Audio Engineering Society

Convention Paper

Presented at the 124th Convention
2008 May 17–20 Amsterdam, The Netherlands

A unifying approach to transform and sinusoidal coding of audio

Maciej Bartkowiak

Poznań University of Technology, Chair of Multimedia Telecommunications and Microelectronics,
Polanka 3, 60-965 Poznań, Poland
mbartkow@multimedia.edu.pl

ABSTRACT

The paper describes a new scenario for low bit rate audio compression that combines two classical techniques: transform coding and sinusoidal coding into a united framework. The main idea is to adaptively decompose the audio signal into subbands whose central frequencies follow continuously the local instantaneous frequencies of certain signal components (formants or individual harmonic partials). The content in each subband is encoded in the baseband after frequency shift towards DC. The technique may be considered either as modified transform coding, i.e. coding along instantaneous frequencies or as extended sinusoidal coding, i.e. modeling with partial envelopes that are represented by transform coefficients. In other words, it is a hybrid scheme offering a continuous operating mode between purely transform and purely sinusoidal compression.

1. INTRODUCTION

Recent advantages in audio compression include pushing the limits of available bit rates for wideband audio down to the range quite recently reserved for speech communication (e.g. 16-24kb/s), while still offering a reasonable quality. At such low transmission rates it is still a domain of traditional perceptual spectral coding, because time-frequency representation offered by filterbanks and transforms such as MDCT allows for flexible shaping of the quantization error according to the perceptual model. It is important to notice that high compression efficiency as offered by the MPEG-4 HE AAC codec [1] has been achieved thanks to extending the purely waveform compression scenario by various parametric extensions like PNS [2], SBR [3] and PS [4]. Within these parametric extensions, various signal components are modeled in the encoder and artificially re-synthesized in the decoder, rather than be compressed in a truthful way. On the other hand, there is a growing family of parametric coding techniques, such as MPEG-4 SSC derived from sinusoidal modeling [5], that demonstrate high efficiency, competing with HE-AAC for certain classes of audio data [6].

Depending on the signal content, both standard techniques show several disadvantages. For example, in case of sparse audio spectra, when the signal energy is concentrated around few narrowband peaks (fig. 1), the

AAC codec often fails in delivering high quality at very low bit rates, even though the signal is not very complex. The reason is that low frequency content is favored by the perceptual model in general, and the limited flexibility of bit allocation strategy does not allow for appropriate distortion control of the “gaps” between spectral peaks.

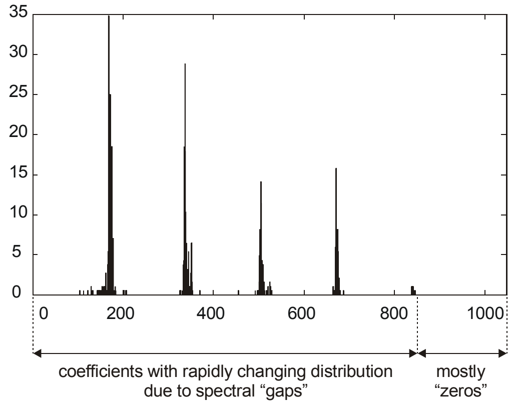

Figure 1: Quantized MDCT coefficients of an audio frame with sparse spectrum

Moreover, in case of non-stationary spectra of sounds with free intonation (e.g. violin, human voice) the underlying model of MDCT transform fails, and the individual harmonic partials are represented quite inefficiently. The side effect is that weaker partials are heavily distorted. Furthermore, the perceptual model uses the original signal as a reference for calculating the masking profile. Thus, at low bit rates it does not take into account the discrepancies in the maskers introduced by the quantization itself.

The SSC codec proved to be very efficient for solo instrument sounds and simple music scores. However it does not cope very efficiently with dense tonal spectra that contain many harmonic partials due to the individual representation of each line. Some partials may be not detected and encoded properly if their frequency is too close to the frequencies of other partials. Furthermore, quantization of partial frequencies introduces unnatural beating of individual components of harmonic series against each other, because their pure harmonic relations are spoiled.

2. THE PROPOSED APPROACH

2.1. Goals and motivations

The new approach described in this paper is a hybrid of transform and sinusoidal coding scenarios. The main goal is to exploit the advantages of the two traditional techniques which are the energy compaction and decorrelation capabilities of MDCT, and the efficiency in representation of non-stationary tonal partials within sinusoidal modeling. The second goal is to create a continuous domain of operation modes that allows for seeking the best compromise between purely sinusoidal and purely transform-based representation (fig. 2). The latter two always serve as fallback options in the case of not achieving a satisfactory coding gain by the hybrid scenario.

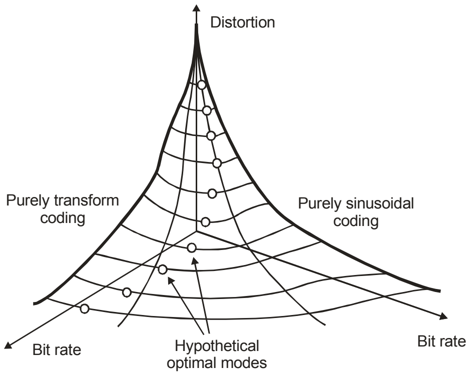

Figure 2: The continuous space of operation modes of the hybrid codec

The idea of combining transform and sinusoidal coding into a hybrid scheme is not new. Levine and Smith [7] consider a compression system with an operating mode switched between sinusoidal+noise modeling and MDCT-based coding, wherein transform technique is employed exclusively for transients. Most of other approaches exploit the idea of analysis-by-synthesis (ABS) coding, wherein subsequent stages operate on a residual obtained as the difference between the original and the reconstructed signal from preceding stage. For example, Daudet and Torresani [8] describe a 3-layer system with MDCT coding being applied in the first order, DWT (wavelet) encoder used for compression of the MDCT coding error, and LPC-based modeling employed for the second residual. Riera-Palou, den Brinker and Gerrits [9] developed a hybrid system for scalable parametric-waveform compression, with parametric (SSC) coder operating on the original signal and a regular pulse excitation (RPE) coder employed to represent the residual signal.

The conceptual difference between our approach and the scenarios mentioned above is that while it is a layered system (since there is a second coding layer for the noise residual), the sinusoidal and transform techniques are unified instead of being used exclusively. Such technique may be considered either as modified transform coding, i.e. coding along instantaneous frequencies or as extended sinusoidal coding, i.e. modeling with special formulation of partial envelopes interpreted as complex-valued narrowband signals [10]. Such envelopes are very efficiently represented by few transform coefficients.

2.2. Codec operation

The encoder (fig. 3) splits the original audio signal into a variable number of subbands with a very flexible configuration. A central frequency for each subband is estimated and tracked. It may be interpreted either as a spectrogram ridge frequency (in the case of a wide subband e.g. corresponding to a whole formant) or as an instantaneous frequency (in the case of a very narrow subband passing only a single sinusoidal partial). All individual subbands are isolated and heterodyned towards DC according to the continuously changing central frequency. The baseband representation of each part of the spectrum is subsequently encoded using an MCLT transform-based technique that makes use of complex-valued basis functions.

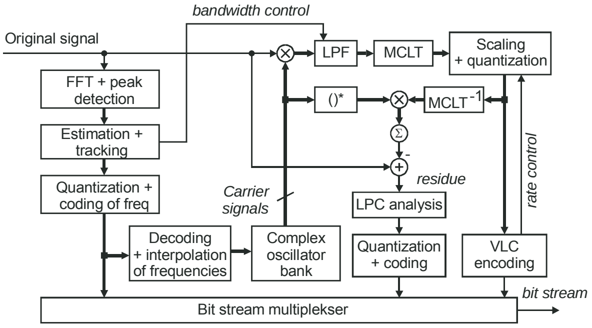

Figure 3: The structure of the hybrid encoder

The most important advantage of the continuous demodulation described above is that natural FM of the spectral content corresponding to pitch changes is reduced, which results in compaction of the tonal energy into very narrow spectral peaks. After "straightening" their frequency evolution, the remaining thickness of each spectral line is mostly related to the AM component. Thus, the compression efficiency of subsequent transform codec is maximized.

A very important feature of the encoder is the subband selection procedure followed by the bandwidth control for the subbands. Acting as a mode control parameter, the subband shape factor (described in 2.3.3) allows for steering of the continuous transition between wideband transform and narrowband sinusoidal coding.

2.3. Representation of tonal energy

2.3.1. Signal analysis

In order to configure the subbands with sufficient temporal resolution, the original signal is analyzed in highly overlapping frames. We use a typical sinusoidal modeling procedure consisting of windowed high resolution FFT (usually 2048 samples, zero padded) and detection of spectral peaks [5]. Furthermore, the QIFFT algorithm [11] is employed for estimation of the peak frequency and corresponding frequency chirp factor of an underlying chirp model for each peak. Estimated peaks matching across consecutive frames are linked into trajectories with the help of a tracking procedure similar to that proposed by Lagrange et al [12], based on linear prediction. The tracking is performed in logarithmic frequency scale in order to achieve a consistent prediction error sensitivity for overtones. At the beginning of each sinusoidal trajectory, we use a combination of linear prediction and forward-backward matching based on chirp factors.

2.3.2. Subband trajectories

Only long trajectories linking strong peaks are considered as the potential subband trajectories. The final decision on which of them are actually used as the central frequencies of subbands is done later, upon the inspection of modulation product spectra.

In order to maintain a consistent frequency shift of particular subbands in the encoder and decoder, quantized frequencies are used for both operations. The trajectories are encoded using an ADPCM method with the same adaptive prediction algorithm as employed in tracking. The frequency prediction error is quantized uniformly and due to the logarithmic scale leads to a relative frequency error in the range of 0.5%-1%. Such a large error would be a source of serious mistuning of harmonic partials in a normal sinusoidal coder, and therefore it would be unacceptable. However, in the proposed approach, the frequency error manifests itself as a complex AM component of the subband (an LF complex envelope) and it is represented by transform coefficients [10].

After decoding from the ADPCM form, the frequencies are interpolated on a sample basis. This interpolation is essential for achieving a continuous compensation of the signal instantaneous frequency changing over time. As opposed to popular implementations of the sinusoidal model, we use a simple cubic spline (Hermite) interpolation of frequencies, because preservation of phase is irrelevant at this stage. First order derivative of frequency is preserved between consecutive segments of each trajectory due to the chirp factor being used in the interpolation formula.

A bank of complex sinusoidal oscillators is driven by the interpolated trajectories. The original signal is heterodyned independently by each of the oscillator outputs, which results in frequency shift towards DC along each of the trajectories. The collection of modulation products is further analyzed using zero-padded FFT in order to determine the optimal bandwidth and decide on which of them are to be encoded as independent subbands.

2.3.3. Adaptive configuration of subbands

For cutting down the computational costs, the modulation products may be initially decimated by a factor related to the maximum assumed bandwidth of the subbands. This factor results from the observations of the operating modes selected by a full-band codec, 4:1 up to 32:1 downsampling ratios showed to be practical in our experiments. Despite of these band limiting operations, the contents of neighbouring subbands may overlap, whenever corresponding sinusoidal trajectories run in close proximity. The key action is to narrow the frequency domain selection within each subband or reject some subbands in such a way that only unique parts of the spectrum are encoded.

Those spectra that exhibit most prominent peaks around DC usually correspond to sinusoidal trajectories related to the strongest partials. They are considered in the first order as the trajectories of candidate subbands for encoding. A following analysis procedure is performed for each of the modulation product spectra:

- Find the strongest peak in the close proximity to DC. This is the central peak of the subband. It may not be exactly centered around DC due to the frequency coding error.

- Determine the main lobe of the central peak, or, in case of several peaks closely spaced to the central one – the resultant lobe of the group. This indicates the initial bandwidth.

- Draw a skirt around the peak with certain amplitude slope. If there are more peaks below the skirt, include those peaks into the subband, and extend the initial bandwidth by the corresponding width of the main lobe of farthest peak (fig. 4). This is the final bandwidth.

- Remove the modulation products related to the peaks already included into the above subband from the list of candidate subbands. There is no need to encode them, since they are covered by the subband just selected.

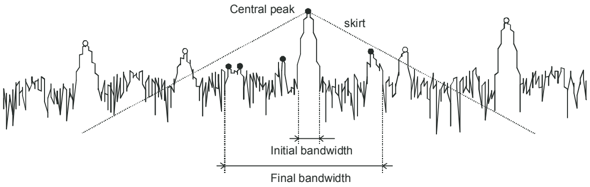

Figure 4: The procedure of bandwidth selection for a subband covering a group of peaks

The shape of the skirt drawn around the central peak determines the tendency of the encoder to select narrower or wider subbands. Therefore a shape parameter (the slope of the skirt) has been selected as the control parameter of the codec operation mode. A special mode control algorithm is needed in order to draw the codec into the most efficient mode. One extreme of this control is a single wide subband that covers the whole spectrum (only possible if no decimation of the modulation products is applied), thus the codec operates very similarly to the standard AAC technique. Another extreme is a very narrowband case, where each subband represents an individual sinusoidal partial. In such case, the number of transform coefficients required to represent the baseband signal (which is in fact the amplitude envelope of the partial) reduces to one complex value: the DC gain + phase error (only possible for subbands with very low frequency coding error). The usual situation is that the control algorithm selects an intermediate setting, and thus it seeks for an optimum point between these two compression scenarios.

2.3.4. Coding of subband contents

An appropriate lowpass filter response is imposed on the product spectra selected for coding, in order to remove the unwanted components that are covered by other subbands. Subsequently, each subband is compressed using a transform technique that takes into account the non-symmetric spectrum. Spectral coding of audio signals is traditionally based on coefficients of MDCT transform. It may be shown that the use of complex kernel transform such as the MCLT proposed by Malvar (1) is an optimal extension for complex-valued signals.

$$X[k] = \sum_{n=0}^{2N-1} x(n)\, w(n)\, \mathrm{e}^{-j\frac{\pi}{N}\left(n+\frac{N+1}{2}\right)\left(k+\frac{1}{2}\right)} \qquad (1\ )$$

where $x(n)$ denotes the time-domain signal, and $w(n)$ denotes a real-valued Hann window function. Since our subband signals are represented in FFT domain, these spectral values are simply translated to the MCLT coefficients by the use of simple relation [13].

The subsequent compression algorithm is very similar to that of the MPEG-4 AAC LC standard. The transform is followed by coefficient perceptual scaling, quantization and entropy coding. In fact, the main difference with AAC is the treatment of the complex-valued coefficients, $X[k]$, and uniform scaling applied to whole subbands. A nonlinear quantizer is used independently for the real and imaginary part, and the degree of quantization is controlled by coefficient scaling,

$$\underline{X_B[k]} = \mathrm{sgn}(X_B[k])\ floor \left\lfloor 2^{(scf_B - gsf)/4} \left| X_B[k] \right|^{3/4} + \varepsilon \right\rfloor \qquad (2\ )$$

Individual scaling factors scf_B are determined for each of the subbands B, plus one global gain factor, gsf controls the degree of distortion of all partials. All coefficients of given subband X_B share the same scaling factor scf_B. Such approach leads to uniform distribution of the quantization noise within each subband so that it may be masked by the energy of central peak. It also allows to adapt an effective bit allocation algorithm primarily developed for an AAC coder. Such simplified approach to distortion control may be considered as inappropriate for wideband coding, however in most situations our codec selects an operating mode with narrow bands.

Due to the narrowband lowpass character of the demodulated subbands, the distribution of quantized values along the data vector is concentrated near its center (corresponding to lowest frequencies). For entropy coding we use a coding scheme taken literally from the MP3 technique. All the "big values" with magnitudes not exceeding 15 are encoded in pairs, using 2D codewords from selected Huffman tables. The whole section is divided into three equal groups, and an optimal Huffman table is selected for each group. Very big values are represented as escape codes. Values from the range of <-1…1> are encoded in quadruples using a dedicated Huffman table.

2.4. Representation of the residual

The compression procedure described in section 2.3.4 does not encompass all the signal energy, since many weaker components are not resolved by the spectral peak detection and tracking. Also, not all weaker parts are included in the subband formation process described in 2.3.3. We observe that the remaining signal energy has usually a stochastic nature and may be easily represented by a typical AR spectral modeling technique.

The residual signal $r(n)$ is obtained in the process of spectral masking, after reconstruction of the compressed tonal energy that takes part in the encoder. For this purpose, the quantized MCLT coefficients are properly re-scaled and an inverse transform is calculated in order to reconstruct the subband signals in time domain (fig. 3). These signals are modulated by the complex conjugated outputs of respective oscillators in order to properly shift them back to the original frequency position. Finally, a masking signal $s(n)$ is calculated as a sum of the reconstructed subbands. This signal is used in the calculation of the residual (3),

$$R_n[k] = X_n[k] \left(clamp\left(\mu \left| S_n[n] \right| \right) * W[k] \right)^{-1}, \qquad (3\ )$$

where k denotes frequency index, μ is a threshold (usually $\mu > 10^5$), $W[k]$ is a smoothing kernel, and *clamp*() is a limiting function such that

$$clamp(x) = \begin{cases} x & x > 1 \\ 1 & x \le 1 \end{cases}, \qquad (4\ )$$

and all operations are done in the SFTF domain with time segments centered around the instant *n*.

A classical frame-based LPC technique is employed for noise modeling of the residual signal. The AR model is fitted in long frames (N=2048) during stationary segments, switched to short (N=256) frames in the presence of transients. The noise power spectrum envelope is represented by 16 PARCOR coefficients that are quantized in log scale with the resolution of 8 bits.

2.5. Automatic control of the operating mode

The goal of the automatic mode control is to select an intermediate operation point in the continuous domain between the two standard techniques in such a way that maximum compression efficiency is achieved for a given signal. This is done by choosing an appropriate skirt shape parameter in the process of subband formation (2.3.3). The skirt shape parameter significantly influences the selectivity of the subbands and thus it changes the number of subbands and their individual bandwidth.

This skirt setting usually should be adapted to the signal on a frame basis and this adaptation imposes a significant computational burden, especially if there is no explicit rule of the parameter control and iterative optimizations are necessary. Furthermore, compression efficiency is hardly measurable for perceptual codecs, especially for parametric codecs at low bit rates where waveform distortion does not reflect perceptual quality at all.

In the current stage of development, our codec tries several settings of the skirt shape parameter and repeats the coding stages for each of the settings, seeking for a best results of a perceptually weighted MSE measure. A more elaborate control method is being developed.

3. SIMULATIONS

3.1. Implementation details

The whole codec has been implemented in software in the Matlab environment as a set of heavily vectorized modules. Unfortunately, due to the high amount of computations required for performing several compression iterations in the pursuit of best efficiency, the operation speed is of the order of 1/80-1/30 of a realtime speed, depending on the complexity of the audio signal.

3.2. Experimental results

For the sake of fair comparison of the proposed technique we decided not to include standard codecs in the listening tests, because our implementation is not yet optimized for best results in all situations. Therefore, instead of a reference ISO or commercial software, the reference was setup using the same implementation as the codec under test. Three variants of the bit stream were produced in each test:

- a forced narrowband setting (codec operation mode essentially equivalent to a standard technique of sinusoidal+noise compression),
- a forced wideband setting (one global subband without frequency shift – equivalent to a streamlined transform codec),
- a hybrid mode, wherein automatic control of skirt shape parameter takes care of the subbands configuration.

A suite consisting of several short music excerpts from the EBU SQAM disc has been used to test the performance of the proposed technique. The reconstructed signals have been compared in a blind listening test by a group of 24 students. The results shown in figure 5 indicate that in most cases the subjective quality of the signal reconstructed from hybrid mode coding was reported as significantly improved with respect to both standard compression scenarios.

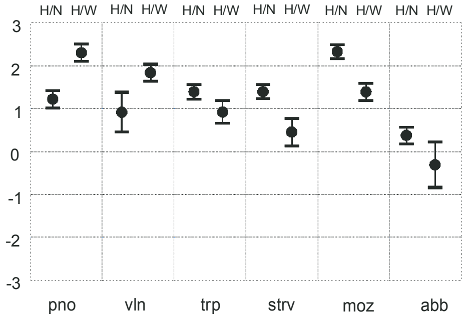

Figure 5: Subjective test results (MOS) for 6 items in 7-point ITU scale. Positive values show a preference of the hybrid over the narrowband (H/N) or wideband (H/W) mode. Bit rate: 24kb/s.

In most cases the listeners reported more natural sound with richer timbre and strongly reduced the various annoying artifacts that were audible in the output of codec forced to operate in one mode.

4. CONCLUSIONS

A novel technique for audio coding is presented in the paper that combines transform and sinusoidal compression scenarios into a single unified approach. Experimental results show that a significant improvement of the subjective quality of reconstructed signal may be achieved by applying an adaptive decomposition of the signal into a variable number of subbands that follow the evolution of nonstationary spectral components and using transform coding within those bands.

The implemented codec still requires an explicit algorithm for controlling the operation mode and bandwidth of the subbands in order to avoid iterative optimizations of the configuration in each audio frame. The computational complexity of the implemented codec is very high, therefore offline compression is the only prospective application of this technique.

5. REFERENCES

[1] ISO/IEC JTC1/SC29/WG11 MPEG, “Int. Standard ISO/IEC 14496-3/Amd1, Coding of Audio-Visual Objects: Audio”, 1999

[2] J. Herre and D. Schulz, “Extending the MPEG-4 AAC codec by perceptual noise substitution”, *104th AES Convention*, 1998, Preprint 4720

[3] M. Wolters, K. Kjörling, D. Homm, and H. Purnhagen, “A closer look into MPEG-4 High Efficiency AAC”, *115th AES Convention*, New York, Oct. 2003

[4] J. Breebaart, S. van de Par, A. Kohlrausch, E. Schuijers, “Parametric Coding of Stereo Audio”, *EURASIP Journal on Applied Signal Processing*, no. 9, 2005, pp 1305-1322

[5] X. Serra, *A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition*, Ph.D. Thesis, Stanford University, Stanford,1989

[6] A.C. den Brinker, E. Schuijers, W. Oomen, “Parametric Coding for High-Quality Audio”, *112th AES Convention,* Apr. 2002, Preprint 5554

[7] S.N. Levine, J.O. Smith III, “Improvements to the switched parametric and transform audio coder” *IEEE Workshop on Appl. of Signal Processing to Audio and Acoustics*, 1999, pp.43-46

[8] L. Daudet, B. Torresani, “Hybrid representations for audiophonic signal encoding”, *Signal Processing*, vol. 82 , no. 11, Nov. 2002

[9] F. Riera-Palou, A.C. den Brinker, A.J. Gerrits, “A hybrid parametric-waveform approach to bit stream scalable audio coding”, *38th Asilomar Conf. on Signals, Systems and Computers,* vol.2, pp. 2250-2254, Nov. 2004

[10] M. Bartkowiak, “A complex envelope sinusoidal model for audio coding”, *10th Int. Conf. on Digital Audio Effects DAFx-07*, Bordeaux, France, 2007

[11] M. Abe, J.O. Smith III, “AM/FM rate estimation for time-varying sinusoidal modelling”, *Int. Conf. Acoustics, Speech and Signal Proc, ICASSP'05,* vol. 3, pp. 201-204, 2005

[12] M. Lagrange, S. Marchand, M. Raspaud, J-B. Rault, “Enhanced Partial Tracking Using Linear Prediction”, *Int. Conf. Digital Audio Effects DAFx-03,* London, UK, 2003

[13] H. Malvar, “Fast Algorithm for the Modulated Complex Lapped Transform”, *IEEE Sig. Proc. Letters*, vol. 10, no. 1, Jan. 2003