.

7. REFERENCES

[1] S. Shlien, "Guide to MPEG-1 Audio Standard", *IEEE Transactions on Broadcasting*, Vol. 40, No. 4, pp. 206-218, December 1994.

[2] J. Herre, B. Grill, "Overview of MPEG-4 Audio and its Applications in Mobile Communications", Audio Department, Erlangen, Germany, *AES 17th International Conference* on High Quality Audio Coding, 1999.

[3] J. Herre, H. Purnhagen, "General Audio Coding", the book "MPEG-4", chapter 11, pp. 487-544, 1999.

[4] A. Spanias, "Perceptual Coding of Audio", *Proceedingsof the IEEE*, Vol. 88, No. 4, April 2000, pp.451-467.

[5] H. Purnhagen, B. Edler, Charalampos Ferekidis, "Object-Based Analysis/Synthesis Audio Coder for Very Low Bit Rates", *AES 104th convention*, preprint 4747, May 1998.

[6] H. Purnhagen, "An Overview of MPEG-4 Audio Version 2", *AES 17th International Conference* on High-Quality Audio Coding, Florence, Italy, September 1999.

[7] B. Edler, H. Purnhagen, "Parametric Audio Coding", *5th International Conference* on Signal Processing (ICSLP 2000), Beijing, August 2000.

[8] Heiko Purnhagen, "Advances in Parametric Audio Coding", University of Hannover, Germany, Proc, 1999 *IEEE Workshop* on Application of Signal Processing to Audio and Acoustic.

[9] H. Purnhagen, N. Meine, "HILN – The MPEG-4 Parametric Audio Coding Tools", University of Hannover, Germany, ISCAS 2000, *IEEE International Symposium* on Circuits and Systems, pp. 201-204, May 2000.

[10] D. Chazan, M. Tzur, R. Hoory, G. Cohen, "Efficient Periodicity Extraction Based on Sine Wave Representation and its Application to Pitch Determination of Speech Signals", *IBM Research*, Israel, 2001.

[11] Kondoz, "Multi-band excitation speech coder", chapter 8, pp.239-272, 1996.

[12] M. Moskovitz, " Improvement of a parametric model for audio signal compression at low bit rates", *M.Sc. thesis*, 2004.

[13] Link for EAQUAL software, "http://www.mp3-tech.org /programmer/misc. html".

[14] T. Ramabadran, A. Smith, M. Jasiuk, "An Itertaive Interpolative Transform Method for Modeling Harmonic Magnitudes", *IEEE Workshop Proceedings*, pp. 38-40, October 2002.