# COMPUTATIONALLY EFFICIENT AMPLITUDE MODULATED SINUSOIDAL AUDIO CODING USING FREQUENCY-DOMAIN LINEAR PREDICTION

*Mads Græsbøll Christensen and Søren Holdt Jensen*

Department of Communication Technology
Aalborg University, Denmark
{mgc,shj}@kom.aau.dk

## ABSTRACT

A method for amplitude modulated sinusoidal audio coding is presented that has low complexity and low delay. This is based on a subband processing system, where, in each subband, the signal is modeled as an amplitude modulated sum of sinusoids. The envelopes are estimated using frequency-domain linear prediction and the prediction coefficients are quantized. As a proof of concept, we evaluate different configurations in a subjective listening test, and this shows that the proposed method offers significant improvements in sinusoidal coding. Furthermore, the properties of the frequency-domain linear prediction-based envelope estimator are analyzed.

## 1. INTRODUCTION

Parametric coding of audio and speech has received considerable attention in the research community and standardization bodies in recent years [1, 2, 3]. In order to achieve good performance at low bit-rates, parametric coding relies on signal models that describe the signal in few physically meaningful parameters. Parametric audio coders perform extremely well when the signal fits the signal model. However, when this is not the case, the coded signal may be of very low perceived quality. This can be observed from subjective listening tests where the scores may vary greatly depending on the signal (see e.g. [4]). In [4] it was shown that even when using rate-distortion optimal segmentation [5], constant-amplitude sinusoids do not lead to satisfactory results for critical transient excerpts such as those from SQAM [6]. It was demonstrated that an amplitude modulated (AM) sinusoidal audio coder lead to significant improvements over a sinusoidal coder for such signals. The coder was based on an analysis-by-synthesis parameter estimation procedure using a perceptual distortion measure. As a consequence, the coder suffered from high complexity and delay. Further, it was also shown in [4] that significant improvements are gained by the combination of rate-distortion optimal segmentation and amplitude modulated sinusoidal audio coding, i.e. that model adaptation and flexible segmentation are complementary tools. The rate-distortion optimal segmentation requires that all possible segment lengths at different starting positions be coded. This, of course, adds considerable complexity and delay, which may be prohibitive for some applications. For example, the MPEG-4 Low Delay Audio Coder [7] does not use block switching and minimizes the use of the bit reservoir due to the delay associated with these methods. A powerful and successful method for efficient coding of transients in the context of transform coding is the so-called temporal noise shaping (TNS) [8], which is part of the MPEG-2/4 AAC and is used in the Low Delay Audio Coder instead of block switching [7]. In TNS, a coding gain is achieved for transient signals by predictive coding of transform coefficients, and the optimal linear predictor can be derived efficiently in the frequency domain using standard methods [9]. While linear prediction has found use in predictive coding of speech, it is also a widely used method for spectral estimation of auto-regressive (AR) stochastic processes. Likewise, TNS can be interpreted as either a method for predictive coding in the frequency domain or as an envelope estimator, or, in modulation theoretical terms, a demodulator. Aside from being an envelope estimator, the frequency-domain linear predictor is also an efficient parametric representation of the envelope that can be quantized using well-known methods.

In this paper, we explore an alternative amplitude modulated sinusoidal coding technique based on frequency-domain linear prediction (FDLP) that has considerably lower complexity and delay than [4]. Further, we also analyze the properties of the FDLP-based envelope estimator. We apply the sinusoidal coding in the subbands of a critically sampled filter bank. The advantage of doing this is twofold. Firstly, it has a lower computational complexity than the full-band system and, secondly, it allows for different envelopes in the individual subbands, which may be desirable for some, but by no means all, signals [4, 10]. The sinusoidal parameters are found, given the subband envelopes, using matching pursuit based on a perceptual distortion measure.

The remaining part of this paper is organized as follows: Section 2 contains an overview of the proposed system. The envelope estimator and its properties are presented in Section 3 and subsequently the matching pursuit algorithm used for sinusoidal parameter estimation is treated in Section 4. In Sections 5 and 6 implementation details and experimental results are presented, and, finally, Section 7 concludes on the work.

## 2. SYSTEM OVERVIEW

The method proposed in this paper is implemented in a system that consists of an analysis and a synthesis system. In the analysis system the input signal is first split into $Q$ critically sampled subbands. We here use a uniform filter bank. In each subband, an envelope is estimated using FDLP and the associated parameters are quantized. Given the subband envelopes a number of sinusoidal parameters are extracted and quantized and finally all parameters are then entropy coded. In the synthesis system, the parameters are reconstructed and the subband signals are synthesized using overlap-add. Finally, the signal is reconstructed using a synthesis filterbank that combined with the analysis filterbank has perfect reconstruction. Now, let us

This research was supported by the ARDOR (Adaptive Rate-Distortion Optimized sound codeR) project, EU grant no. IST–2001–34095, and the Intelligent Sound project, Danish Technical Research Council grant no. 26–02–0092.

introduce some definitions and the notation. First, we define the subband signal for subband $q$ as $x_q(n) = \sum_{m=0}^{M-1} h_q(m)x(n-m)$ for $n = 0, \dots, N-1$ with $h_q(n)$ being the impulse response of the $q$th analysis filter of length $M$. From these subband signals, the input signal can be reconstructed (with a delay $d$) as $\sum_{q=1}^{Q}\sum_{m=0}^{M-1} g_q(m)$-$x_q(n-m) = x(n-d)$ using the impulse responses of the synthesis filters $g_q(n)$. In the envelope estimation and in the sinusoidal parameter estimation, we will make use of the so-called discrete-time analytic signal for a particular segment and subband $q$. For $n = 0, \dots, K-1$ with $K = N/Q$ this is defined as

$$a_q(n) = x_q(Qn) + j\mathcal{H}\left\{x_q(Qn)\right\}, \qquad (1)$$

where $\mathcal{H}\{\cdot\}$ denotes the Hilbert transform. Here we have assumed that the segment length $N$ is an integer multiple of $Q$. Note that the calculation of the analytic signal may be integrated into the filterbank implementation since $h_q(n) * (x(n) + j\mathcal{H}\{x(n)\}) = x(n) * (h_q(n) + j\mathcal{H}\{h_q(n)\})$. The model of the analytic subband signal used in this paper can be written as the following sum of sinusoids where the sinusoids are modulated by the amplitude modulating signal $\hat{\gamma}_q(n)$,

$$\hat{a}_q(n) = \sum_{l=1}^{L_q} \hat{\gamma}_q(n) A_{q,l} e^{j\omega_{q,l}n + j\phi_{q,l}}, \qquad (2)$$

where each sinusoid is characterized by a frequency $\omega_{q,l}$, a phase $\phi_{q,l}$ and an amplitude $A_{q,l}$. This signal model is built using a matching pursuit algorithm [11] based on the perceptual distortion measure presented in [12] and a redundant dictionary consisting of modulated complex sinusoidal components. This dictionary can be seen as being signal-adaptive since the amplitude modulating signal $\hat{\gamma}_q(n)$ varies with the signal over time and subbands.

In the decoder, the subband signal is recovered by taking the real-value of (2), upsampling the signal by a factor of $Q$ and subsequent filtering by the synthesis filter $g_q(n)$.

## 3. ENVELOPE ESTIMATION

In this section we briefly present the main results of the envelope estimator based on the FDLP principle first introduced in [8] as temporal noise shaping for transform coding. Further, we also provide some additional analysis of the properties of this estimator. It must be emphasized that the application of FDLP considered here is fundamentally different from that of [8]. First, we define the Fourier transform of the analytic signal for $k = 0, \dots, K-1$ as

$$A_q(k) = \sum_{n=0}^{K} a_q(n) e^{-j2\pi\frac{k}{K}n}. \qquad (3)$$

We then write the frequency-domain prediction error $E_q(k)$ as a linear combination of $A_q(k)$ having the $I$ (complex) prediction coefficients $b_i \in \mathbb{C}$:

$$E_q(k) = A_q(k) - \sum_{i=1}^{I} b_i A_q(k-i). \qquad (4)$$

The optimal prediction coefficients are then found in such a way that the squared prediction error is minimized, i.e.,

$$\{b_i\} = \operatorname{argmin} \sum_{k=0}^{K-1} \left|A_q(k) - \sum_{i=1}^{I} b_i A_q(k-i)\right|^2, \qquad (5)$$

which can be solved efficiently using well-known methods [9]. Then, by taking the Fourier transform of the squared instantaneous envelope, we get the spectral autocorrelation sequence estimate $C_q(\tau)$:

$$\sum_{n=0}^{K-1} |a_q(n)|^2 e^{-j2\pi\frac{\tau}{K}n} = \frac{1}{K}\sum_{k=0}^{K-1} A_q(k) A_q^*(k-\tau) \qquad (6)$$

$$= C_q(\tau), \qquad (7)$$

from which the prediction coefficients also can be found. Taking the inverse Fourier transform of both sides of (4), we get

$$e_q(n) = a_q(n)\left[1 - \sum_{i=1}^{I} b_i e^{j2\pi\frac{i}{K}n}\right]. \qquad (8)$$

Rearranging this, we get the (complex) envelope estimate for $n = 0, \dots, K$,

$$\hat{\gamma}_q(n) = \frac{a_q(n)}{e_q(n)} = \frac{1}{1 - \sum_{i=1}^{I} b_i e^{j2\pi\frac{i}{K}n}}. \qquad (9)$$

Specifically, the squared instantaneous envelope estimate is

$$|\hat{\gamma}_q(n)|^2 = \frac{1}{|1 - \sum_{i=1}^{I} b_i e^{j2\pi\frac{i}{K}n}|^2}. \qquad (10)$$

As the prediction filter is minimum-phase, the phase of $\hat{\gamma}_q(n)$ can be determined uniquely from $\log|\hat{\gamma}_q(n)|$ since they form a Hilbert transform pair, i.e.,

$$\angle\hat{\gamma}_q(n) = \mathcal{H}\{\log|\hat{\gamma}_q(n)|\}. \qquad (11)$$

Using Parseval's theorem, we can write the minimization in (5) of the sum of the squared prediction in the time domain:

$$\min \sum_{k=0}^{K-1} |E_q(k)|^2 = \min \sum_{n=0}^{K-1} \frac{|a_q(n)|^2}{|\hat{\gamma}_q(n)|^2}. \qquad (12)$$

From these equations, we can make a number of interesting observations. From (12) we see that minimizing the prediction error a least-squares fashion corresponds to minimizing the sum of the ratio between the squared instantaneous envelope and the estimate. The frequency-domain linear predictor models the time-domain envelope in exactly the same way as the time-domain linear predictor models the spectral envelope, and, hence, they share the same properties and suffer from the same problems (e.g. the cancellation of errors and overemphasis on peaks [9]). One notable property is that the envelope estimate converges to the squared instantaneous envelope as the model order $I$ grows. From (7) and (12) we see that a decorrelation of the transform coefficients results in a flattening of the squared instantaneous envelope since $C_q(\tau) = 0$ for $\tau \neq 0$ implies a flat envelope. It can easily be shown that the squared instantaneous envelope of two sinusoids that are modulated by the same signal contains cross-terms that are due to the sinusoidal carriers (see [10]). In sinusoidal modeling these cross-terms are modeled by the sinusoids and should hence not be captured by the envelope estimator. For sinusoids that are well-separated in frequency these cross-terms will occur as long-term correlation in $C_q(\tau)$, and hence FDLP has the (at least in this particular application) undesirable property that it will seek to model these cross-terms. Consequently, the model order should be chosen sufficiently low such that this does not happen and this order cannot, contrary to common practice in transform coding, simply be chosen from the prediction gain. Moreover, the envelope estimator will fail when the sinusoids and the modulating signal are not well-separated in frequency (since Bedrosian's theorem [13] does not hold in this case) or when the sinusoids are closely spaced [10].

## 4. SUBBAND MATCHING PURSUIT

The individual sinusoidal parameters, i.e. frequencies, phases and amplitudes, are found in each subband using a psychoacoustic matching pursuit [11] given the subband envelopes $\hat{\gamma}_q(n)$. The subband envelope adapts the dictionary to the subband signal, and, as a consequence, a higher rate of convergence, in terms of the distortion as a function of the number of components, can be achieved. In each iteration, with $i$ being the iteration index, the algorithm operates on the $F$ point Fourier transform of the subband residuals $R_{q,i}(k)$ which are initialized for $k = 0, \ldots, F-1$ as

$$R_{q,1}(k) = \sum_{n=0}^{K-1} w(n) a_q(n) e^{-j2\pi\frac{k}{F}n}, \qquad (13)$$

with $w(n)$ being the analysis/synthesis window. The algorithm finds in each iteration the subband and the parameters that minimize the weighted squared absolute value of the Fourier transform of the residual, i.e., $D_{q,i} = \sum_{k=0}^{F-1} P_q(k)|R_{q,i}(k)|^2$, where $P_q(k)$ is a perceptual weighting function for the frequency region associated with subband $q$. This weighting function is derived, for each segment, from the auditory masking model presented in [12]. The combination of frequency (index) $\hat{f}$ and subband $\hat{q}$ that minimizes the perceptual distortion are chosen as:

$$\{\hat{q}, \hat{f}\} = \underset{\{q,m\}}{\operatorname{argmax}} \frac{|\Psi_q(m)|^2}{\Phi_q(m)}, \qquad (14)$$

with the numerator containing the inner product

$$\Psi_q(m) = \sum_{k=0}^{F-1} P_q(k) Z_q^*(k-m) R_{q,i}(k), \qquad (15)$$

and the denominator the norm

$$\Phi_q(m) = \sum_{k=0}^{F-1} P_q(k) Z_q^*(k-m) Z_q(k-m). \qquad (16)$$

$Z_q(k)$ is defined as the Fourier transform of the windowed subband envelope, i.e.,

$$Z_q(k) = \sum_{n=0}^{K-1} w(n) \hat{\gamma}_q(n) e^{-j2\pi\frac{k}{F}n}. \qquad (17)$$

The optimum phase and amplitude associated with the estimated complex sinusoid of frequency $\hat{f}$ in subband $\hat{q}$ can be found as

$$A_{\hat{q},i} e^{j\phi_{\hat{q},i}} = \frac{\Psi_{\hat{q}}(\hat{f})}{\Phi_{\hat{q}}(\hat{f})}. \qquad (18)$$

Finally, having found the subband, frequency, phase and amplitude we update the Fourier transform of that subband residual as

$$R_{\hat{q},i+1}(k) = R_{\hat{q},i}(k) - A_{\hat{q},i} e^{j\phi_{\hat{q},i}} Z_{\hat{q}}(k-\hat{f}). \qquad (19)$$

Note how the numerators of equations (18) and (14) contain the same inner product. It can be seen from the following that, like for the constant-amplitude case treated in [11], these inner products can be efficiently computed for different $m$ using FFTs:

$$\Psi_q(m) = \sum_{n=0}^{K-1} v_q(n) w(n) \hat{\gamma}_q^*(n) e^{-j2\pi\frac{m}{F}n}, \qquad (20)$$

with $v_q(n) = \sum_{k=0}^{F-1} P_q(k) R_{q,i}(k) e^{j2\pi\frac{k}{F}n}$. Similarly, the denominator of equations (18) and (14) can be found using Fourier transforms:

$$\Phi_q(m) = \sum_{n=0}^{K-1} \left[ \sum_{k=0}^{F-1} |Z_q(k)|^2 e^{-j2\pi\frac{k}{F}n} \right] p_q(n) e^{-j2\pi\frac{m}{F}n}, \qquad (21)$$

with $p_q(n)$ being the inverse Fourier transform of the perceptual weighting function, i.e., $p_q(n) = \sum_{k=0}^{F-1} P_q(k) e^{j2\pi\frac{k}{F}n}$. Finally, we note that since the subband signals are orthogonal, the total distortion is simply the sum of the subband distortions and can hence be subject to rate-distortion optimization. It follows from the perfect reconstruction of the filter bank and the convergence of the matching pursuits in the subbands that the entire system will converge.

## 5. IMPLEMENTATION DETAILS

In assessing the improvement that the higher update-rate of the amplitude, i.e. amplitude modulation, results in, we compare two different configurations of the proposed system: The first uses only constant-amplitude sinusoids (denoted CA) while the second uses multiband amplitude modulation with Q=8 (denoted AM). The optimal segment length has been determined empirically by informal listening tests to be about 35 ms for the CA configuration using a von Hann window with 50% overlap. This segment length was also used for the AM configuration. The frequencies and amplitudes are quantized using the logarithmic quantizer described in [4] while the phases are quantized uniformly using 5 bits. At average this results in approximately 15 bits per sinusoidal component. The complex prediction coefficients were quantized by mapping the reflection coefficients to the log-area ratios which were then quantized uniformly. The associated rate has been estimated from the entropy of the quantization indices, which resulted in approximately 9 bits per complex prediction coefficient, and a 5th order complex prediction filter was used in the simulations. Both configurations were set to run at 30 kbps. In order to achieve efficient coding of stationary sinusoids, the envelope is not used when it has a correlation coefficient of more than 90% with a constant envelope. Alternatively, the rate-distortion optimization-based coder switching architecture proposed in [4] can be used for this. Subband FFTs of 1024 points were used for the AM configuration while the CA configuration uses 8192 point FFTs in the matching pursuit. Further, we have used an FFT-based implementation of the Hilbert transform. In practice we have found the orthogonality between subbands not to be of critical importance for the task at hand. Hence, we have used a uniform 8-band pseudo QMF filter bank for the AM configuration with a prototype filter of length 512.

## 6. RESULTS AND DISCUSSION

In Figure 1 we illustrate the order selection problem of the envelope estimator for two sinusoids $2\cos(2\pi 0.1n) + \sin(2\pi 0.11n + \pi/3)$. It can be seen that the FDLP models the cross-terms that are due to the carriers and that the low-order model is better than the high-order model when cross-terms are present.

The subband processing has been verified to produce good results compared to a full-band system using constant-amplitude sinusoids, i.e. no AM. This verifies that the subband processing does not introduce any noticeable artifacts, although it has some inherent drawbacks. There is a significant reduction of complexity of the subband system compared to the full-band system. Where the full-band

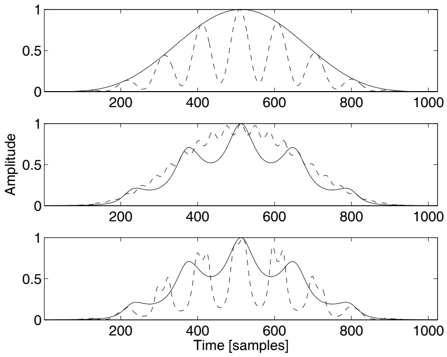

**Fig. 1**. Illustration of the problem of cross-terms in the squared instantaneous envelope for multiple sinusoids. The top panel shows the true squared envelope (solid) and the squared instantaneous envelope (dashed). The squared instantaneous envelope of one sinusoid estimated using a 5th order predictor (solid), and using a 25th order predictor (dashed) are shown in the middle panel. And similarly, in the bottom panel, for two sinusoids estimated using a 5th order predictor (solid), and using a 25th order predictor (dashed).

| Statistic | AM | CA | Anchor 1 | Anchor 2 | HR |
|---|---|---|---|---|---|
| Mean | 62 | 54 | 29 | 54 | 95 |
| Conf. (±) | 7.4 | 6.6 | 4.0 | 5.8 | 2.8 |

**Table 1**. Results of MUSHRA test. Scores for all excerpts and listeners (means and 95% confidence intervals).

system would require FFTs of size $F$, the subband system requires FFTs of size $F/Q$ and only one subband has to be updated per iteration of the matching pursuit.

As proof of concept of the proposed method, we use a subjective listening test. Specifically, we use the MUSHRA test [14] for quantifying the improvements of the AM configuration compared to the CA configuration. For each excerpt, the listeners were asked to rank 5 differently processed versions relative to a known reference on a score from 0 to 100. These included the hidden reference (denoted HR), an anchor low-pass filtered at 3.5 kHz and an anchor low-pass filtered at 7 kHz (denoted Anchors 1 and 2). The remaining two versions were the different configurations, AM and CA. The excerpts were 5 different critical 10 s transient (mono) signals from SQAM [6], namely castanets, claves, glockenspiel, triangle and xylophone. 12 inexperienced listeners participated and the test was conducted using headphones. In Table 1 the results of the listening test are shown. As can be seen from the anchors, the sometimes large confidence intervals can largely be attributed to variations over the excerpts and listeners. Testing instead for the differences in scores, the mean of the difference between the AM and the CA configuration was found to be 8.3 with a 95% confidence interval of $\pm 7.4$. Since the confidence interval does not include zero, we conclude that the AM configuration performs significantly better than the CA configuration. In interpreting the results it should noted that due to the temporal integration in the human auditory system, events of a short duration, such as onsets, may only result in small improvements in scores. We also note that, as shown in [4], both the CA and AM configurations may be further improved and even combined using rate-distortion optimal segmentation and allocation [5] at the cost of significantly increased delay and complexity.

## 7. CONCLUSION

We have presented a method for amplitude modulated sinusoidal audio coding based on frequency-domain linear prediction for estimation and efficient coding of time-domain envelopes. This has been found, in a subjective listening test, to improve on sinusoidal coding for critical transient signals. Further, the properties of the envelope estimator have been analyzed and it has been demonstrated that special care must be taken in selecting the model order.

## 8. REFERENCES

[1] B. Edler, H. Purnhagen, and C. Ferekidis, “ASAC – Analysis/Synthesis Audio Codec for Very Low Bit Rates,” in *100th Conv. Aud. Eng. Soc.*, 1996, paper preprint 4179.

[2] ISO/IEC, *Coding of audio-visual objects – Part 3: Audio (MPEG-4 Audio Edition 2001)*. ISO/IEC Int. Std. 14496-3:2001, 2001.

[3] E. G. P. Schuijers, A. W. J. Oomen, A. C. den Brinker, and J. Breebart, “Advances in parametric coding for high-quality audio,” in *114th Conv. Aud. Eng. Soc.*, 2003, paper preprint 5852.

[4] M. G. Christensen and S. van de Par, “Efficient parametric coding of transients,” *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14(4), July 2006, to appear.

[5] P. Prandoni and M. Vetterli, “R/D optimal linear prediction,” *IEEE Trans. Speech and Audio Processing*, pp. 646–655, 8(6) 2000.

[6] European Broadcasting Union, *Sound Quality Assessment Material Recordings for Subjective Tests*. EBU, Apr. 1988, Tech. 3253.

[7] E. Allamanche, R. Geiger, J. Herre, and T. Sporer, “MPEG-4 Low Delay Audio Coding based on the AAC Codec,” in *106th Conv. Aud. Eng. Soc.*, May 1999, paper preprint 4929.

[8] J. Herre and J. D. Johnston, “Enhancing the performance of perceptual audio coders by using temporal noise shaping (TNS),” in *101st Conv. Aud. Eng. Soc.*, 1996, paper preprint 4384.

[9] J. Makhoul, “Linear Prediction: A Tutorial Review,” *Proc. IEEE*, vol. 63(4), pp. 561–580, Apr. 1975.

[10] M. G. Christensen, S. V. Andersen, and S. H. Jensen, “Amplitude modulated sinusoidal models for audio modeling and coding,” in *Knowledge-Based Intelligent Information and Engineering Systems*, ser. Lecture Notes in Artificial Intelligence, V. Palade, R. J. Howlett, and L. C. Jain, Eds. Springer-Verlag, 2003, vol. 2773, pp. 1334–1342.

[11] R. Heusdens, R. Vafin, and W. B. Kleijn, “Sinusoidal modeling using psychoacoustic-adaptive matching pursuits,” *IEEE Signal Processing Lett.*, vol. 9(8), pp. 262–265, Aug. 2002.

[12] S. van de Par, A. Kohlrausch, R. Heusdens, J. Jensen, and S. H. Jensen, “A perceptual model for sinusoidal audio coding based on spectral integration,” vol. 9, pp. 1292–1304, 2005.

[13] E. Bedrosian, “A product theorem for Hilbert transforms,” in *Proc. IEEE*, vol. 55(1), May 1963, pp. 868–869.

[14] *ITU-R BS.1534*, ITU Method for subjective assessment of intermediate quality level of coding system, 2001.