Restructured Audio Encoder for Improved Computational Efficiency

5103 (F- 4)

Ye Wang (1), Miikka Vilermo (1), Mauri Väänänen (1) and Leonid Yaroslavsky (2)
(1) Nokia Research Center, Speech and Audio Systems Lab.,Tampere, Finland.
(2)Department of Interdisciplinary Studies, Tel Aviv University, Ramat Aviv 69978, ISRAEL.

Presented at
the 108th Convention
2000 February 19-22
Paris, France

This preprint has been reproduced from the author's advance manuscript, without editing, corrections or consideration by the Review Board. The AES takes no responsibility for the contents.

Additional preprints may be obtained by sending request and remittance to the Audio Engineering Society, 60 East 42nd St., New York, New York 10165-2520, USA.

AN AUDIO ENGINEERING SOCIETY PREPRINT

RESTRUCTURED AUDIO ENCODER FOR IMPROVED COMPUTATIONAL EFFICIENCY

Ye Wang[1], Leonid Yaroslavsky[2], Miikka Vilermo[1], Mauri Väänänen[1]

[1] Nokia Research Center
Speech and Audio Systems Lab.
Tampere, Finland
ye.wang@nokia.com
miikka.vilermo@nokia.com
mauri.vaananen@nokia.com

[2] Department of Interdisciplinary Studies
Tel Aviv University
Ramat Aviv 69978, ISRAEL
yaro@eng.tau.ac.il

Abstract

In an audio encoder such as AAC, the input PCM samples are sent in parallel to an MDCT and a psychoacoustic model that performs time-to-frequency decomposition once again with a FFT. To avoid this redundancy, we restructure the encoder to take the output of the MDCT as the input of the psychoacoustic model. Preliminary tests show this idea is feasible.

1. Introduction

In practically all state-of-the-art audio encoders, two basic coding tools are necessary:

- A suitable time-frequency decomposition serves to be an efficient digital representation of audio signals in a transform domain;
- A psychoacoustic model is representative of the most important auditory characteristics - masking threshold.

For deriving the psychoacoustic model, a perceptually meaningful time-frequency representation of the audio signal is necessary to capture the most relevant signal information. Most audio encoders available today, including those in MPEG-1 and MPEG-2 standards, use a parallel DFT for this purpose.

This paper describes a restructured perceptual audio encoder for improved computational efficiency. Similar effort has been reported in [1]. The optimization presented here is relevant for the real time implementation of an audio encoder such as MPEG-2 AAC. During this work, we have gained some insights into the MDCT (Modified Discrete Cosine Transform) transformation and the impact of the optimization to coding performance.

In order to replace the DFT with MDCT, we examine the relationship between MDCT, Shifted DFT (SDFT) [2] and DFT in the next section.

2. MDCT and Shifted DFT

The MDCT of a signal sequence a_k of $2N$ samples is defined as [3][4]:

$$\alpha_r = \sum_{k=0}^{2N-1} h_k a_k \cos\left[\pi \frac{\left(k + \frac{N+1}{2}\right)\left(r + \frac{1}{2}\right)}{N}\right], \tag{1}$$

where h_k is a window function. We assume an identical analysis-synthesis time window. The constraints of perfect reconstruction are [1]:

$$h_k = h_{2N-1-k} \tag{2}$$

$$h_k^2 + h_{k+N}^2 = 1 \tag{3}$$

A sine window is widely used in audio coding, because it offers good stop-band attenuation, provides good attenuation of the block edge effect and allows perfect reconstruction [1]. This sine window is defined as:

$$h_k = \sin\left[\pi \frac{(k + 1/2)}{2N}\right], \tag{4}$$

In the following we will prove that the MDCT is equivalent to a Shifted Discrete Fourier Transform (SDFT) [2] of a modified input signal. SDFT is a generalization of DFT that allows a possible arbitrary shift in position of the samples in the time and frequency domain with respect to the signal and its spectrum coordinate system [2].

The direct and inverse Shifted Fourier transforms are defined as [2]:

$$\alpha_r^{u,v} = (1/2N)^{1/2} \sum_{k=0}^{2N-1} a_k \exp[i2\pi(k+u)(r+v)/2N], \tag{5}$$

$$a_k^{u,v} = (1/2N)^{1/2} \sum_{r=0}^{2N-1} \alpha_r^{u,v} \exp[-i2\pi(k+u)(r+v)/2N], \tag{6}$$

where u and v represent the time and frequency domain shifts respectively.

We will now prove that the MDCT of a windowed signal $\tilde{a}_k$ of 2N samples is a SDFT of the alias-embedded signal $\hat{a}_k$ with $u = (N+1)/2, v = 1/2$.

For compactness we omit the normalization factor $(1/2N)^{1/2}$ in the following derivation, and introduce $SDFT((N+1)/2, 1/2)$ as the SDFT with the time domain shift of (N+1)/2 and the frequency domain shift of ½.

Denote $\tilde{a}_k = h_k \cdot a_k$, Then

$$\alpha_r = \sum_{k=0}^{2N-1} \tilde{a}_k \cos\left[\pi \frac{\left(k+\frac{N+1}{2}\right)\left(r+\frac{1}{2}\right)}{N}\right], \tag{7}$$

Represent the cosine term via complex exponents and split the summation in Eq. (7) into four parts:

$$\begin{aligned}
\alpha_r &= \frac{1}{2}\sum_{k=0}^{2N-1} \tilde{a}_k \left\{ \exp\left[i\pi \frac{\left(k+\frac{N+1}{2}\right)\left(r+\frac{1}{2}\right)}{N}\right] + \exp\left[-i\pi \frac{\left(k+\frac{N+1}{2}\right)\left(r+\frac{1}{2}\right)}{N}\right]\right\} = \\
&\frac{1}{2}\sum_{k=0}^{N-1} \tilde{a}_k \exp\left[i\pi \frac{\left(k+\frac{N+1}{2}\right)\left(r+\frac{1}{2}\right)}{N}\right] + \frac{1}{2}\sum_{k=0}^{N-1} \tilde{a}_k \exp\left[-i\pi \frac{\left(k+\frac{N+1}{2}\right)\left(r+\frac{1}{2}\right)}{N}\right] + \\
&\frac{1}{2}\sum_{k=N}^{2N-1} \tilde{a}_k \exp\left[i\pi \frac{\left(k+\frac{N+1}{2}\right)\left(r+\frac{1}{2}\right)}{N}\right] + \frac{1}{2}\sum_{k=N}^{2N-1} \tilde{a}_k \exp\left[-i\pi \frac{\left(k+\frac{N+1}{2}\right)\left(r+\frac{1}{2}\right)}{N}\right],
\end{aligned} \tag{8}$$

Replace the summation index k in the second and fourth terms of Eq. (8) with $N-1-k$ and $3N-1-k$ respectively. This results in:

$$\begin{aligned}
\alpha_r &= \frac{1}{2}\sum_{k=0}^{N-1} \tilde{a}_k \exp\left[i\pi \frac{\left(k+\frac{N+1}{2}\right)\left(r+\frac{1}{2}\right)}{N}\right] + \\
&\frac{1}{2}\sum_{k=0}^{N-1} \tilde{a}_{N-1-k} \exp\left[-i\pi \frac{\left(2N-\left(k+\frac{N+1}{2}\right)\right)\left(r+\frac{1}{2}\right)}{N}\right] + \\
&\frac{1}{2}\sum_{k=N}^{2N-1} \tilde{a}_k \exp\left[i\pi \frac{\left(k+\frac{N+1}{2}\right)\left(r+\frac{1}{2}\right)}{N}\right] + \\
&\frac{1}{2}\sum_{k=N}^{2N-1} \tilde{a}_{3N-1-k} \exp\left[-i\pi \frac{\left(4N-\left(k+\frac{N+1}{2}\right)\right)\left(r+\frac{1}{2}\right)}{N}\right],
\end{aligned} \tag{9}$$

or, described in another way:

$$\alpha_r = \frac{1}{2}\sum_{k=0}^{N-1}\tilde{a}_k \exp\left[i\pi\frac{\left(k+\frac{N+1}{2}\right)\left(r+\frac{1}{2}\right)}{N}\right]+$$

$$\frac{1}{2}\sum_{k=0}^{N-1}\tilde{a}_{N-1-k}\exp\left[-i2\pi\left(r+\frac{1}{2}\right)\right]\exp\left[i\pi\frac{\left(k+\frac{N+1}{2}\right)\left(r+\frac{1}{2}\right)}{N}\right]+$$

$$\frac{1}{2}\sum_{k=N}^{2N-1}\tilde{a}_k \exp\left[i\pi\frac{\left(k+\frac{N+1}{2}\right)\left(r+\frac{1}{2}\right)}{N}\right]+$$

$$\frac{1}{2}\sum_{k=N}^{2N-1}\tilde{a}_{3N-1-k}\exp\left[-i4\pi\left(r+\frac{1}{2}\right)\right]\exp\left[i\pi\frac{\left(k+\frac{N+1}{2}\right)\left(r+\frac{1}{2}\right)}{N}\right], \qquad (10)$$

Eq. (10) can be further simplified by substitution,

$$\exp\left[-i2\pi\left(r+\frac{1}{2}\right)\right]=-1, \qquad (11)$$

$$\exp\left[-i4\pi\left(r+\frac{1}{2}\right)\right]=1 \qquad (12)$$

Introduce time aliased signal:

$$\hat{a}_k = \begin{cases}\tilde{a}_k - \tilde{a}_{N-1-k}, & k=0,\ldots,N-1\\ \tilde{a}_k + \tilde{a}_{3N-1-k}, & k=N,\ldots,2N-1\end{cases} \qquad (13)$$

and finally obtain:

$$\sum_{k=0}^{2N-1}\tilde{a}_k\cos\left[\pi\frac{\left(k+\frac{N+1}{2}\right)\left(r+\frac{1}{2}\right)}{N}\right]=\frac{1}{2}\sum_{k=0}^{2N-1}\hat{a}_k\exp\left[i2\pi\frac{\left(k+\frac{N+1}{2}\right)\left(r+\frac{1}{2}\right)}{2N}\right] \qquad (14)$$

The left side of Eq. (14) is *MDCT* of the windowed signal $\tilde{a}_k$. The right side of Eq. (14) is $SDFT((N+1)/2,1/2)$ of the signal $\hat{a}_k$ formed from the initial windowed signal $\tilde{a}_k$ according to Eq. (13). For simplicity we have omitted the normalization factor $(1/2N)^{1/2}$ in both *MDCT* and $SDFT((N+1)/2,1/2)$ expressions. Physical interpretation of Eq. (13) is straightforward. Between 0 and (N-1) time samples the signal is mirrored and then inverted before being subsequently added to the original signal. Between N and (2N-1) time samples the signal is also mirrored and added to the original signal. Apparently, the mirrored terms in Eq. (13) are aliases. Therefore, MDCT coefficients can be obtained by adding the SDFT coefficients of the initial windowed signal and the alias. In other words, we can rewrite Eq. (14) as:

$$MDCT(signal) = SDFT_{(N+1)/2,1/2}(signal) + SDFT_{(N+1)/2,1/2}(alias) \tag{15}$$

The MDCT can be expressed by means of the conventional DFT as:

$$\begin{aligned}&\sum_{k=0}^{2N-1} \hat{a}_k \exp\left[i2\pi \frac{(k+(N+1)/2)(r+1/2)}{2N}\right] \\ &= \left\{\sum_{k=0}^{2N-1}\left[\hat{a}_k \exp\left(i2\pi \frac{k}{4N}\right)\right]\exp(i2\pi \frac{kr}{2N})\right\}\exp\left(i2\pi \frac{((N+1)/2)(r+1/2)}{2N}\right)\end{aligned} \tag{16}$$

To the right side of Eq. (16), the first exponential function corresponds to a modulation of $\hat{a}_k$ with a spectrum shift in frequency domain by ½ of the frequency-sampling interval, the second exponential function corresponds to the conventional DFT, and the third exponential function corresponds to a phase shift in the time domain. The above analysis has shown that MDCT of a signal is equivalent to DFT of a time aliased copy of this signal defined by Eq. (13) and therefore it is not point wise equivalent to DFT of the signal proper. Nevertheless one can assume that, with lower resolution, it may satisfactorily approximate the DFT spectrum. The following work is based on this assumption.

3. Restructured audio encoder structure

The modifications of the encoder are relatively minor, and affect mainly the psychoacoustic model part (See Figures 1 and 2). To avoid performing time-to-frequency decomposition twice in an AAC type audio encoder, we restructure the encoder to take the output of the MDCT-block as the input of the psychoacoustic model instead of FFT input. The psychoacoustic model is an excitation-pattern model discussed in our previous paper [5].

Since tonal and non-tonal components have very different masking characteristic [6][7], it is desirable to include a tonality measure in the masking threshold calculation. An unpredictability measure in the model described in the MPEG-2 AAC standard [8], which counts both magnitude and phase prediction errors within three consecutive frames of DFT spectra, is a good indicator of tonality measure for most test signals. We modify the calculation for the MDCT to count only magnitude prediction. However, this unpredictability measure clearly fails to serve as a tonality measure as with the DFT in the AAC recommended model. Therefore we have studied and compared the spectral characteristics of the DFT and MDCT, and investigated different tonality measures to solve this problem.

4. Some experimental results

We have performed some preliminary tests with a fixed bitrate of 64 kbps/channel, and have compared the resulting audio quality. Test results show that the tonality measure is critical with signals such as pitchpipe (clearly degraded audio quality, if the unpredictability measure is not used), and less critical with signals such as symphony orchestra.

Our test results show that a simple tonality measure similar to MPEG-1 model 1 works well. We have achieved virtually the same coding performance with reduced computational complexity, with the exception of speech signals, which show some slight quality degradations. Several samples such as pitchpipe, speech and symphony orchestra signals have been tested. Tests have been conducted by integrating the modified psychoacoustic model into an MPEG-2 AAC type audio codcr that contains only the basic coding tools. In order to achieve better results, further optimization is still needed.

In Figure 3 – 4, we directly compare the DFT and MDCT power spectra. The MDCT seems to have a slightly better frequency resolution in comparison with the DFT.

In Figure 5.- 6, we compare the DFT and SDFT((N+1)/2, 1/2) power spectra. They are quite similar to each other.

In Figure 7 – 8, we compare the excitation levels in each scale factor band using the MDCT and DFT respectively. It is clear that their energies in each scale factor band are very close to each other. The differences of the two methods are within 2 dB. This is valid for all test signals we have tested.

5. Conclusion

We have successfully replaced the DFT block in the psychoacoustic model with the MDCT coefficients already available in the main time-frequency decomposition. Although DFT and MDCT spectra are not identical, their energies in respective Equivalent Rectangular Band (ERB) are very close to each other. This makes the above replacement feasible.

6. Acknowledgement

L. Yaroslavsky acknowledges Tampere International Center for Signal Processing, Tampere University of Technology (TUT) for supporting this research during his visiting professorship at TUT.

7. Reference

[1] Ferreira, A., “Spectral Coding and Post-Processing of High Quality Audio”, PhD thesis, *http://telecom.inescn.pt*, 1998.

[2] Yaroslavsky, L., Eden, M., “Fundamentals of Digital Optics”, Birkhauser, Boston, 1996.

[3] Princen, J., Bradley A., “Analysis/Systhesis Filter Bank Design Based on Time Domain Aliasing Cancellation”, IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. ASSP-34, No.5, October 1986.

[4] Malvar, H., “Signal Processing with Lapped Transform”, Artech House, Boston, 1991

[5] Wang, Y., Vilermo, M. "En excitation level based psychoacoustic model for audio compression", The 7th ACM International Multimedia Conference, October 30 to November 4, 1999 Orlando, Florida, USA.

[6] Moore B. C. J., (1997) "An Introduction to the Psychology of Hearing", 4. Edition, Academic Press, London.

[7] E. Zwicker, H. Fastl, "Psychoacoustics, Facts and Models", Springer-Verlag, 1990.

[8] ISO/IEC JTC1/SC29/WG11, "Coding of moving pictures and audio- MPEG-2 Advanced Audio Coding AAC", ISO/IEC 13818-7 International Standard, 1997.

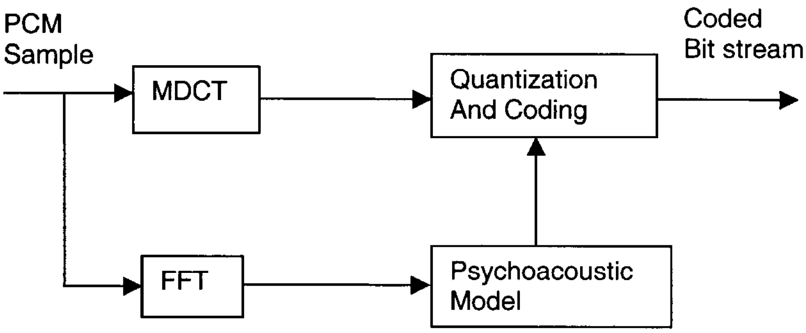

Figure 1. A simplified conventional audio encoder structure

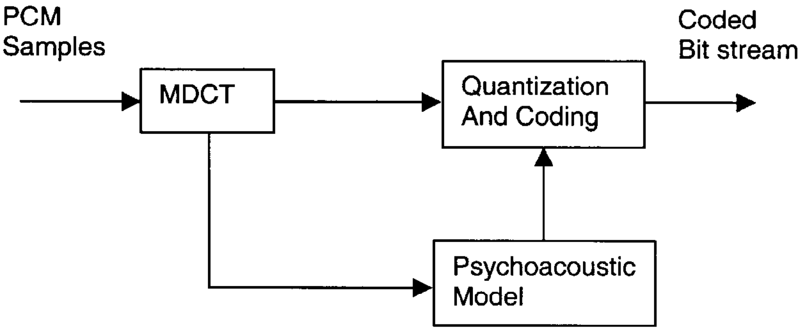

Figure 2. A simplified restructured audio encoder structure

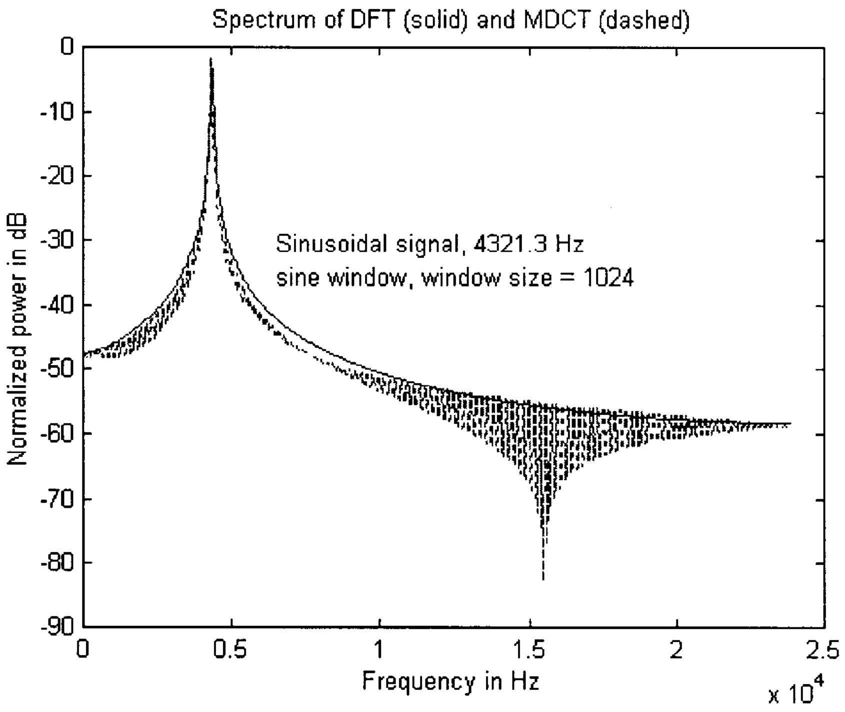

Figure 3. Frame-wise comparision of the DFT and MDCT power spectra of a sinusoidal signal

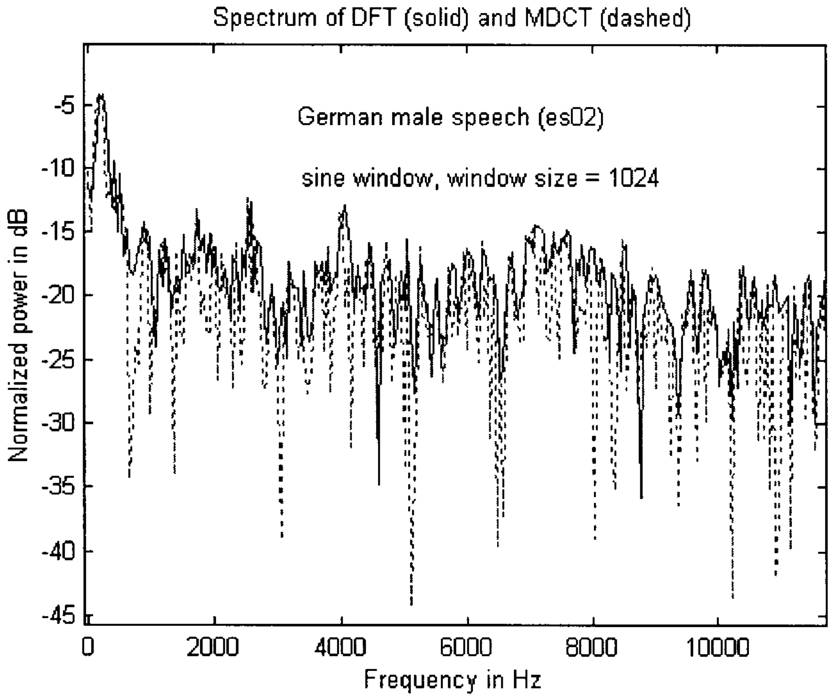

Figure 4. Frame-wise comparision of the DFT and MDCT power spectra of a German male speech signal

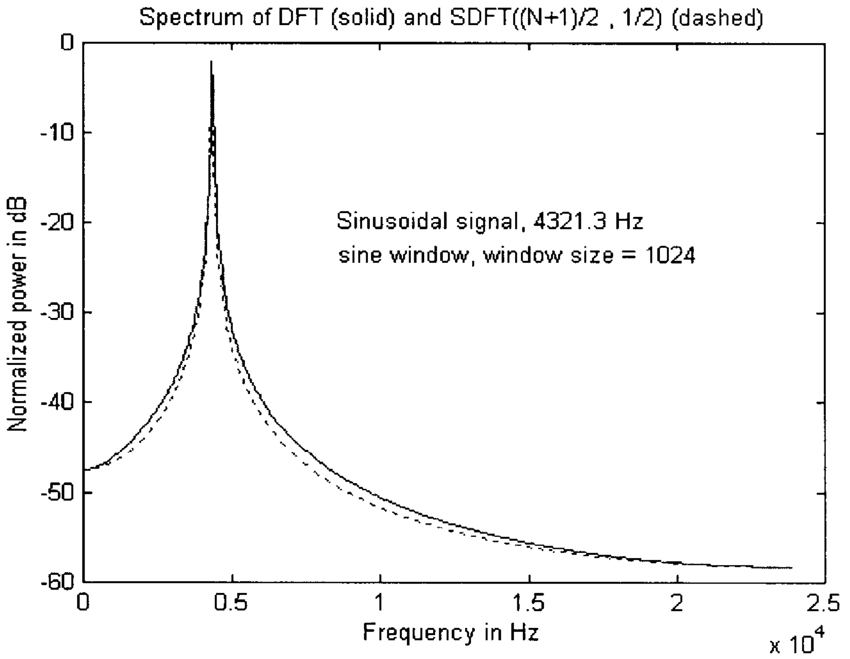

Figure 5. Frame-wise comparision of the DFT and SDFT((N+1)/2, 1/2) power spectra of a sinusoidal signal

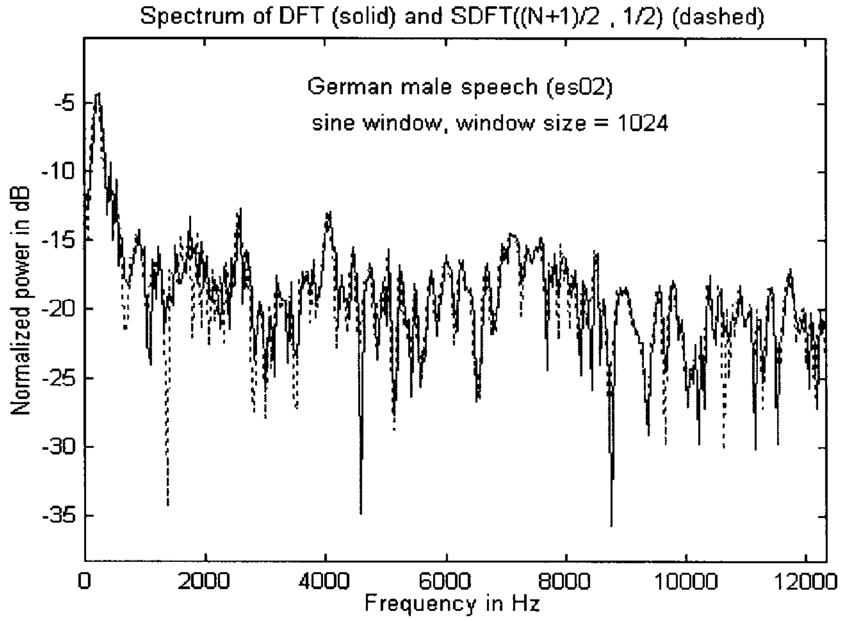

Figure 6. Frame-wise comparision of the DFT and SDFT((N+1)/2, 1/2) power spectra of a German male speech signal

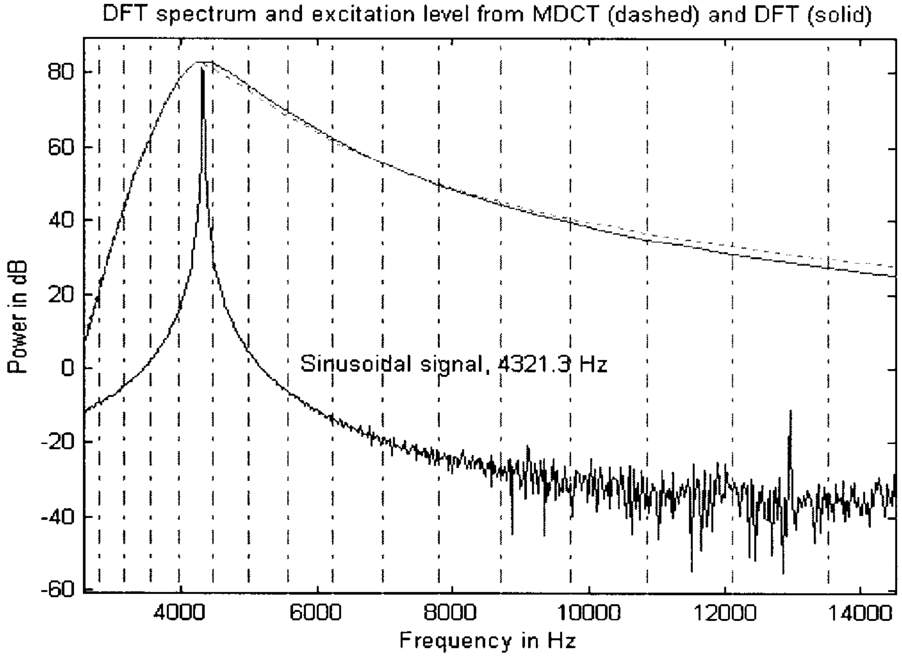

Figure 7. Frame-wise comparision of the DFT spectrum and the excitation levels using the MDCT and DFT respectively. A sinusoidal signal is used. The dashdot lines indicate the Equivalent Rectangular Bands (ERBs).

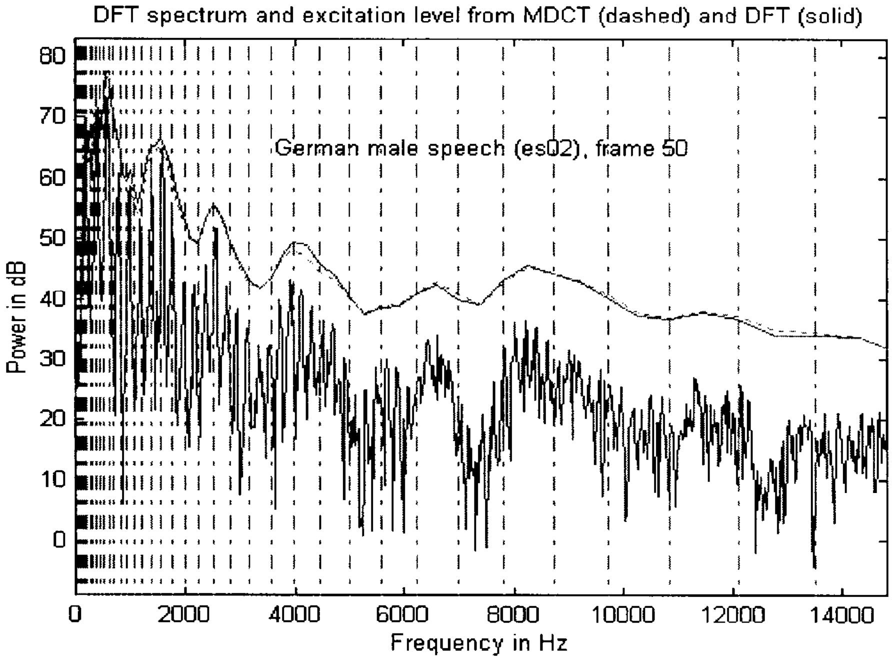

Figure 8. Frame-wise comparision of the DFT spectrum and theexcitation levels using the MDCT and DFT respectively. A German male speech signal is used. The dashdot lines indicate the Equivalent Rectangular Bands (ERBs).