IMPROVED MODELING OF AUDIO SIGNALS BY MODIFYING TRANSIENT LOCATIONS

Renat Vafin[1], *Richard Heusdens*[2], *Steven van de Par*[3], *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

[3]Digital Signal Processing Group
Philips Research Laboratories
5656 AA Eindhoven, The Netherlands
Steven.van.de.Par@philips.com

ABSTRACT

We propose a method for obtaining an improved representation of transients in audio signals. The representation is based on a damped sinusoidal model. To improve the representation, transient locations are modified in such a way that a transient can start only at the beginning of a sinusoidal segment. The introduced modifications facilitate a reduction of the number of damped sinusoids needed to model a transient well and eliminate pre-echo artifacts. We verify with a listening test that the modifications do not result in a perceptual difference between the original and 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, 5]. In a parametric audio coder, a signal is represented by a 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 (e.g., [1, 4, 5]). Having a dedicated model for the transient component proved to be beneficial for audio signals with sharp attacks, because sinusoidal and noise models cannot represent those perceptually important events efficiently [6].

We propose a method for improving the representation of transients. It was shown in [7] that transients can be modeled efficiently using sinusoids with exponentially-modulated amplitudes (damped sinusoids). 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 does not start at the beginning of a segment. Compared to the case where a transient starts at the beginning of a segment, the number of damped sinusoids needed to model the transient with a certain quality increases considerably. If a transient is not modeled properly, the modeling error is distributed over the entire segment, resulting in audible pre-echoes. Different methods have been used to solve this problem:

- 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 [8]). This method can reduce the pre-echo problem, but does not solve it completely.
- Allow a one-sample-precision (full-precision) variable segmentation of the signal, such that transients will always start at the beginnings of segments (e.g., [1]). However, in a hybrid audio coder, where a search for optimal segmentation is performed through a few analysis models, a full-precision variable segmentation can result in a large computational complexity.

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

- The restricted segmentation significantly simplifies the analysis procedure in an audio coder.
- 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. The procedure to modify transient locations is described in Section 2. Modeling with damped sinusoids is described in Section 3. Results of computer simulations and listening tests are presented in Section 4. Finally, conclusions are summarized in Section 5.

2. MODIFICATION OF TRANSIENT LOCATIONS

In [9] we presented a method for modifying transient locations in an audio signal. The transient component of the audio signal is estimated using a model based on duality between the time and frequency domains, as presented in [10]. This transient model is good for very short transients, i.e., with a sharp attack and a fast decay. Transient locations are modified by modifying parameters of a frequency-domain representation of the transient component.

This paper presents an improved method for modifying transient locations. In this new method, an audio signal is modified in the following steps:

1. The beginnings and ends of transients are detected using an energy-based approach with two sliding rectangular windows, as presented in [11].
2. The samples between the beginning and the end of each transient are shifted (essentially cut-and-paste) to the locations specified by a sinusoidal segmentation.
3. The signal parts in-between transients are time-warped to fill the intervals between the shifted transients.

The advantages of the new transient modification method over the one presented in [9] are the following:

This work was partially supported by Philips Research.

- The transient detection model of [11] provides good results also for transients with slow decay.
- The time-warping of the signal parts in-between transients is based on knowledge of properties of sound perception, such as pitch perception and temporal masking effects.
- The new modification method results in a lower computational complexity.

The transient detection approach of [11] used in step 1 is based on the evaluation of the criterion function, $C(n)$:

$$C(n) = \log\left(\frac{E_R(n)}{E_L(n)}\right) \cdot E_R(n), \qquad (1)$$

$$E_L(n) = \sum_{k=n-K}^{n-1} s^2(k), \qquad E_R(n) = \sum_{k=n+1}^{n+K} s^2(k),$$

where $E_L(n)$ and $E_R(n)$ are the energies of the input signal s within length-K rectangular windows on the left- and right-hand side of a time sample n. Significant peaks of the criterion function $C(n)$ correspond to the beginnings of transients.

Step 2 of the new transient modification method is straightforward. We now describe step 3 of the modification method. Due to modification of transient locations, the distance between two transients can become longer or shorter. To fill the interval between the shifted transients, the signal part in-between has to be time-warped correspondingly. The time-warping of the signal is done in such a way that it preserves the correct amplitudes of the edge points of the signal part in-between the transients. Thus, no discontinuities are introduced just before or after a transient. The signal in-between transients is stretched or compressed in time. To compute the amplitudes at the new integer sampling instances based on the known amplitudes of the original samples, an approximation of the ideal bandlimited interpolation based on $sinc$ functions is used. To compute the amplitude of each new sample, amplitudes of eight original samples are used, four at each side of the new sample. A hanning window is used to constrain the time support of the $sinc$ functions.

For tonal signals, a stretching or compressing of the signal in time results in a corresponding change of fundamental frequency, f_0. The goal of the modification procedure is to ensure that the induced modification of f_0 is not audible. Therefore, the following algorithm is proposed for time-warping a signal part in-between two shifted transients (the steps are illustrated in Figure 1 for the case where the length of the signal between two shifted transients is longer than the original; the opposite case is treated similarly):

a) If the required change in length of a signal part in-between two transients results in a change of f_0 by less than 0.2 %, then simply use the time-warping method as described above (Figure 1a). Else go to step b.

   Motivation: from the literature on psychoacoustics, it is known that changing f_0 of a tonal sound by 0.2 % can be audible [12]. Our experiments verified this result.

b) Split the signal part in-between two transients into two nonoverlapping intervals: the first interval is located directly after the end of the first transient and lasts 10 ms (interval 1 in Figure 1b), and the second interval is the remaining part, i.e., it lasts until the beginning of the second transient (interval 2 in Figure 1b). The lengths of the two intervals are modified by a different amount. If the required change in length of the signal part in-between two transients can be done by changing f_0 in the interval 1 by less than 2 % and in the interval 2 by less than 0.2 %, then time-warp the signal in the two intervals correspondingly. Else go to step c.

   Motivation: the interval directly after the end of a transient is characterized by a strong masking effect from the transient. Therefore, larger changes of the signal in this interval are possible before they become audible. Our experiments verified that a change of f_0 by less than 2 % in the 10 ms interval directly after the end of a transient is inaudible.

c) Time-warp the signal in the two intervals such that the resulting change of f_0 is no more than 2 % in the interval 1 and no more than 0.2 % in the interval 2. If the resulting change in length is not sufficient to fill the distance between

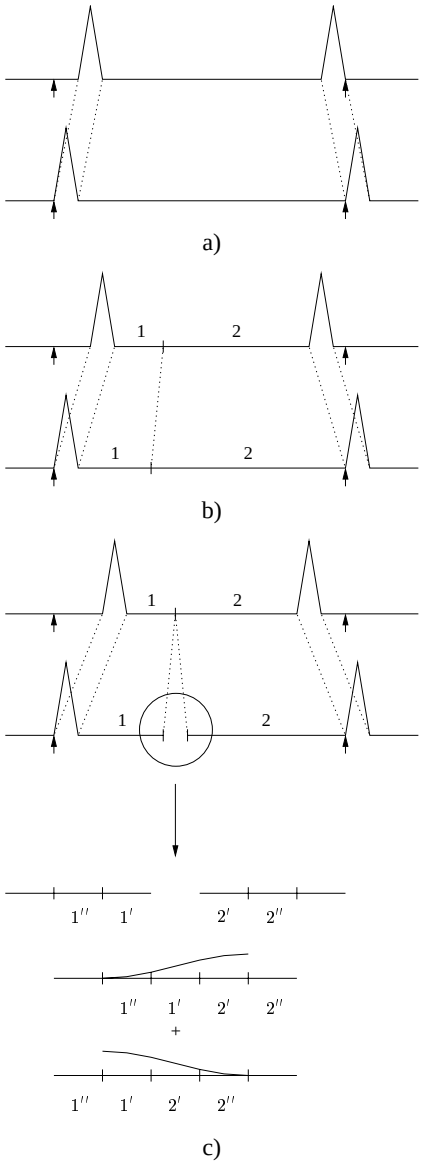

Figure 1: Modification of transient locations. The new locations of transient beginnings are depicted with small arrows. The signal part in-between two transients becomes longer. Steps a, b, c are explained in Section 2.

the shifted transients, then apply an overlap-add procedure with a hanning window using samples from the two intervals in order to increase or decrease the length of the signal. To ensure a smooth transition between two intervals, the length of the overlap-add region is chosen to be larger than required to obtain a correct length of the signal in-between two transients (Figure 1c).

3. MODELING WITH DAMPED SINUSOIDS

It was shown in [7] that a transient can be modeled efficiently using a damped sinusoidal model. This model aims at approximating a signal s by a sum of, say M, sinusoids with exponentially-modulated amplitudes, i.e.,

$$\hat{s}(n) = \sum_{m=1}^{M} a_m e^{d_m n} \cos(\omega_m n + \varphi_m), \quad n = 0, \ldots, N-1, \quad (2)$$

where $a_m, d_m, \omega_m, \varphi_m \in \mathbb{R}$ denote the amplitude, damping coefficient[1], angular frequency and phase of the m-th sinusoidal component, respectively. $N \in \mathbb{N}$ is the segment length.

The sinusoidal parameters a_m, d_m, ω_m and φ_m can be selected with a number of methods, including spectral peak-picking, subspace-based analysis techniques and analysis-by-synthesis methods. For the experiments described in this paper we used the matching-pursuit algorithm [13], a particular analysis-by-synthesis method. The matching-pursuit algorithm is a greedy iterative algorithm which projects at each iteration a signal onto the function (in our case a damped sinusoid) that best matches the signal and subtracts this projection to form a residual signal to be approximated in the next iteration.

To find an optimal time segmentation, we used the algorithm proposed in [14]. By "optimal" we mean optimal in a rate-distortion sense. This algorithm divides the input signal s into non-overlapping segments and tries, by combining these segments, to find the partitioning of s that minimizes the distortion given a target bit budget or a given number of sinusoidal components. Under the assumption of additivity of rate and distortion over the constituent segments, the global optimal segmentation is found by first optimizing the rate versus distortion for each segment independently, and then, using dynamic programming, searching for the optimal segmentation by combining these optimally encoded segments. By doing so, the algorithm gives the optimal time segmentation of s, as well as the number of sinusoidal components to allocate to the individual segments.

4. EXPERIMENTAL RESULTS

Below, we present results of computer simulations and listening tests with audio signals. The signals are mono, sampled at 44.1 kHz. The test excerpts include castanets, bass, ABBA, Celine Dion, Metallica, harpsichord, Suzanne Vega. Transient locations are modified according to a time grid of 220 samples (ca 5 ms).

It is important to verify that the introduced modifications do not result in an audible difference between the original and modified audio signals. To do that we performed a subjective listening test in which signal triplets AOB were presented to listeners. Here

[1]The damping coefficient d_m can be any real number. Positive values of d_m, therefore, correspond to expanding amplitudes rather than to truly damped amplitudes.

Excerpt	Duration, s	# detected transients	Correct responses, %
castanets	7.1	43	57.5
bass	10.8	16	52.5
ABBA	9.9	29	45.0
Celine Dion	12.8	26	52.5
Metallica	10.1	19	52.5
harpsichord	11.7	9	40.0
Suzanne Vega	10.1	13	42.5

Table 1: Results of the listening test on audibility of signal modifications which include shifting transients and time-warping the signal parts in-between transients.

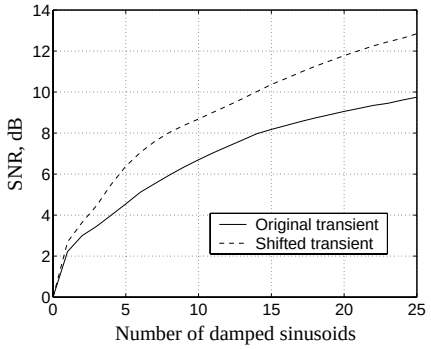

Figure 2: Performance of a damped sinusoidal model in the case of a restricted segmentation for an original and a shifted transient. The minimum segment length is 5 ms.

O is the original signal, A or B is the original signal and B or A is the modified signal. The task of a listener was to respond whether the modified signal was A or B. For each test excerpt, the triplets AOB were presented to a listener 5 times, each time the position of the modified signal (A or B) was changed randomly. Eight experienced listeners participated in the test. The results averaged over all listeners are presented in Table 1. They confirm that, in general, subjects are not able to detect a difference between the original and the modified signal. Two of our most sensitive listeners, however, remarked that they could hear minor differences for one or two excerpts which was supported by a high percentage of correct responses.

Next, we illustrate the improvement due to the modification procedure. We study the performance of a damped sinusoidal model for an original signal (i.e., transients start at arbitrary locations) and for a modified signal (i.e., transients can start only at the beginnings of sinusoidal segments). The methods used to evaluate the performance are the same as in [9]. The performance is studied in terms of signal-to-noise ratio (SNR) versus the number of damped sinusoids and is well represented in Figure 2, where it is shown for a particular transient of the castanets signal. It is evident that more sinusoids are needed to model the transient with a certain quality in the case where the transient does not start at the beginning of a sinusoidal segment. The lower plots of Figures 3 and 4 show the reconstruction with 25 damped sinusoids of the original and the modified transients, respectively. Also, the optimal segmentation defined by the minimum segment length of 5 ms (solid vertical lines) and sinusoidal allocation are shown. The original transient does not start at the beginning of the segment and, as a result, the modeling error is distributed to samples before the transient. This results in a clearly audible pre-echo. On the other hand,

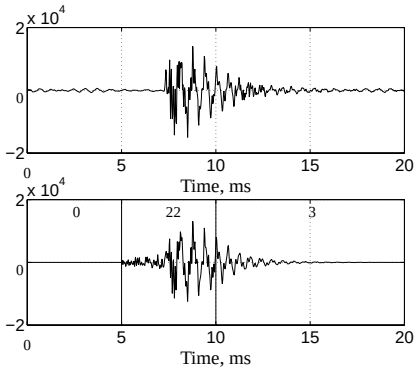

Figure 3: The original transient and its reconstruction with 25 damped sinusoids. The minimum segment length is 5 ms.

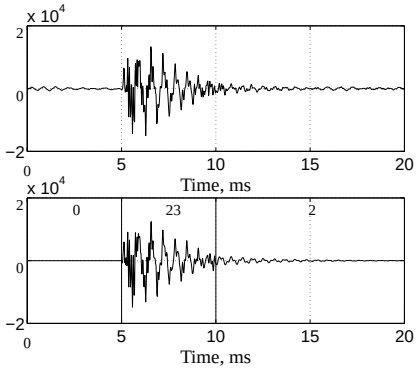

Figure 4: The shifted transient and its reconstruction with 25 damped sinusoids. The minimum segment length is 5 ms.

the modified transient starts at the beginning of the segment and, as a result, the pre-echo problem is eliminated. It has to be noted that a similar improvement would be achieved with a full-precision variable segmentation (and no signal modification). However, in a hybrid coder, the restricted segmentation and the signal modification result in a significantly lower total computational complexity.

5. CONCLUSIONS

In this paper, we elaborated on the idea of modifying transient locations for improved modeling and coding of audio. Transient locations in an audio signal are modified in such a way that a transient can start only at the beginning of an analysis segment. We presented a new method for modifying transient locations. The method is based on shifting transients and time-warping the signal parts in-between transients. The introduced modifications facilitate an efficient representation of transients with damped sinusoids and eliminate pre-echo artifacts. We verified with a listening test that the modifications are, in general, inaudible.

It has to be noted, however, that a straightforward application of the modification procedure is not suitable for stereo signals. The reason for this is that an independent modification of transient locations in the two channels may destroy the original stereo image. A possible solution to the problem is to send side information describing the transient shifts to the decoder, and restore the original transient locations by shifting and time-unwarping the corresponding intervals of synthesized signals. The number of bits used to describe the shift of a transient is, in general, significantly lower than the bit saving obtained due to the reduction of the number of damped sinusoids.

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] A. W. J. Oomen and A. C. den Brinker, "Sinusoids plus noise modeling for audio signals," in *Proc. Audio Eng. Soc. 17th Conf. "High Quality Audio Coding"*, (Florence, Italy), pp. 226–232, 1999.

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] R. Vafin, R. Heusdens, and W. B. Kleijn, "Modifying transients for efficient coding of audio," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Proc.*, vol. 5, (Salt Lake City, Utah, USA), pp. 3285–3288, 2001.

[10] 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.

[11] J. Kliewer and A. Mertins, "Audio subband coding with improved representation of transient signal segments," in *Proc. EUSIPCO*, (Rhodos, Greece), pp. 2345–2348, 1998.

[12] B. C. J. Moore, *An Introduction to the Psychology of Hearing*. Academic Press, 1997.

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

[14] 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.