MODIFYING TRANSIENTS FOR EFFICIENT CODING OF AUDIO

Renat Vafin[1], *Richard Heusdens*[2], *W. Bastiaan Kleijn*[1]

[1]Department of Speech, Music and Hearing
KTH (Royal Institute of Technology)
S-10044 Stockholm, Sweden
{renat, bastiaan}@speech.kth.se

[2]Department of Mediamatics
Delft University of Technology
2628 CD Delft, The Netherlands
R.Heusdens@its.tudelft.nl

ABSTRACT

In this paper, we propose a method for efficient representation of transients in audio signals. We estimate the transient component of an original audio signal and modify the locations of the transients in such a way that the transients can occur only at locations defined by a relatively coarse time grid. This procedure allows an efficient representation of transients with damped sinusoids. We also verify that the introduced modifications do not result in a perceptual difference between the original and the modified audio signals.

1. INTRODUCTION

Parametric coding of audio is a popular tool for representing audio signals at very low bit rates [1, 2, 3, 4]. In a parametric audio coder, a signal is represented by a source model, and parameters of the model are estimated and encoded. A popular parametric representation of audio signals is based on a decomposition of an original signal into three components: a transient component, a tonal (sinusoidal) component, and a noise component. Each component is then represented by a corresponding set of parameters (e.g., [1, 3, 4]). Having a dedicated model for the transient component proved to be beneficial for parts of audio signals with sharp attacks, because sinusoidal and noise models cannot represent those perceptually important events efficiently [5].

We propose a method that improves the efficiency of representing the transient component of an audio signal. In [6] it is shown that transients can be modeled efficiently using sinusoids with exponentially-modulated amplitudes. (Below we refer to such sinusoids as damped sinusoids. However, the damping coefficient can be any real number, and positive values correspond to increasing amplitudes rather than to truly decreasing amplitudes.) In that work, an audio signal is analyzed on a segment-by-segment basis, and each segment is represented as a sum of damped sinusoids. A problem occurs when a transient starts in the middle of a segment. Compared to the case where a transient starts in the beginning of a segment, the number of damped sinusoids needed to model the transient well increases considerably. If a transient is not modeled properly, the modeling error is distributed over the whole segment resulting in audible pre-echoes. Different ways of solving this problem have been used:

- Allow a one-sample-precision (full-precision) variable segmentation of the signal, such that transients will always start in the beginning of segments (e.g., [1]).
- Allow a switching between a long and a short window defining analysis segments, such that short windows are used for parts of an audio signal with sharp attacks (e.g., MPEG-1 Layer III audio coding algorithm [7]). In this case, the segmentation is defined simply by the lengths of the long and the short windows.

In this paper, we use a restricted time segmentation. By restricted segmentation we mean that the segments are defined by integer multiples of a predefined minimum segment size, say 5 ms. Given such a restricted time segmentation, we modify the transient component of the audio signal such that transients can start only at the beginning of the segments. This will result in an efficient representation of transients with damped sinusoids. Our approach has the following advantages over a full-precision variable segmentation of the signal:

- If rate-distortion control is used to distribute coding resources between transient, sinusoidal, and noise models, a restricted segmentation for the corresponding signals simplifies the analysis process significantly.
- The transient modification procedure has a lower computational cost compared to the search for optimal full-precision segmentation with damped sinusoids.
- The restricted segmentation results in a reduction of the number of bits needed to describe the segmentation.

The remainder of this paper is organized as follows. In Section 2, a transient-location modification procedure is presented. In Section 3, a modeling with damped sinusoids is described. Results of computer simulations and informal listening tests are presented in Section 4. Finally, conclusions are presented in Section 5.

2. MODIFICATION OF TRANSIENT LOCATIONS

2.1. Outline

To modify the locations of transients in an audio signal, we proceed with the following steps:

This work was partially supported by Philips Research.

1. Estimate the transient component in the original audio signal and subtract it from the original audio signal to form a residual signal.
2. Modify the locations of the estimated transients in such a way that the transients can occur only at locations specified by the grid.

We also verify that, when the modified transient signal is added to the residual signal obtained in step 1, there is no perceptual difference between the thus obtained signal and the original audio signal.

In order to modify transient locations we have to estimate the transient component in the original audio signal first. Different transient models have been used in parametric coding of audio. In our current work, we use the elegant transient model based on duality between the time and the frequency domain presented in [8].

2.2. Transient estimation and modification

The transient estimation model presented in [8] is based on the duality between the time and the frequency domain. A delta-impulse in the time domain corresponds to a sinusoid in the frequency domain. Furthermore, a sharp transient in the time domain corresponds to a frequency-domain signal which can be represented efficiently by a sum of sinusoids. Specifically, the transients are estimated using the following steps:

1. The discrete cosine transform (DCT) is used to transform a time-domain segment to the frequency domain. The segment size (equivalently, the DCT size) should be sufficiently large to ensure that a transient is a short event in time (thus, transformed to the frequency domain, it can be modeled efficiently by sinusoids). A block size of about 1 s is sufficient.
2. The frequency-domain (DCT-domain) signal is analyzed with a sinusoidal model. In our current work, we use a consistent iterative sinusoidal analysis/synthesis with Hanning-windowed sinusoids as presented in [9].

The sinusoidal analysis of a DCT-domain segment is done on a segment-by-segment basis. As a result, the DCT-domain segment is represented as

$$S_i(l) = \sum_{j=1}^{J} h(l) A_{i,j} \cos\left(\omega_{i,j}(l - \frac{L-1}{2}) - \phi_{i,j}\right), \qquad (1)$$

$$l = 0, \ldots, L-1, \quad i = 1, \ldots, I,$$

where L is the length of the sinusoidal segments (the shift between sinusoidal segments is $L/2$). The length of the sinusoidal segments, L, is a small fraction of the DCT size, N. $h(l)$ are samples of the Hanning window, and $\{A_{i,j}, \omega_{i,j}, \phi_{i,j}\}$ are amplitudes, frequencies and phases of the estimated sinusoids, respectively. The index i denotes a particular sinusoidal segment within the DCT-domain segment, while the index j denotes a particular sinusoid within the sinusoidal segment. The information about the location of a transient in a time-domain segment is contained in the frequency parameters of the corresponding sinusoids. A transient in the beginning of a segment results in low sinusoidal frequencies, while a transient in the end of a segment results in high sinusoidal frequencies. The frequency resolution of the sinusoidal model depends on the required resolution in estimation of transient locations. If the required time resolution is one sample then the required frequency resolution is defined by the reciprocal of the DCT size.

Due to the duality between the transient location in a time-domain segment and the frequencies of the corresponding sinusoids, the obvious way to modify the transient location is to modify the corresponding frequencies (plus a correction in the phase parameters). Let us denote by n_0 the transient location in the time-domain segment and $\hat{n}$ the closest allowed location from a time grid. Then the desired time shift is defined as

$$\Delta n = n_0 - \hat{n}. \qquad (2)$$

In order to modify the transient location by Δn the frequencies $\omega_{i,j}$ and phases $\phi_{i,j}$ corresponding to the transient should be modified as follows:

$$\begin{aligned} \hat{\omega}_{i,j} &= \omega_{i,j} - \frac{\Delta n \pi}{N}, && (3) \\ \hat{\phi}_{i,j} &= \phi_{i,j} + \frac{\Delta n \pi}{N}\left(\frac{L-1}{2} + (i-1)\frac{L}{2}\right). && (4) \end{aligned}$$

No modification of amplitudes $A_{i,j}$ is needed.

Note that the above procedure is different from independent quantization of sinusoidal parameters. All frequencies corresponding to one transient are modified by the same amount. This, together with the phase correction of Eq. (4), ensures that the shape of the time-domain transient is preserved, only the location is modified.

2.3. Details of the modification procedure

Because the DCT size is relatively large (1 s), more than one transient can occur in a time-domain segment. In this case, the model has to identify sinusoidal parameters corresponding to different transients. This is done by declaring close sinusoidal frequencies $\omega_{i,j}$ to represent the same transient. Specifically, two sinusoids having frequencies differing by not more than ε_ω are declared to represent the same transient and two sinusoids having frequencies differing by more than ε_ω are declared to represent different transients. Then locations of all transients are modified separately. Below when we talk about a group of frequencies $\omega_{i,j}$ we mean frequencies corresponding to a particular transient.

A transient can occur at the beginning or at the end of a time-domain segment. In this case, the modification of sinusoidal frequencies can yield frequencies below 0 or above π. This results in a distortion of the shape of the time-domain transient. To account for this, we allow an overlap between time-domain segments (0.1 s). In this case a transient can appear in two overlapping segments, i.e., in the region of mutual overlap. Because the overlap is sufficiently large, if the transient is located very close to a border of one of the overlapping segments, then it is located at a safe distance from a border of the other segment. It is straightforward to identify the transient location from sinusoidal frequencies, and therefore it is easy knowing the estimated sinusoidal frequencies in the two overlapping segments to identify when a transient is represented in two segments. If such a situation occurs, we cancel the corresponding sinusoids in the segment where the transient is closer to the corresponding border.

A typical transient lasts for more than one time sample. A natural question is then what the location n_0 of the transient is. After the modification of location the corresponding sample of the transient will be placed at location $\hat{n}$ corresponding to the beginning of a segment defined by the time grid. Therefore, it is important that the estimated value n_0 corresponds to the start of the transient. The time-domain approach described below proved to yield good results. First, we identify the time samples $n_{\min}$ and $n_{\max}$ corresponding to the frequency values $\min(\omega_{i,j})$ and $\max(\omega_{i,j})$, where $\omega_{i,j}$ are frequencies of sinusoids corresponding to a particular transient. Next, we find the highest amplitude of the estimated transient signal in the time interval $[n_{\min}, n_{\max}]$. Then, the start sample of the transient n_0 is defined to be the first sample in the interval $[n_{\min}, n_{\max}]$ having amplitude higher than 10 % of the highest amplitude.

Typically, the estimated transient component of an audio signal contains samples of small amplitudes before the sample n_0. Because we declare the time sample n_0 to be the first sample of the transient and that no transients can occur at a distance defined by ε_ω before the transient, we force the corresponding samples before n_0 to have zero amplitude. As a result, those samples go to the residual signal with their original amplitudes.

3. MODELING WITH DAMPED SINUSOIDS

A damped sinusoidal model aims at approximating a signal s with a sum of sinusoids with exponentially-modulated amplitudes, i.e.,

$$\begin{aligned} \hat{s}(n) &= \sum_{m=1}^{2M} B_m e^{\alpha_m n} \cos(\upsilon_m n + \psi_m) \\ &= \sum_{m=1}^{M} r_m p_m^n, \quad n = 0, \ldots, K-1, \end{aligned} \qquad (5)$$

where $r_m, p_m \in \mathbf{C}$. $K \in \mathbf{N}$ is the segment length. Equation (5) expresses $\hat{s}(n)$ as the sum of M damped (complex) exponentials. The parameter r_m determines the initial phase and amplitude, while p_m determines the frequency and damping. In order to determine the parameters r_m and p_m for the M exponentials, we used the matching pursuit algorithm [10]. Matching pursuit approximates a signal by a finite expansion into elements chosen from a redundant dictionary. Let $\mathcal{D} = (g_\gamma)_{\gamma \in \Gamma}$ be a complete dictionary of unit-norm elements. The matching pursuit algorithm is a greedy iterative algorithm which projects a signal s onto the dictionary element g_γ that best matches the signal and subtracts this projection to form a residual signal to be approximated in the next iteration. Finding the best matching dictionary element consists of computing the inner products $\langle s, g_\gamma \rangle$ and selecting the element that maximizes the inner product. In order to find the parameters r_m and p_m, we construct a dictionary consisting of damped exponentials,

$$g_{\alpha,\upsilon} = c e^{\alpha n} e^{i\upsilon n}, \quad n = 0, \ldots, K-1, \qquad (6)$$

where the constant c is introduced for having unit-norm dictionary elements, and compute the inner products of the residual signal at iteration m, s_m, and the dictionary elements defined in Eq. (6):

$$\langle s_m, g_{\alpha,\upsilon} \rangle = c \sum_{n=0}^{K-1} s_m(n) e^{\alpha n} e^{-i\upsilon n}. \qquad (7)$$

By doing this for different values of α, we evaluate the transfer function $S_m(z)$ on circles in the complex z-plane having radius e^α.

4. EXPERIMENTAL RESULTS

In this section, we present results of computer simulations and informal listening tests with audio signals. The audio excerpts are a castanets signal, songs by ABBA, Celine Dion, Metallica, and a vocal by Suzanne Vega. The signals are sampled at 44.1 kHz. The DCT size is 44288 samples (ca 1 s) and the overlap between time-domain segments is 4410 samples (0.1 s). The sinusoidal analysis of the DCT-domain signals is done using Hanning windows of length 512 samples and mutual overlap 256 samples. The transient component of the signal was estimated and subtracted to form the residual signal. Next, the transient locations were modified according to a time grid of 220 samples (ca 5 ms).

It is important to verify that the modification of transient locations does not introduce any audible distortion. To check that, we added the modified transient signal to the residual signal. Our listening tests verified that there is no perceptual difference between the thus obtained signal and the original audio signal.

Next, we illustrate the improvement due to the modification procedure. We study the performance of a damped sinusoidal model with the restricted segmentation for an original transient signal (i.e., generally, a transient starts at an arbitrary location) and a modified transient signal (a transient starts in the beginning of a segment). The optimal restricted time segmentation (with the minimum segment size of 220 samples) for damped sinusoids is found using the technique proposed in [11]. The performance is studied in terms of signal-to-noise ratio (SNR) versus number of damped sinusoids and is well illustrated by Figure 1, where we present our results for a particular transient of the castanets signal. The modification procedure results in a considerably smaller number of damped sinusoids needed to represent the transient with a certain quality. Lower plots of Figures 2 and 3 show the reconstruction with 25 damped sinusoids of the original and the modified transients, respectively. The original transient is not located in the beginning of the segment and, as a result, the modeling error is distributed to samples before the transient. This results in an audible pre-echo. On the other hand, the modified transient is located in the beginning of the segment and, as a result, the pre-echo problem is eliminated.

5. CONCLUSIONS

In this work, we presented a method for improving the damped sinusoidal modeling of transients in an audio signal. The method is based on modification of transient locations

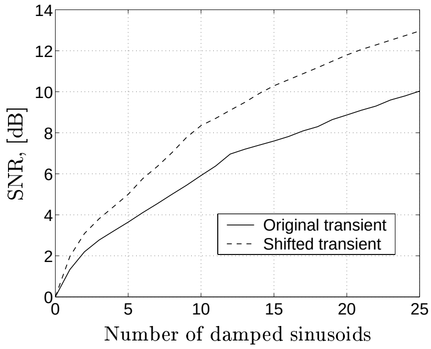

Figure 1: Performance of a damped sinusoidal model in the case of a restricted segmentation for an original and a shifted transient. The minimum segment size is 5 ms.

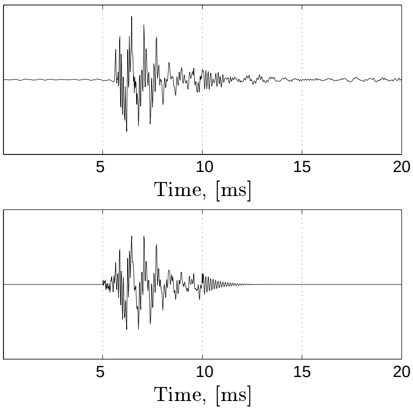

Figure 2: The original transient and its reconstruction with 25 damped sinusoids. The minimum segment size is 5 ms.

such that transients can start only at locations specified by a relatively coarse time grid. A particular advantage of restricting the transient locations is that it simplifies an analysis procedure in an audio coder (involving transient, sinusoidal and noise models), and reduces the side information associated with the corresponding segmentation.

It has to be noted, however, that more investigation is needed for stereo audio signals. In the case of stereo, time delays between channels are important for spatial localization of objects. Therefore, modifying transient locations independently in two channels can result in a stereo image different from the original. The experiments presented in this paper were conducted with mono audio signals only, and more study is to be performed for stereo.

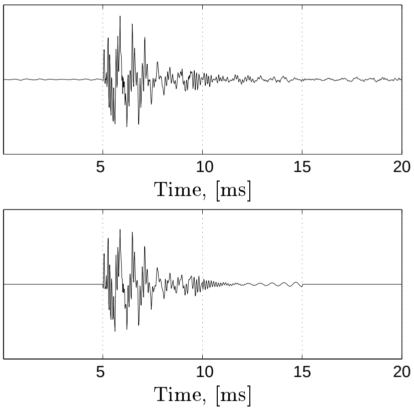

Figure 3: The shifted transient and its reconstruction with 25 damped sinusoids. The minimum segment size is 5 ms.

6. REFERENCES

[1] K. N. Hamdy, M. Ali, and A. H. Tewfik, "Low bit rate high quality audio coding with combined harmonic and wavelet representation," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Proc.*, vol. 2, (Atlanta, Georgia, USA), pp. 1045–1048, 1996.

[2] B. Edler, H. Purnhagen, and C. Ferekidis, "ASAC - analysis/synthesis audio codec for very low bit rates." Preprint 4179 (F-6) 100th AES Convention, Copenhagen, Denmark, 1996.

[3] H. Purnhagen, "Advances in parametric audio coding," in *Proc. 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, (New Paltz, New York, USA), pp. W99–1–W99–4, 1999.

[4] T. S. Verma and T. H. Y. Meng, "A 6 kbps to 85 kbps scalable audio coder," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Proc.*, vol. II, (Istanbul, Turkey), pp. 877–880, 2000.

[5] M. Goodwin, "Residual modeling in music analysis-synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Proc.*, vol. 2, (Atlanta, Georgia, USA), pp. 1005–1008, 1996.

[6] J. Nieuwenhuijse, R. Heusdens, and E. F. Deprettere, "Robust exponential modeling of audio signals," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Proc.*, vol. 6, (Seattle, Washington, USA), pp. 3581–3584, 1998.

[7] K. Brandenburg and G. Stoll, "ISO-MPEG-1 Audio: a generic standard for coding of high-quality digital audio," *J. Audio Eng. Soc.*, vol. 42, pp. 780–792, October 1994.

[8] T. S. Verma, S. N. Levine, and T. H. Y. Meng, "Transient modeling synthesis: a flexible analysis/synthesis tool for transient signals," in *Proc. Int. Computer Music Conf.*, (Thessaloniki, Greece), pp. 25–30, 1997.

[9] K. Vos, R. Vafin, R. Heusdens, and W. B. Kleijn, "High-quality consistent analysis-synthesis in sinusoidal coding," in *Proc. Audio Eng. Soc. 17th Conference "High Quality Audio Coding"*, (Florence, Italy), pp. 244–250, 1999.

[10] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Proc.*, vol. 41, pp. 3397–3415, December 1993.

[11] Z. Xiong, K. Ramchandran, C. Herley, and M. T. Orchard, "Flexible tree-structured signal expansions using time-varying wavelet packets," *IEEE Trans. Signal Proc.*, vol. 45, pp. 333–345, February 1997.