Improvement of a parametric model for audio signal compression at low bit rates

Michael Moskovitz

Improvement of a parametric model for audio signal compression at low bit rates

Submitted in Partial Fulfillment of The
Requirements for the Degree of Master of Science in
Electrical Engineering.

Michael Moskovitz

Submitted to the Senate of
the Technion - Israel Institute of Technology

ADAR, 5764 HAIFA APRIL 2004

Acknowledgments

The research thesis was carried out under the supervision of Dr. Dan Chazan and Prof. David Malah in the department of Electrical Engineering.

It is a pleasure for me to thank Prof. David Malah and Dr. Dan Hazan for their grate involvement, devoted guidance and invaluable help throughout all stages of the research.

I am particularly indebted to Dr. Purnhagen for running his standard HILN coder on our audio data set used in our comparison tests.

I also thank my colleague Alex Kobzanchev for the helpful cooperation in the development of the closely spaced frequencies separation algorithm.

I also thank the Signal Processing Lab for the technical help.

The generous support of the Technion is gratefully acknowledged.

Contents

Contents (continue)

Contents (continue)

Contents (continue)

List of Figures

List of Tables

Abstract

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 two decades, there has been a wide 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). The later standards have produced coding techniques for audio signal compression at very low bit rates (16 kbps and below), at the price of 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. This is reflected for example in the fact that some sounds are masked by the certain louder sounds. This means that masked sounds do not have 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 work 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.
The HILN coder is a version of MPEG-4 Audio 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. Those audio objects are individual sinusoids, Harmonic tones and Noise.

An individual sinusoid object is described by its frequency, amplitude and phase parameters.
Because of the low phase sensitivity of the human ear, phase information for sinusoids is not transmitted, on the other hand it is essential to provide phase continuity of sinusoidal tracks.
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 for a small number of components can be transmitted. Therefore a perception model is employed to select those components which are most important for the perceptual quality of the signal.

The first stage of the parametric model analysis divides the signal into frames. This is based on the assumption that most audio signals are quasi-stationary i.e., their properties change slowly with time. For each time frame 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 signal perception by the human auditory system. In each iteration, the sinusoid which is most prominent above the masking threshold is found, thus those components that are most important 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 production of the sinusoids is followed by a step of fundamental frequency extraction, which describes the frequencies of many harmonics as multiples of the fundamental frequency. The rest of the sinusoids which do not match an integer multiple of the fundamental frequency create a set of individual sinusoids. The remaining residual signal (removing all the extracted sinusoids from the input signal) is considered a noise-like signal.

The HILN model has several disadvantages, as follows:
Firstly a limited number of sinusoids are extracted in the analysis by synthesis loop, due to the lack of transmission bits. The improved model presented in this work

overcomes this limitation and extracts all the sinusoidal components from the input signal.

In addition two closely spaced sinusoids will be detected as one sinusoid by HILN, because of frequency resolution limitation. In the improved model, a new technique for identifying closely spaced components is presented. The technique is based on maximizing the correlation between the sinusoidal representation of the estimated components and the input signal, in a reduced frequency band.

The HILN calculates in each iteration a masking threshold evoked by the sinusoids extracted in the previous iterations. The 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.

The HILN uses a single pitch, which gives poor representation for complex audio signals such as multi-pitch ones. 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.

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 high deviation from the real amplitudes, which 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, which adjusts the harmonic amplitudes to the model amplitudes (sampled from the envelope), in addition to reducing the amplitudes dynamic range and the inclusion of the perceptual properties of the human auditory system in the calculations. 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 can be achieved by warping the frequency scale to devote a larger portion of the total spectrum modeling accuracy to the perceptually more important frequencies. In addition, we added the optional using two spectral envelopes, where the bit rate permits this.

The components' parameters are finally quantized and multiplexed to form a bit-stream, which is transmitted to the decoder. This work is mainly involve in the 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 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 which is continued 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 to the previous harmonic in terms of an integer multiples of the fundamental frequency and is quantized using a Huffman code.

The proposed system operates at both fixed and variable rates in the range of 12 to 16 kbps.

The improved model was tested for perceptual quality using the EAQUAL software which provides an objective quality measure for reconstructed audio files as compared to the original. The most interesting parameter output 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 showed 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 the cost of about twice the run-time.

References רשימת מקורות

[1] ISO/IEC 11172-3 international standard, "Information Technology – Coding of moving pictures and associated audio for digital storage media up to about 1.5Mbit/s", 1993.

[2] Karlheinz Brandenburg, "MP3 and AAC Explained", AES 17th International Conference on High Quality Audio Coding, Fraunhofer Institute for Integrated Circuits FhS-IISA, Erlangen, Germany, 1999.

[3] Karlheinz Brandenburg, "Low Bitrate Audio Coding- State of the Art, Challenges and Future Directions", Proceeding of ICSLP2000.

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

[5] K. R. Rao, J. J. Hwang, "Techniques and Standards for Image, Video, and Audio Coding", chapter 10, pp. 242-272, 1996.

[6] Karlheinz Brandenburg, Oliver Kunz, Akihiko Sugiyama, "MPEG-4 natural audio coding", Signal Processing: Image Communiction, 15 (2000), pp. 423-444.

[7] Tetsuya Takahashi, Takashi Morita, "Card Size Portable Audio Player Using High Quality Audio Coding Technology TWIN VQ", Cyber Space Laboratories, Japan, 2000, pp. 907-913.

[8] Jurgen Herre, Bernhard 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.

[9] N. Iwakami, T. Moriya and S. Miki, "High quality audio coding at less than 64 kbit/s by using transform-domain weighted interleave vector quantization (TwinVQ)", Proc. ICASSP-95, May 1995, pp. 3095-3098.

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

[11] Bernd Edler, Heiko Purnhagen, "Concepts for Hybrid Audio Coding Schemes Based on Parametric Techniques", Preprint 4808, 105th AES Convention, San Francisco, September 1998.

[12] Heng-Ming Tai, Shudei Jiang, "MPEG-4 Parametric Audio Coding and its Implementation", Department of Electrical Engineering, University of Tulsa, pp. 762-766, 1999.

[13] Bernd Edler, Heiko Purnhagen, "Parametric Audio Coding", 5th International Conference on Signal Processing (ICSLP 2000), Beijing, August 2000.

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

[15] Heiko Purnhagen, Nikolaus 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.

[16] Heiko Purnhagen, Bernd Edler, Charalampos Ferekidis, "Object-Based Analysis/Synthesis Audio Coder for Very Low Bit Rates", University of Hannover, Germany, AES 104th convention, preprint 4747, May 1998.

[17] ISO/JTC 1/SC 29/WG11 International standard, "Information Technology – Very Low Bitrate Audio-Visual Coding", October 1996.

[18] Heiko Purnhagen, Nikolaus Meine, Bernd Edler, "Speeding up HILN – MPEG-4 Parametric Audio Encoding with Reduced Comlexity", University of Hannover, Germany, AES 109th convention, September 1999, http://www.tnt.uni-hannover.de/org/whois/wissmit/purnhage/publications.html

[19] Chris A. Lanciani, "Auditory Perception and the MPEG Audio Standard", thesis, Gorgia Institute of Technology, August 1995.

[20] Andreas Spanias, "Perceptual Coding of Audio", Proceedings of the IEEE, Vol. 88, No. 4, April 2000, pp. 451-467.

[21] Shinfeld Yehuda, "The Encyclopedia of Human Body – Ear and Hearing", pp. 8-21, 1986.

[22] Matti Karjalainen and Tero Tolnen, "Multi-Pitch and Periodicity Analysis Model for Sound Separation and Auditory Scene Analysis", Helsinki University of Technology, Finland, 1999.

[23] Tero Tolnen, Matti Karjalainen, "A Computationally Efficient Multipitch Analysis Model", IEEE transactions on speech and audio processing, Vol. 8, No. 6, pp. 708-716, November 2000.

[24] Anssi P. Klapuri, "Multipitch Estimation and Sound Separation By The Spectral Smoothness principle", Tampere University of Technology, Finland, 2001.

[25] Dan Chazan, Meir Tzur, Ron Hoory, Gilad Cohen, "Efficient Periodicity Extraction Based on Sine Wave Representation and its Application to Pitch Determination of Speech Signals", IBM Research, Israel, 2001.

[26] Tuomas Virtanen, Anssi Klapuri, "Separation of Harmonic Sounds Using Multipitch Analysis and Iterative Parameter Estimation", Signal Processing Laboratory, Tampere University of Technology, Finland, October 2001.

[27] K.K. Paliwal, B.S. Atal, " Efficient Vector Quantization of Lpc Parameters at 24 Bits/Frame", IEEE Tans. On Speech snd Audio Processing, Vol. 1, No. 1,

Jan 1993, pp. 661-664.

[28] Yoseph Linde, Andreas Buzo, Robert M. Gray, "An Algorithm for Vector Quantizer Ddesign"IEEE, Transactions on commusnications, Vol. 28, No.1, January 1980, pp. 84-95.

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

[30] Pushkar Patwardhan, Preeti Rao, “Frequency warped all-pole modeling of vower spectral dependence on voice and vowel quality”, Proceedings of Workshop on Spoken Language Processing, January 2003.

[31] Diemo Schwarz, Xavier Rodet, "Spectral Envelope Estimation and Representation for Sound Analysis-Synthesis", Proceedings of the International Computer Music Conference, 1999.

[32] Tenkasi Ramabadran, Aaron Smith, Mark Jasiuk, "An Itertaive Interpolative Transform Method for Modeling Harmonic Magnitudes", IEEE Workshop Proceedings, pp. 38-40, October 2002.

[33] T. F. Quatieri, *Speech Signal Processing*, chapter 8, "Speech Coding", 2000.

[34] R. J. McAulay, T. F. Quatieri, "Sinusoidal Coding", Chapter 4, Speech Systems Technology Group, MIT Lincoln Laboratory, pp. 121-173, 1995.

[35] Jurgen Herre, Heiko Purnhagen, "General Audio Coding", the book "MPEG-4", chapter 11, pp. 487-544, 1999.

[36] Matthew A. Watson, Peter Buettner, "Design and Implementation of AAC Decoders", Dolby Laboratories, pp 408-409, 2000.

[37] Taeko Miwa, Yoshiaki Tadokoro, "Musical Pitch Estimation and Discrimination of Musical Instruments using Comb Filters for Transcription", Toyohashi University of Technology, Japan, pp. 105-108, 1999.

[38] Anssi Klapuri, "Pitch Estimation Using Multiple Independent Time-Frequency Windows", Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, October 1999, pp. 115-118.

[39] Frank Baumgarte, "A Physiological Ear Model for Specific Loudness and Masking", Proceedings of the 1997 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, October 1997.

[40] James A. Moorer, "Signal Processing Aspects of Computer Music- A Survey", Department of Music, Stanford University, California, February 1977.

[41] Heiko Purnhagen, Nikolaus Meine, Bernd Edler "Sinusoidal Coding Using Loudness Based Componenet Selection", Proc. ICASSP2002, May 2002, pp. 1817-1820.

[42] Hossien Najafzadeh-Azghandi, Peter Kabal, "Perceptual Coding Of Narrowband Audio Signals At 8 Kbit/s", Proc. IEEE Workshop on Speech

Coding for Telecom, pp. 109-110, September 1997.

[43] Ted Painter, , "Perceptual Segmentation and Component Selection in Compact Sinusoidal Representation of Audio", Ph.D. thesis (advisor: Andreas Spanias), 2000.

[44] Ye Wang, Leonid Yaroslavsky, Miika Vilermo, "Some Peculiar Properties of the MDCT", Proceedings of ICSP2000, pp. 61-64.

[45] Das A. and Gersho A,, "Variable dimension spectral coding of speech at 2400 bps and below with phonetic classification", Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp. 492-495, April 1995.

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

[47] Rodet, Xavier, "Musical Sound Signal Analysis/Synthesis: Sinusoidal+Residal and Elementary Waveform Models". IEEE Time-Frequency and Time-Scale Workshop 1997, Coventry, Grande Bertagne.

[48] D.P. Kemp, J.S. Collura, T.E. Tremain, "LPC parameter quantization at 600, 800 and 1200 bits per second", Proceedings of the Tactical Communications Conference, 1992, pp. 71-75.