MULTIBAND AMPLITUDE MODULATED SINUSOIDAL AUDIO MODELING

Mads Græsbøll Christensen[†]*, Steven van de Par*[‡]*, Søren Holdt Jensen*[†]*, and Søren Vang Andersen*[†]

[†] Dept. of Communication Technology
Aalborg University, Denmark
{mgc,shj,sva}@kom.auc.dk

[‡] Philips Research Laboratories
Eindhoven, The Netherlands
steven.van.de.par@philips.com

ABSTRACT

In this paper, we investigate the importance of taking frequency-dependent temporal phenomena into account in audio coding. We do this in the context of sinusoidal modeling of audio signals by applying amplitude modulation to the sinusoidal components. Traditionally, audio coders use a fixed time-segmentation for all frequencies despite that it is well-known that the time-frequency resolution of the human auditory system is not constant. The well-known window switching is an example of this. We compare multiband amplitude modulated sinusoidal models to a singleband model using different audio excerpts. Based on both comparative listening tests and a psychoacoustical distortion measure it is concluded that an improvement is generally gained using multiband amplitude modulation, although specific single sources are well-modeled using a singleband model.

1. INTRODUCTION

A well-known problem in perceptual audio coding and modeling is what is known as pre-echo distortion or pre-echos (see e.g. [1]). Pre-echos can be defined as the introduction of a modeling error or quantization error that occurs before a transient signal. These occur in block-based modeling when there is an onset or attack at the end of a segment.

The importance of pre-echo control in audio coding and modeling can be understood by considering the temporal masking properties of the human auditory system. In audio coding the original signal serves as a masker of the error-signal. This masking is very effective when the error-signal is simultaneous with, or directly follows the masker. However, when the error-signal preceeds the masker, very little masking is observed. This is depicted in Figure 1 showing masking thresholds as a function of time. Pre-masking can be measured to typically last only about 20 ms, whereas post-masking can last longer than 100 ms [2]. Trained listeners, however, may exhibit little or no pre-masking except for very short signals [3]. This means that any artifacts introduced before an onset are very poorly masked compared to the situation where a signal is present. The point that motivates this work is that masking phenomena occur on a critical band basis. Singleband techniques such as window switching [4] or AM as used in [5] do not take this into account. What may happen is that the choice of window length or the estimation of the amplitude modulating signal may be dominated by a stationary low-frequency component while a transient occurs at high frequencies, whereby audible artifacts are caused. Or it may happen that a short window is chosen because of some high-frequency transient while stationary low-frequency parts may suffer because of the decreased frequency resolution.

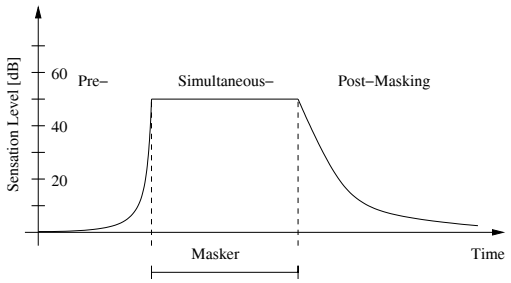

Fig. 1. Temporal or nonsimultaneous masking properties. Pre-masking occurs before the onset of the masker and post-masking occurs afterwards (after [2]).

In sinusoidal coding of speech [6] and audio, fixed segmentation for all frequencies has also been the standard solution, although multiresolution sinusoidal modeling was considered in [7]. Rate-distortion optimal time-segmentation [8] leads to an improved sinusoidal modeling, but still provides only a partial solution because a) the segmentation is still fixed over frequency and b) the minimum segment size is constrained because of the computational complexity involved in finding the optimum segmentation. Also, the use of overlap between segments inevitably smears any modeling error into neighboring frames.

In [9] amplitude modulated sinusoidal models for audio modeling and coding were introduced and in this paper we build further on this work. We achieve frequency-dependent temporal modeling using multiband amplitude

This work was conducted within the ARDOR project, EU grant no. IST–2001–34095.

modulation, where different amplitude modulating signals are used at different frequencies. Amplitude modulated sinusoidal models for audio modeling and coding are attractive for modeling of transient phenomena because constant-amplitude sinusoidal models converge slowly in terms of rate-distortion for transient signals thus performing badly for low bit-rates.

The paper is organized as follows. In Section 2 the amplitude modulated sinusoidal analysis-synthesis system is presented. This includes two parts, namely estimation of amplitude modulating signals and estimation of the parameters of the sinusoidal carriers. In Section 3 the multiband model is compared to the singleband model by listening tests as well as a perceptual distortion measure. Finally, conclusions about the work are presented in Section 4.

2. AM SINUSOIDAL ANALYSIS-SYNTHESIS

We use an amplitude modulated sinusoidal model, that looks as follows:

$$\hat{x}(n) = \sum_{q=1}^{Q} \gamma_q(n) \sum_{l=1}^{L_q} A_{l,q} \cos(\omega_{l,q} n + \phi_{l,q}), \qquad (1)$$

where $\gamma_q(n)$ is the amplitude modulating signal in the q'th subband and L_q is the number of sinusoids in that subband. $\omega_{l,q}$, $A_{l,q}$ and $\phi_{l,q}$ are the frequencies, amplitudes and phases of the sinusoids. We distinguish between a singleband model ($Q = 1$) and a multiband model ($Q > 1$).

The task is now to find $\gamma_q(n)$ for each subband. We start the estimation, which is based on [9], by splitting the input signal into subbands using the perfect reconstruction non-uniform filterbank described in [10]. Then we have for each subband a signal $x_q(n)$ and a model of that signal $\hat{x}_q(n)$. The instantaneous envelope of the model can then easily be shown to be

$$|\hat{x}_{q,c}(n)|^2 = \sum_{l=1}^{L_q} \sum_{k=1}^{L_q} \gamma_q^2(n) A_{q,l} A_{q,k} \times \exp(j(\phi_{q,k} - \phi_{q,l})) \exp(j(\omega_{q,k} - \omega_{q,l})n), \qquad (2)$$

with subscript c denoting the analytic signal (see e.g. [9]). The squared instantaneous envelope is thus composed of a set of auto-terms ($l = k$) that identifies the amplitude modulating signal and a set of interfering cross-terms ($l \neq k$). From this it can be seen that the frequencies of these cross-terms in the instantaneous envelope is given by the distances between the sinusoidal components. Thus, the lowest frequency in the squared instantaneous envelope caused by the interaction of the sinusoids is given by the minimum distance between two adjacent sinusoids.

These cross-terms can be reduced by constraining the minimum distance between sinusoids and then lowpass filter the squared instantaneous envelope of the input signal as shown in [9], i.e.

$$\gamma_q^2(n) = \alpha e_q^2(n) * h_{LP}(n), \qquad (3)$$

where $e_q^2(n) = x_q^2(n) + \mathcal{H}\{x_q(n)\}^2$ with $\mathcal{H}\{\cdot\}$ denoting the Hilbert transform. Moreover, α is a positive scaling factor and $h_{LP}(n)$ is the impulse response of an appropriate lowpass filter with a stopband frequency below half the minimum distance between two sinusoids, i.e.

$$2BW < \min_{l \neq k} |\omega_{q,l} - \omega_{q,k}|. \qquad (4)$$

For quasi-harmonic (pitched) sounds such as voiced speech, this spacing is simply the fundamental frequency. For a discussion on design issues regarding this filter see e.g. [9, 5].

Given that the amplitude modulating signal has been estimated for the q'th subband, we can then find the sinusoidal carriers of the subband (note that for convenience we now change the notation from indexing by subband to indexing by iteration). These are found by applying the estimated amplitude modulating signals to an overcomplete dictionary containing complex sinusoids resulting in a subband dictionary $\mathcal{D}_q$ containing atoms $g_{k,q}(n)$. We then perform matching pursuit [11], where in each iteration the maximizer of the normalized inner product between the atom and the residual is chosen, i.e.

$$\mathbf{g}_{i,q} = \arg \max_{g_{k,q} \in \mathcal{D}_q} \frac{| < \mathbf{g}_{k,q}, \mathbf{r}_{i,q} > |^2}{\|\mathbf{g}_{k,q}\|_2^2}, \qquad (5)$$

where $\mathbf{g}_{k,q} = [g_{k,q}(0) \ \ldots \ g_{k,q}(N-1)]^T$ and $\mathbf{r}_{i,q} = [r_{i,q}(0) \ \ldots \ r_{i,q}(N-1)]^T$ with $r_{i,q}(n)$ being the residual of the i'th iteration. Now, writing out the inner product using the AM model, we get

$$< \mathbf{g}_{k,q}, \mathbf{r}_{i,q} > = \sum_{n=0}^{N-1} \gamma_q(n) \exp(-j\omega_k n) r_{i,q}(n). \qquad (6)$$

It can be seen, that by defining $\tilde{r}_{i,q}(n) = \gamma_q(n) r_{i,q}(n)$, the greedy estimation can be carried out efficiently using an FFT of $\tilde{r}_{i,q}(n)$. In a similar way, we can apply the window $w(n)$ twice to the input and find the solution using an FFT, whereby the error is minimized in a weighted least-squares sense, i.e.

$$\min \sum_{n=0}^{N-1} w^2(n) r_{i+1}^2(n) = \min \sum_{n=0}^{N-1} (w(n) c_k g_{k,q}(n) - w(n) r_{i,q}(n))^2, \qquad (7)$$

where c_k is the coefficient (phase and amplitude in this case) of the k'th atom. This causes not only the input but also the model to be weighted. This takes the use of windowing in both analysis and synthesis into account.

Equation (5) minimizes only the subband residual. When minimizing over the entire signal, we simply pick the maximum of the spectral subband maxima. This leads to the iterative (i being the iteration index) FFT-based algorithm (the FFT is denoted $\mathcal{FFT}\{\cdot\}$) described below, where the frequencies, phases and amplitudes of the sinusoidal model are found. We initialize the residuals with $r_{1,q}(n) = x_q(n)\ \forall q$.

1. Find subband

$$q_i = \arg\max_q \left(\frac{|\mathcal{FFT}\{\gamma_q(n)w^2(n)r_{i,q}(n)\}|^2}{\sum_{n=0}^{N-1} \gamma_{q_i}^2(n)w^2(n)} \right)$$

and corresponding frequency

$$\omega_i = \arg\max_\omega |\mathcal{FFT}\{\gamma_{q_i}(n)w^2(n)r_{i,q_i}(n)\}|^2.$$

2. Estimate phase and amplitude by the inner product:

$$c_i = \frac{\sum_{n=0}^{N-1} r_{q_i,i}(n)w^2(n)\gamma_{q_i}(n)\exp(-j\omega_i n)}{\sum_{n=0}^{N-1} \gamma_{q_i}^2(n)w^2(n)},$$

which can be found from the subband FFT.

3. Generate new subband residual:

$$r_{i+1,q_i}(n) = r_{i,q_i}(n) - 2\gamma_{q_i}(n)|c_i|\cos(\omega_i n + \angle c_i).$$

This procedure is continued until some stopping criterion is reached. Although the estimation procedure is dependent on the amplitude modulating signal $\gamma_q(n)$, the algorithm still converges if we restrict $\gamma_q(n)$ to be strictly positive. Hereby the subband dictionaries $\mathcal{D}_q$ still form overcomplete bases and the algorithm converges on a subband level [11] and because of the perfect reconstruction filterbank, the entire system converges.

The above algorithm can be implemented much more efficiently than in the form above. The FFTs of the individual subbands and their maxima can be computed once at initialization. Then, in each iteration we only need to update the FFT of the subband residual and find the spectral maximum of it. The search in step 1 then reduces to searching among the Q spectral maxima.

3. EXPERIMENTAL RESULTS

The importance of multiband temporal modeling has been investigated using both listening tests in the form of AB preference tests as well as an objective distortion measure. We compare the singleband model ($Q = 1$) to the multiband model ($Q > 1$).

Settings			
	Value		
Parameter	ABBA	GLCK	SPCH
Sampl. freq. [kHz]	44.1	44.1	8
Filterbank order	200	200	200
Filters	25	12	5
LP Filter order	100	100	100
Sinusoids	40	40	40
Cutoff freq. [Hz]	100	500	25

Table 1. Parameter values for different excerpts.

The excerpts used in the tests are: glockenspiel (GLCK), ABBA (ABBA), and Danish female speech (SPCH). They are all mono signals and have a length in the range of 5-10 s. These represent very different signal types from single source signals to complex music containing multiple sources.

The settings of the sinusoidal analysis-synthesis system for the different excerpts are shown in Table 1. In all cases a segment size of 20 ms and overlap-add with a 50% overlap von Hann window was used. Also, the FFT size was 8192. For the demodulation filter (3), we use an FIR filter designed using the window method (Hamming window).

In Table 2 the results of the AB preference tests are listed for the individual excerpts. 9 experienced listeners were used. It can be seen clearly, that there is a strong preference for the multiband model in the two cases, where the signals contain several sources, namely ABBA and glockenspiel, whereas for the case of speech, the preference tends toward equal. Significance has been determined by a small-sample case sign test (binomial distribution) using a 0.05 level of significance.

Results of Listening Tests			
	Preference		
Excerpt	Singleband	Multiband	Significant
ABBA	11%	89%	Yes
GLCK	11%	89%	Yes
SPCH	56%	44%	No

Table 2. Results of AB-preference tests.

Also, the results were verified using an objective measure. As a suitable perceptual model that also include temporal masking phenomena, we used the Dau et al. model [12]. This model consists of a filterbank that resembles critical band filtering, followed by an inner-haircell model and adaptation loops which account for the temporal masking that occurs in the auditory system. The resulting internal representation is low-pass filtered and used for a perceptual distortion prediction by calculating the mean squared difference between the internal representations of the original

Results using the Dau et al. Model		
	Distortion	
Excerpt	Singleband	Multiband
ABBA	7270	5431
GLCK	931	644
SPCH	412	426

Table 3. Distortions calculated using the Dau et al. model.

and modified signal. The distortions are listed in Table 3. The Dau et al. model confirmed the results of the listening tests with the multiband model outperforming the singleband model in two first cases while the difference for the speech is very small. The overall distorition is highest for the ABBA excerpt, because it is a complex signal, whereas the total distortion is lowest for the speech signal, due to its limited bandwidth. That there is a slightly higher distortion for the multiband model for the case of SPCH can be attributed to the additional processing of the multiband system and the shape of the filters of the filterbank.

The conclusion is that for particular single sources such as speech, the singleband model performs very well. This is in line with [5], where also members of brass, woodwind, and string instrument families are mentioned as sources being well modeled by the singleband model. For more complex signals such as superpositions of multiple sources, there is a great need for multiband modeling and coding, which is clearly indicated by the high preference for multiband modeling of ABBA.

That the singleband model works well for single sources is an indication that the model in Eq. (1) can indeed form the basis of compression not only in terms of subbands, but also in terms of sinusoids sharing an amplitude modulating signal, i.e. by a decomposition into sources.

4. CONCLUSION

In this paper, we have investigated the need for taking temporal phenomena in audio modeling and coding into account in a way that is frequency dependent. This has been done in the context of sinusoidal modeling, where we have applied amplitude modulation in order to achieve better temporal modeling. We have presented a multiband sinusoidal analysis-synthesis system that utilizes amplitude modulation to achieve frequency dependent temporal resolution. Finally, we have compared this multiband model to a commonly used singleband model and it has been demonstrated using both an objective perceptual distortion measure as well as listening tests, that significant improvements are achieved by this for complex signals containing multiple sources such as general audio, and that the singleband model performs very well for particular single sources.

5. REFERENCES

[1] T. Painter and A. S. Spanias, "Perceptual Coding of Digital Audio," in *Proc. IEEE*, Apr. 2000, vol. 88(4).

[2] E. Zwicker and H. Fastl, *Pyschoacoustics - Facts and Models*, Springer, 2nd edition, 1999.

[3] B. C. J. Moore, "Masking in the human auditory system," in *Collected Papers on Digital Audio Bit-Rate Reduction*, N. Gilchrist and C. Grewin, Eds. 1996, Audio Eng. Soc.

[4] B. Edler, "Codierung von audiosignalen mit überlappender transformation und adaptiven fensterfunktionen," in *Frequenz*, 1989, pp. 1033–1036.

[5] E. B. George and M. J. T. Smith, "Analysis-by-synthesis/overlap-add sinusoidal modeling applied to the analysis-synthesis of musical tones," in *J. Audio Eng. Soc.*, 1992, vol. 40(6).

[6] R. J. McAulay and T. F. Quatieri, "Sinusoidal coding," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds., chapter 4. Elsevier Science B.V., 1995.

[7] S. N. Levine, T. S. Verma, and J. O. Smith III, "Alias-free, multiresolution sinusoidal modeling for polyphonic, wideband audio," in *Proc. IEEE Workshop Appl. of Signal Process. to Aud. and Acoust.*, 1997.

[8] R. Heusdens and S. van de Par, "Rate-distortion optimal sinusoidal modeling of audio using psychoacoustical matching pursuits," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2002.

[9] 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*, V. Palade, R. J. Howlett, and L. C. Jain, Eds., vol. 2773 of *Lecture Notes in Artificial Intelligence*, pp. 1334–1342. Springer-Verlag, 2003.

[10] M. M. Goodwin, "Nonuniform filterbank design for audio signal modeling," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1997.

[11] S. Mallat and Z. Zhang, "Matching pursuit with time-frequency dictionaries," in *IEEE Trans. Signal Processing*, Dec. 1993, vol. 40.

[12] T. Dau, D. Püschel, and A. Kohlrausch, "A quantitative model of the effective signal processing in the auditory system. i. model structure," in *J. Acoust. Soc. Am.*, June 1996, vol. 99(6).