COMBINED SPECTRAL ENVELOPE NORMALIZATION AND SUBTRACTION OF SINUSOIDAL COMPONENTS IN THE ODFT AND MDCT FREQUENCY DOMAINS

Aníbal J. S. Ferreira

School of Engineering of the University of Porto / INESC Porto
ajf@inescporto.pt

ABSTRACT

Recent research in high-quality audio coding seeks not only improved coding gains but also new functionalities such as easy semantic access to compressed audio material and audio modification in the compressed domain. These objectives imply the decomposition of the audio signal into several components of specific semantic value, such as sinusoidal components, that take advantage of selective coding and parametrization tools.

In this paper we presume an MDCT based audio coding environment and present a new technique combining spectral envelope normalization with accurate subtraction of sinusoidal components in the MDCT frequency domain. It is shown how a parametrization of L stationary sinusoids in the complex ODFT spectrum can lead to the effective subtraction in the real MDCT spectrum, of 3L spectral lines. A demonstration of the implementation of the technique is available on the Internet.

1. INTRODUCTION

Current state-of-the-art high-quality audio coders are in general frequency domain coding schemes that are based on the perceptual audio coding paradigm [1]. This is the case of many proprietary and standardized audio coding algorithms such as AC-3 and MPEG-2 AAC [2].

Traditionally, perceptual audio coders are signal adaptive with respect to the stationarity of the signal and address such functionalities as high compression ratio, low delay coding, error resilience and bit-stream scalability.

However, an interest is emerging for non-conventional functionalities such as easy semantic segmentation, classification and access to audio material using information naturally embedded in the compressed audio representation, and such as easy audio modification in the compressed domain (*e.g.,* pitch modification or time-scale modification). These functionalities are particularly interesting in the perspective of the forthcoming MPEG-7 standard whose objective is to standardize a description of audio/visual information allowing its easy classification, access and retrieval.

This new trend in high quality audio coding has recently started to be addressed by a new generation of audio coders that look into the audio signal in a semantic sense, trying to isolate individual signal components and assigning to each one the most efficient and appropriate coding tools since the associated psychoacoustic rules may differ significantly. For example, ASC, a former MPEG-4 candidate, is a perceptual audio coder that combines the parametrization of an existing relevant harmonic profile in the audio signal (within the analysis/synthesis framework of the coder), with a perceptually based quantization technique, in order to reach good coding quality for both resolved and unresolved partials [2]. Other coders implement an explicit decomposition of the audio signal typically into three individual components: sinusoids, stationary noise, and non-stationary noise (transients) [3, 4]. Each component is estimated, is parametrized, and is removed (*i.e.,* subtracted) from the original signal, creating a residual that undergoes further analysis and parametrization regarding the remaining signal components.

Concerning the signal components in focus in this paper (the sinusoidal part), all the references we had access to, address sinusoidal estimation and synthesis using an analysis/synthesis framework (*e.g.,* the McCAulay and Quatieri sinusoidal addition method [5]) different from that of the remaining two signal components. Furthermore, both the subtraction of the sinusoidal components from the original signal in the encoder, as well as their addition in the decoder to the remaining audio components, is implemented in the time domain.

Presuming quasi-stationary conditions and taking in consideration that the MDCT is the most commonly used filter bank in audio coding, we propose a new technique combining spectral envelope normalization and accurate subtraction in the (complex) ODFT or (real) MDCT frequency domains, of an existing relevant sinusoidal structure in the audio signal, thus avoiding the need for multiple time/frequency transformations of residuals in the encoder, as well as the associated overhead in computation and increased system delay.

Our emphasis in this paper on the sinusoidal components of an audio signal is due to the fact that according to a study on typical audio material and using analysis audio frames having a duration of 23 ms., it has been concluded that more than 80% of all audio frames can be classified as quasi-stationary and exhibit at least three relevant sinusoidal components harmonically related [2].

The structure of this paper is as follows. In section 2 we present the main expressions used for sinusoidal modeling in the ODFT or MDCT frequency domains using only a reliable estimate of the frequency, magnitude and phase of a single ODFT spectral line. In section 3 we address spectral envelope normalization after LPC analysis or cepstral analysis and in the perspective of its combination with spectral subtraction using the proposed technique. In section 4 we describe the complete algorithm of spectral flattening using the MDCT filter bank and insuring perfect reconstruction. We also illustrate in this section the effectiveness of the spectral subtraction technique in the absence of spectral envelope normalization, and discuss the implications of combining spectral envelope normalization with spectral subtraction. Finally, in section 5 we present the main conclusions of this paper.

This work has been supported by the Portuguese Foundation for Science and Technology (FCT) under project FCT34018/99.

2. SINUSOIDAL MODELING IN THE FREQUENCY DOMAIN

The real coefficients of the MDCT analysis filter bank are defined as (we assume that N is an even number):

$$X_M(k) = \sum_{n=0}^{N-1} h(n)x(n)\cos\left[\frac{2\pi}{N}\left(k+\frac{1}{2}\right)(n+n_0)\right], \quad (1)$$

where $n_0 = \frac{1}{2} + \frac{N}{4}$, $x(n)$ is the input signal, $h(n)$ is the time analysis window whose length is N, and k is the coefficient index, with $0 \leq k \leq \frac{N}{2} - 1$. Throughout this paper we will take for the time analysis window the following function which is commonly referred to as the "sine window":

$$h(n) = \sin\frac{\pi}{N}(n+\frac{1}{2}) \quad , \quad 0 \leq n \leq N-1. \quad (2)$$

It can be shown that the MDCT coefficients can also be conveniently obtained as [2]:

$$X_M(k) = \Re e\left\{X_O(k)\right\}\cos\theta(k) + \Im m\left\{X_O(k)\right\}\sin\theta(k) \quad (3)$$

where $\theta(k) = \frac{\pi}{N}\left(k+\frac{1}{2}\right)\left(1+\frac{N}{2}\right)$ and $X_O(k)$ represents the coefficients of the (complex) Odd-DFT transform (ODFT) which is defined as:

$$X_O(k) = \sum_{n=0}^{N-1} h(n)x(n)e^{-j\frac{2\pi}{N}(k+\frac{1}{2})n}. \quad (4)$$

It can be concluded from (3) that the magnitude of the MDCT coefficients are upper-bounded by the magnitude of the ODFT coefficients:

$$\mid X_M(k) \mid = \mid X_O(k) \mid \mid \cos\left[\angle X_O(k) - \theta(k)\right] \mid, \quad (5)$$

where $\theta(k)$ is defined as above and $\angle X_O(k)$ represents the phase of the k^{th} ODFT coefficient. The significance of this result is that the estimation of sinusoidal components is much more reliable in the ODFT domain than in the MDCT domain.

Throughout this paper we will use for illustration purposes a short segment of a real audio signal ("Tom's Dinner" - Suzanne Vega). This audio signal is available, together with a demonstration Matlab M-file, at the Web page indicated in section 5. The magnitude of the ODFT coefficients of a 1024-samples segment of this audio signal after being windowed by (2), as well as the magnitude of the resulting MDCT coefficients, are plotted in Fig. 1. Besides confirming the conclusion resulting from equation (5), this figure also indicates the position of 25 sinusoidal components harmonically related that have been detected using an algorithm presented in [2]. It is shown in a companion paper [6] and also demonstrated by means of a related Matlab M-file available on the Internet (http://www.inescn.pt/∼ajf/waspaa01/accurate.html) that the parameters regarding magnitude (A), frequency $\left(\frac{2\pi}{N}(\ell+\Delta\ell)\right)$, and phase ($\phi$) of a stationary sinusoid defined as:

$$x(n) = A\sin\left[\frac{2\pi}{N}(\ell+\Delta\ell)n + \phi\right] , \quad (6)$$

where ℓ and $\Delta\ell$ are respectively the integer part and the fractional part of the frequency of a sinusoid on the "bin" ODFT frequency scale, can be estimated, with 99% accuracy, using the complex ODFT filter bank and the sine window.

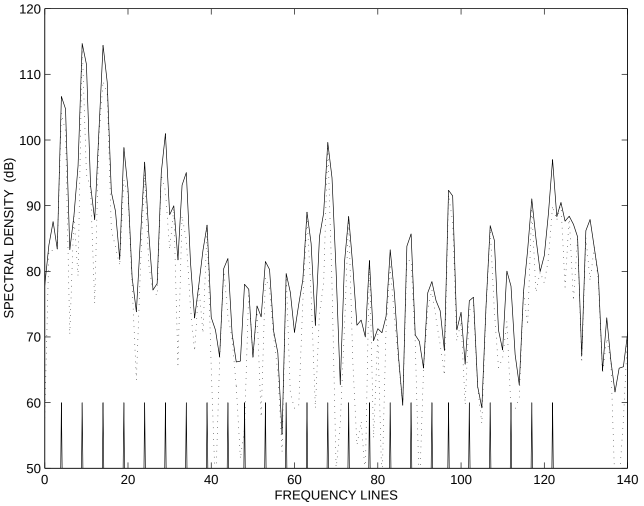

Figure 1: Magnitude of the ODFT spectrum (solid line) and of the MDCT spectrum (dotted line) of a real audio signal. The spikes at the bottom indicate the integer position of sinusoidal components harmonically related.

Once the parameters A, ℓ, $\Delta\ell$, and ϕ are known and assuming they correspond to a stationary sinusoid, is is possible to reconstruct both the magnitude and phase of three ODFT spectral lines (or subbands) whose indexes are $\ell-1$, ℓ and $\ell+1$ and that represent the main lobe of the frequency response of $h(n)$ centered on the frequency of the sinusoidal signal [6, 2]:

$$|X_O(\ell-1)| \simeq \frac{|NA|}{4}\left|\frac{2}{\sqrt{3}}\cos\frac{\pi}{6}(2\Delta\ell+1)\right|^F, \quad (7)$$

$$\angle X_O(\ell-1) = \phi - \frac{\pi}{2N} + \pi\Delta\ell\left(1-\frac{1}{N}\right), \quad (8)$$

$$|X_O(\ell)| \simeq \frac{|NA|}{4}\left|\frac{2}{\sqrt{3}}\cos\frac{\pi}{6}(2\Delta\ell-1)\right|^F, \quad (9)$$

$$\angle X_O(\ell) = \phi - \pi\left(1-\frac{1}{2N}\right) + \pi\Delta\ell\left(1-\frac{1}{N}\right), \quad (10)$$

$$|X_O(\ell+1)| \simeq \frac{|NA|}{4}\left|\frac{2}{\sqrt{3}}\cos\frac{\pi}{6}(2\Delta\ell-3)\right|^F, \quad (11)$$

$$\angle X_O(\ell+1) = \angle X_O(\ell-1), \quad (12)$$

where F is a constant. These expressions were obtained taking advantage of the properties of the ODFT filter bank and using a simple model for the main lobe of the frequency response of the analysis window (whose width is $6\pi/N$) [6]. This is the reason why expressions (7), (9), and (11) are approximate, while expressions (8), (10), and (12) are exact. It should also be noted that

- the magnitude of $|X_O(\ell)|$ corresponds always to a local maximum,
- these results are still approximately valid even when a sinusoid is moderately modulated in amplitude [6].

As a consequence and for each sinusoidal component, the above expressions can be used to synthesize three spectral lines that can be directly subtracted from the complex ODFT spectrum, or from the real MDCT spectrum that derives from the former using (3).

3. SPECTRAL ENVELOPE NORMALIZATION

Effective flattening of the spectral representation of the audio signal is a desired feature in a perceptual audio coder. For example, Twin-VQ, a successful coding proposal to the MPEG-4 standardization activities, relies on four levels of normalization of the MDCT spectrum prior to weighted interleave vector quantization, in order to maximize the quantization/coding gain of this latter technique [7]. On the other hand, it is also known that audio coders using Huffman coding loose efficiency when the quantized coefficients to be entropy encoded exhibit sudden large magnitudes.

For this reason, we combine the spectral subtraction technique proposed in this paper with spectral envelope normalization and we compare two techniques: LPC (All-Pole) analysis and cepstral analysis. LPC-AP analysis is well-known from the speech processing area [8] and is also used in audio coding [9]. However, all-pole modeling is also frequently regarded as unsuitable to model the spectral envelope of audio signals essentially for two reasons:

- LPC-AP modeling does not model accurately the zeros of the spectral envelope of typical audio material,
- due to high dynamic range of typical audio material, LPC-AP modeling does not provide the same smoothness across the whole audio range (*e.g.,* high-level formants are modeled by a strong peak in the LPC spectrum).

These conclusions can be easily verified taking the example of the ODFT spectrum depicted in Fig. 1. We have modeled its spectral envelope using a 16^{th} order LPC-AP filter (which is considered in the literature to be adequate for audio signals [9]) and also using the first 16 cepstral coefficients (which have shown to provide a smooth and accurate spectral modeling in all our tests with typical audio material) according to the algorithm of Fig. 2. The spectral envelope approximation resulting from either approach is depicted in Fig. 3. As we will discuss in the next section, due to the "peaky" tendency of all-pole based spectral modeling, a "smoother" alternative such as cepstral based modeling is more convenient when combined with spectral subtraction.

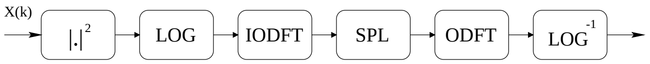

Figure 2: Spectral envelope modeling by *short-pass liftering* the cepstral coefficients.

4. MDCT SPECTRAL ENVELOPE NORMALIZATION AND SINUSOIDAL SUBTRACTION ALGORITHM

The complete algorithm presuming the MDCT filter bank and featuring spectral normalization by a suitable spectral envelope model, as well as accurate sinusoidal subtraction, while insuring perfect reconstruction in the absence of quantization, is represented in Fig. 4. We note that in this context, perfect reconstruction can only be achieved when the algorithm of Fig. 4 is included in the overlap-add procedure underlying the MDCT filter bank, so as to insure cancellation of time-domain aliased terms [10]. This is due to the fact that the "ODFT2MDCT" operator, which represents equation (3), is linear but not invertible and as a consequence, even in the absence of quantization, spectral normalization and subtraction, it can be shown that $\hat{x}(n) \neq x(n)$. This is also the reason why in the decoder the sinusoidal addition must take place in the MDCT domain (instead of the ODFT domain). In contrast, the inverse spectral normalization can be indistinctly implemented in the MDCT or in the ODFT frequency domains.

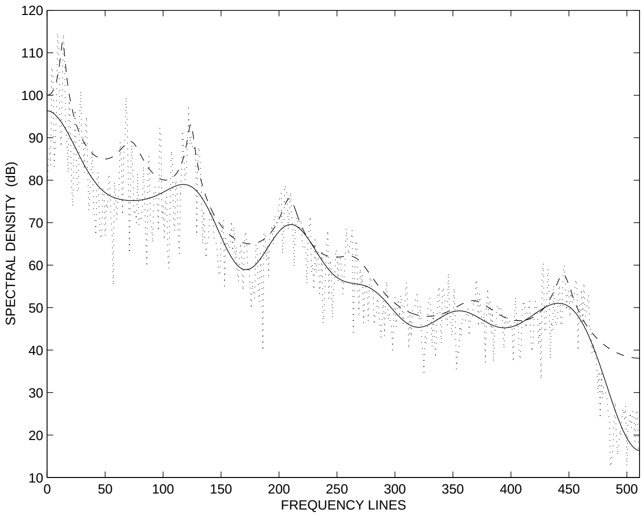

Figure 3: Spectral envelope modeling of the ODFT spectrum (dotted line) using a 16^{th} order LPC-AP predictor (dashed line) and a 16 coefficient cepstral based model (solid line).

A clear advantage of the spectral subtraction technique is that using only accurate estimates of the frequency, magnitude and phase of L sinusoidal components of the ODFT spectrum, it is possible to accurately subtract 3L spectral lines from the complex ODFT or real MDCT spectra, if the audio signal is quasi-stationary. As sinusoidal components consist generally in strong spectral peaks, as a consequence of this technique, the spectrum will be effectively flattened. This is illustrated in Fig. 5 for the case of the ODFT and MDCT spectra shown if Fig. 1, and in the absence of spectral envelope normalization. It can be seen that even for a real audio signal as the one of this example (and also used in the demonstration Matlab file), the three spectral lines defining the peak of each sinusoidal component, are effectively reduced by as much as 20 dB or 30 dB, or even more when $k = \ell$.

Taking as a reference the residual MDCT coefficients after implementing the spectral subtraction operation just described, we evaluated the impact of normalizing the spectral magnitudes prior to spectral subtraction using the two spectral envelope approximations illustrated in Fig. 3. For each case, the denormalized spectral magnitudes of the residual relative to the reference residual is represented in Fig. 6. It can be seen that the highest deviation occurs when the LPC spectral envelope model is used, specifically in a region where the LPC envelope model peaks abruptly. This result is expected since effective spectral subtraction is only achieved when the spectral envelope model is very smooth in the frequency region involving the three spectral lines that are synthesized using the information of only one spectral line. As an average measure for the results of Fig. 6, the total deviation of the relative spectral density for the LPC envelope model is 40% higher than for the cepstral envelope model. In this sense, cepstral based modeling is more appropriate than LPC based modeling.

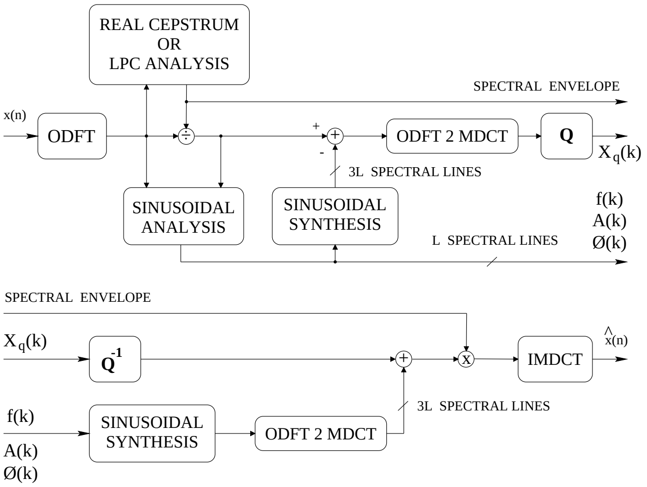

Figure 4: Encoder and decoder sections of the algorithm allowing spectral envelope normalization and subtraction of sinusoidal components.

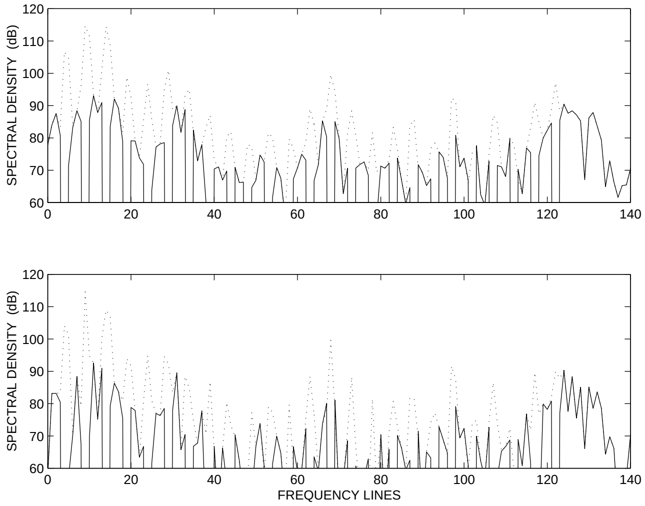

Figure 5: Magnitude of the ODFT residual (top) and of the MDCT residual (bottom) after spectral subtraction of 25 sinusoidal components. The original spectra are represented by dotted lines.

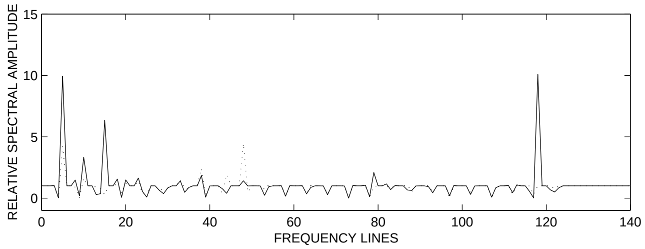

Figure 6: Magnification of the MDCT residual due to LPC (solid) or cepstral based (dotted) spectral envelope normalization.

5. CONCLUSION

In this paper we have presented a technique of spectral flattening in the ODFT or MDCT frequency domains that combines spectral envelope normalization with accurate spectral subtraction of sinusoidal components. Expressions have been presented for the synthesis of three spectral lines defining a sinusoid using only the frequency, the magnitude and phase of a single ODFT spectral line. We believe the technique presented has the potential to improve the coding efficiency of perceptual coders as well as to embed parametric information in the compressed/coded representation, in a natural way. Preliminary tests with an MDCT based codec (ASC) indicate that by synthesizing uniquely the tonal/harmonic parametric information, a useful signal is obtained that can be used for preview purposes. A Web page has been prepared to detail and illustrate the implementation of the proposed technique: http://www.inescn.pt/∼ajf/waspaa01/flattening.html.

6. REFERENCES

[1] Nikil Jayant, James Johnston, and Robert Safranek, “Signal Compression Based on Models of Human Perception,” *Proceedings of the IEEE*, vol. 81, no. 10, pp. 1385–1422, October 1993.

[2] Aníbal J. S. Ferreira, *Spectral Coding and Post-Processing of High Quality Audio*, Ph.D. thesis, Faculdade de Engenharia da Universidade do Porto-Portugal, 1998, http://telecom.inescn.pt/doc/phd_en.html.

[3] Scott N. Levine, *Audio Representations for Data Compression and Compressed Domain Processing*, Ph.D. thesis, Stanford University, 1998.

[4] Tony S. Verma, *A Perceptually Based Audio Signal Model with Application to Scalable Audio Compression*, Ph.D. thesis, Stanford University, 1999.

[5] R. J. McCaulay and T. F. Quatieri, “Speech Analysis/Synthesis based on a Sinusoidal Speech Model,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 34, no. 4, pp. 744–754, August 1986.

[6] Aníbal J. S. Ferreira, “Accurate Estimation in the ODFT Domain of the Frequency Phase and Magnitude of Stationary Sinusoids,” in *2001 Workshop on Applications of Signal Processing to Audio and Acoustics*, October 21-24 2001, submitted.

[7] T. Moriya, N. Iwakami, K. Ikeda, and S. Miki, “A Design of Transform Coder for Both Speech and Audio Signals at 1 bit/sample,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1997, pp. 1371–1374.

[8] Andreas S. Spanias, “Speech Coding: A Tutorial Review,” *Proceedings of the IEEE*, vol. 82, no. 10, pp. 1541–1582, October 1994.

[9] S. Boland and M. Deriche, “Hybrid LPC and Discrete Wavelet Transform Audio Coding with a Novel Bit Allocation Algortihm,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1998, pp. 3657–3660.

[10] John P. Princen and Alan Bernard, “Analysis/Synthesis Filter Bank Design Based on Time Domain Aliasing Cancellation,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-34, no. 5, pp. 1153–1161, October 1986.