HARMONIC-BAND COMPLEX WAVELET TRANSFORM AUDIO ANALYSIS AND SYNTHESIS

*A.R. Gillan, *P.R. Hill*

Electrical and Electronic Engineering, The University of Bristol, Bristol, BS8 1UB, UK

ABSTRACT

A novel improvement to an existing wavelet representation is presented within the context of analysis and synthesis of harmonic audio signals. The approach replaces the discrete wavelet transform stage in the harmonic-band wavelet transform with the dual-tree complex wavelet transform. The harmonic-band wavelet transform lacks explicit phase information, is subject to considerable transform domain coefficient aliasing and its filters exhibit poor frequency selectivity and high sidelobe levels. Integration of the dual-tree complex wavelet transform mitigates all of these disadvantages and significantly improves the accuracy of signal synthesis where signals are synthesised through autoregressive analysis and linear prediction of transform domain coefficients.

Index Terms— Wavelet, Audio, Synthesis

1. INTRODUCTION

A particular approach to characterising acoustically-generated harmonic sounds has been adopted through the signal-adaptive spectral representation of the harmonic-band wavelet transform (HBWT) [1]. This transform, which employs a hybridisation of the discrete wavelet transform (DWT) with the modified discrete cosine transform (MDCT), provides a compelling model for analysis and synthesis of such signals and the new transform presented here is based on its implementation. Power spectra that the basic HBWT [2] is particularly suited to analyse are comprised of approximately harmonically-related peaks with the power spectrum neighbouring each harmonic decaying as an inverse power of the distance in frequency away from it. The motivation behind the transform's definition was to formulate a new means of spectral segmentation for representing the harmonic and stochastic components of these signals. Accordingly, the outcome is such that each harmonic may be individually analysed and that spectral intervals surrounding each harmonic (the stochastic component) are divided into sub-bands, each of which may also be individually analysed.

Fractal additive synthesis (FAS) – a method based on the HBWT – has been described to improve upon widely adopted STFT-based models [2], in particular, spectral modelling synthesis (SMS). In SMS [3] the harmonic component is synthesised based on sinusoidal analysis of the original signal and then subtracted from it to produce a residue. The stochastic component is generated through approximating this residue's entire spectral envelope. In contrast, synthesis of the stochastic component by FAS is dependent upon an explicit, spectrally-segmented, HBWT model of the stochastic component in the original signal. A more detailed and explicit analysis was seen to overcome limitations posed by SMS in the effective synthesis of the stochastic component – essential to emulating the specific and vibrant natural complexity of acoustically-generated harmonic sounds.

*Corresponding Author: paul.hill@bristol.ac.uk

1.1. The Modified Discrete Cosine Transform and Cosine-Modulated Filter Banks

The MDCT [4][5] is a lapped transform based on the type-IV discrete cosine transform (DCT-IV). An MDCT is in fact exactly equivalent to a DCT-IV of size N, where the 2N inputs have been preprocessed with N additions and subtractions. This transform can effectively reduce block boundary artefacts and due to time-domain aliasing cancellation, perfect reconstruction can be achieved [6][7]. By applying an MDCT to a discrete time signal $s(l)$, an expansion in terms of a set of sequences is obtained by shifting a cosine-modulated, low-pass prototype filter $w(l)$ [1]. To compute the expansion, a P-channel cosine-modulated filter bank (CMFB) with re-sampling factor P can be employed. The synthesis filter impulse response is obtained by time reversal of the analysis filter impulse response [1]. P here is set to the length in samples of the average fundamental period of a roughly or pseudo-periodic input signal, so it follows that the disclosed method is defined only for pitch-stable sounds. Each filter bank channel except for the first and last is tuned to a single sideband of a harmonic of the input signal. P-order decimation of the CMFB without aliasing caused by the decimation is made possible as the bandwidth of each of its outputs is $1/(2P)$. Integer factor P is calculated by

$$P = \left\lfloor \frac{f_s}{f_0} + 0.5 \right\rfloor \tag{1}$$

where f_s is the sampling frequency of the pseudo-periodic input signal and f_0 is its fundamental frequency. The MDCT basis functions [1] can be written

$$q_{p,r}(l) = q_{p,0}(l - rP) \qquad p = 0,...,P-1;\ r \in Z \tag{2}$$

with

$$q_{p,0}(l) = w(l) cos\left[\left(l + \frac{P+1}{2}\right)\left(p + \frac{1}{2}\right)\frac{\pi}{P}\right] \tag{3}$$

where the low-pass prototype filter

$$w(l) = \frac{1}{\sqrt{2P}} sin\left[\left(l + \frac{1}{2}\right)\frac{\pi}{2P}\right] \tag{4}$$

satisfies the symmetry conditions in [8].

1.2. The Dual-Tree Complex Wavelet Transform

The dual-tree complex wavelet transform (DT-ℂWT) has been shown to overcome drawbacks inherent in the discrete wavelet transform (DWT) [9] and has been successfully employed for both image and audio texture synthesis [10] [11]. Advantages that the DT-ℂWT offers over the DWT include:

- near shift invariance,
- the availability of explicit phase information [12],
- substantially reduced aliasing in the transform domain,

- an enhanced emphasis of positive frequencies and rejection of negative frequencies (or vice-versa).

Any finite energy signal $s(l)$ may be expanded on a DT-$\mathbb{C}$WT set

$$s(l) = \sum_{n=1}^{N} \sum_{m=-\infty}^{\infty} d_n^c(m)\psi_{n,m}^c(l) + \sum_{m=-\infty}^{\infty} c_N^c(m)\varphi_{N,m}^c(l) \quad (5)$$

where index n represents scale and index m represents the time-shift. The signal $s(l)$ is projected over a basis comprised of scaled and time-shifted versions of a complex fundamental wavelet $\psi^c = \psi^r + j\psi^i$ and a time-shifted version of a corresponding complex scaling function $\varphi^c = \varphi^r + j\varphi^i$. $d_n^c(m) = d_n^r(m) + jd_n^i(m)$ and $c_N^c(m) = c_N^r(m) + jc_N^i(m)$ are the DT-$\mathbb{C}$WT complex wavelet coefficients and corresponding complex scaling coefficients respectively. An efficient implementation of the DT-$\mathbb{C}$WT is a filter bank structure composed of two real DWTs with iterated Hilbert pairs of filters producing the real and imaginary components of the transform [9].

2. IMPLEMENTATION

2.1. Cosine-Modulated Filter Banks

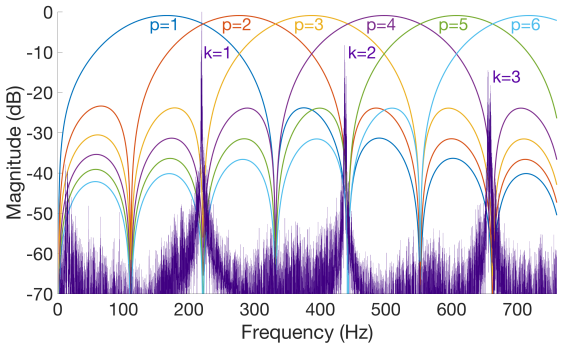
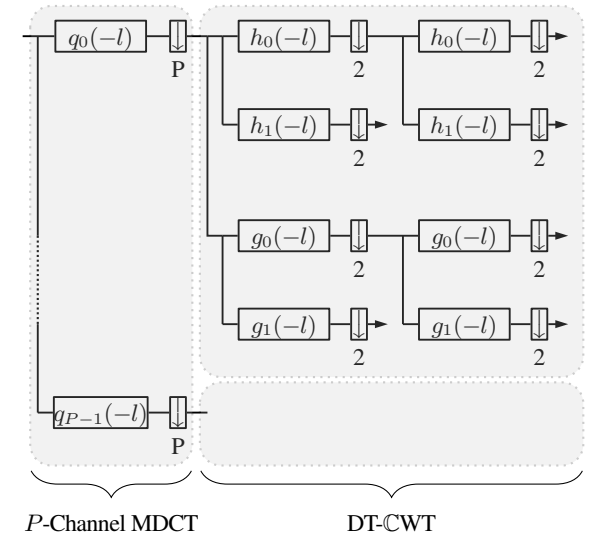


Fig. 1. The frequency response of CMFB channels $p=1$ to $p=6$ with $P=200$, and the magnitude Fourier transform of an input signal showing its first 3 harmonic peaks, $k=1$ to $k=3$.

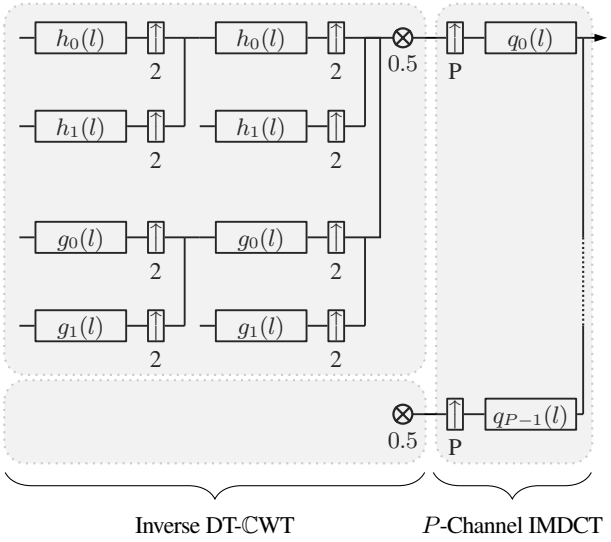
All signals within the class of pseudo-periodic signals for which this method is defined exhibit spectra comprised of harmonic peaks

$$k = 1,2...,\lfloor P/2 \rfloor - 1$$

approximately separated by intervals of the fundamental frequency f_0. As mentioned previously, the bandwidth of a P-channel CMFB filter is $\Delta f = 1/(2P)$. This results in the two sidebands of each of the k harmonics being passed by a pair of filters i.e. a sideband per filter. The index of the filter corresponding to the lower sideband is $p = 2k-1$ and the index corresponding to the upper sideband is $p = 2k$ [2]. Our implementation uses a fast modulated lapped transform provided by the LT-toolbox [13] for MATLAB® where the length L of filters is constrained to $L = 2P$. Figure 1 shows a plot of 6 forward transform CMFB filter responses overlaid on the spectrum of an input signal (String Section in Table 1) with $f_s = 44.1$ kHz, $f_0 = 220$ Hz and therefore $P = 200$ as per (1).



(a) Analysis filter bank.



(b) Synthesis filter bank.

Fig. 2. A $N=2$-level harmonic-band dual-tree complex wavelet transform filter bank, cascaded as a P-channel cosine-modulated filter bank with P dual-tree complex wavelet transform filter banks.

2.2. The Harmonic-Band Dual-Tree Complex Wavelet Transform

The crux of this synthesis by analysis technique is the formulation of a transform method which can 'intelligently' decompose the spectrum of the class of signals being processed, allow for effective manipulation of its transform domain coefficients and perfectly reconstruct its input. The harmonic-band dual-tree complex wavelet transform (HBDT-$\mathbb{C}$WT) is introduced as an improvement to the HBWT, integrating the advantages of the DT-$\mathbb{C}$WT into the existing HBWT framework.

The HBDT-$\mathbb{C}$WT is implemented as the cascade of a P-channel cosine-modulated filter bank (CMFB) with P DT-$\mathbb{C}$WT filter banks depicted in Figure 2. The inverse HBDT-$\mathbb{C}$WT reverses the process of the forward HBDT-$\mathbb{C}$WT through appropriate upsampling and inverse filtering by the transform stages discussed previously, perfectly

reconstructing input signals. The discrete-time harmonic-band dual-tree complex wavelet function is defined by

$$\xi^c_{n,m,p}(l) = \sum_r \psi^c_{n,m}(r) q_{p,r}(l) \tag{6}$$
$$n = 1,2,...,N;\ m \in Z\ p = 0,1,...,P-1$$

where n is wavelet scale, m is time-shift, $\psi^c_{n,m}(r)$ is the discrete time complex wavelet function and $q_{p,r}(l)$ is the MDCT basis of (2). The discrete time harmonic-band complex scaling function is defined by

$$\zeta^c_{N,m,p}(l) = \sum_r \varphi^c_{N,m}(r) q_{p,r}(l) \tag{7}$$
$$m \in Z;\ p = 0,1,...,P-1$$

where $\varphi^c_{N,m}(r)$ is the discrete time complex scaling function. Based on the orthogonality and completeness conditions for the HBWT given in [1] it can be shown that any signal $s(l) \in l^2$ can be expanded on a HBDT-ℂWT set

$$s(l) = \sum_{p=0}^{P-1} \left(\sum_{n=1}^{N} \sum_m b^c_{p,n}(m) \xi^c_{n,m,p}(l) + \sum_m a^c_{p,N}(m) \zeta^c_{N,m,p}(l) \right) \tag{8}$$

where $b^c_{p,n}(m)$ are the HBDT-ℂWT complex wavelet coefficients and $a^c_{p,N}(m)$ are the HBDT-ℂWT complex scaling coefficients.

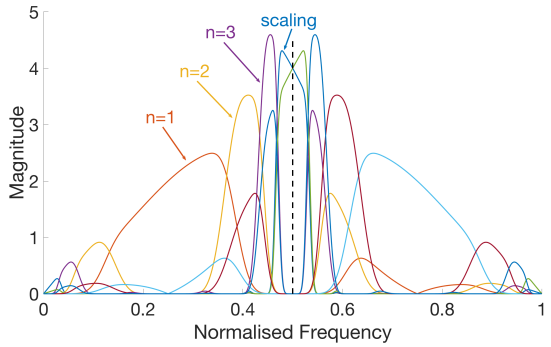
It is to be noted that for odd HBDT-ℂWT channels ($p = 1,3,...,P-1$) the complex conjugate of the DT-ℂWT filter impulse responses is convolved with the CMFB filter impulse response for that channel. For even HBDT-ℂWT channels ($p = 0,2,...,P-2$), the original DT-ℂWT filter impulse responses are convolved with the CMFB filter impulse response for that channel. The motivation for this operation is to, for adjacent channels, generate a symmetrical filter frequency response (Figure 3 (b)) around the spectral location of harmonic peaks in the input signal, as with the HBWT. In our implementation, transform coefficients produced by odd HBDT-ℂWT channels are subject to complex conjugation rather than the DT-ℂWT filters themselves.

Figure 3 depicts sets of filter bands shaped to decompose two sidebands and a harmonic of a signal with a fundamental frequency in the middle of the spectrum. An important advantage of the HBDT-ℂWT over the HBWT is illustrated in these plots. Namely, the emphasis of a single side of the frequency spectrum and the rejection of the opposite side – a feature of the DT-ℂWT described in [14]. Combined with relatively high sidelobe levels on the required side of each individual harmonic location, a substantial wavelet filter sidelobe intrusion into spectral territory on the opposing side of individual harmonic locations is obvious in (a) of Figure 3. Additionally, the scaling filter responses seen in this plot show virtually no spectral discrimination.

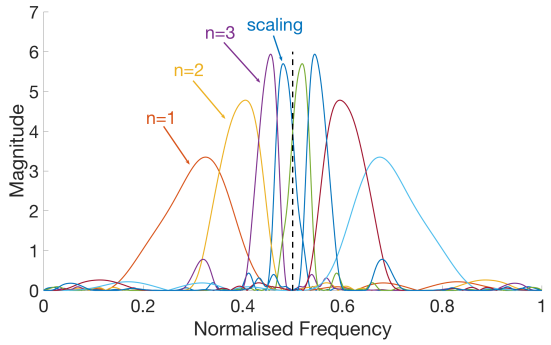
In contrast, the HBDT-ℂWT filter responses in (b) of Figure 3 exhibit low overall sidelobe levels, limited sidelobe intrusion into territory on the opposing side of the harmonic location and significant improvement in scaling filter spectral discrimination. The synergy of these characteristics accounts for an improvement in the precision of the targeted decomposition of harmonics and their sidebands offered by the HBWT.

2.3. Transform Coefficient Analysis and Synthesis

Once HBDT-ℂWT coefficients are obtained from an input signal, new coefficients can be generated based on their analysis and subsequently fed into an inverse HBDT-ℂWT to produce a synthetic signal. To achieve this,



(a) HBWT filters.



(b) HBDT-ℂWT filters.

Fig. 3. Frequency response of (a) HBWT filters and (b) HBDT-ℂWT filters, for $P = 4$ and $N = 3$, plotted for channels $p = 1$ and $p = 2$. The sub-band decomposition is of a single harmonic (depicted by the dashed black line) and shows the lower and upper sideband filters. Filters are labeled only for channel $p = 1$.

the known $b^c_{p,n}$ and $a^c_{p,N}$ coefficients are first analysed by autoregressive (AR) spectral estimation. Using the estimated AR parameters, new coefficients $\hat{b}^c_{p,n}$ and $\hat{a}^c_{p,N}$ are generated by driving AR finite difference equations with white Gaussian noise. As the amplitude histograms of the scaling coefficients representing the harmonic component of our input signals were observed to be normally distributed, it was assumed that they could be synthesised autoregressively as well as the wavelet coefficients.

A linear difference equation describing the most general autoregressive moving average (ARMA) model of a time series [15] is defined as

$$x(m) = -\sum_{i=1}^{v} \alpha(i) x(m-i) + \sum_{i=0}^{w} \beta(i) u(m-i), \tag{9}$$

relating an input driving sequence $u(m)$ to a wide-sense stationary (WSS) discrete random output process $x(m)$. $\alpha(i)$ are the AR parameters and $\beta(i)$ are the moving average (MA) parameters. (9) also represents a linear time-invariant infinite impulse response (IIR) difference equation where w is the feedforward filter order, $\beta(i)$ are the feedforward filter coefficients, v is the feedback filter order, $\alpha(i)$ are the feedback filter coefficients, $u(m)$ is the input signal and $x(m)$ is the output signal [16].

The AR or all-pole model used in our implementation is defined by setting $\beta(0) = 1$ and $\beta(i) = 0, \forall i > 0$. $A(z) = \sum_{i=0}^{v} \alpha(i) z^{-i}$ where $\alpha(0) = 1$ is assumed to have all of its zeros within the z-plane unit circle guaranteeing $H(z) = \frac{1}{A(z)}$, the rational transfer function between the input $u(m)$ and output $x(m)$, to be a stable and causal filter [15].

	RMSE		$\overline{\mathbf{C_{xy}}}$	
Input Sound	$\mathbb{C}$	$\mathbb{R}$	$\mathbb{C}$	$\mathbb{R}$
Alto Flute	0.178	0.324	0.512	0.398
Contrabass Clarinet	0.153	0.273	0.518	0.402
Viola	0.162	0.324	0.596	0.447
Cello	0.110	0.181	0.494	0.386
Woodwind Section	0.143	0.268	0.397	0.323
Horn Section	0.089	0.325	0.470	0.359
Euphonium Section	0.141	0.309	0.313	0.234
String Quartet	0.101	0.205	0.505	0.378
String Section	0.073	0.282	0.519	0.406
Full Orchestra	0.077	0.258	0.443	0.365

Table 1. **RMSE** and magnitude squared coherence averaged across the entire spectrum ($\overline{\mathbf{C_{xy}}}$) for the acoustic-instrument-based input signals listed. All comparisons are between synthesised signals and the portion of the original signals which they are predicting. $\mathbb{C}$ denotes results from the HBDT-$\mathbb{C}$WT and $\mathbb{R}$ denotes results from the HBWT.

Driving sequence $u(m)$ is assumed to be a zero-mean white Gaussian noise process with variance σ^2 and thus power spectral density (PSD) σ^2. Setting $\beta(0)=1$ allows the gain term of the AR filter to become the variance σ^2 of $u(m)$. Finding estimates of the optimal AR parameters $\{\hat{\alpha}(1),\hat{\alpha}(2),\ldots,\hat{\alpha}(v)\}$ and noise variance $\hat{\sigma}^2$ enables $AR(v)$ linear prediction

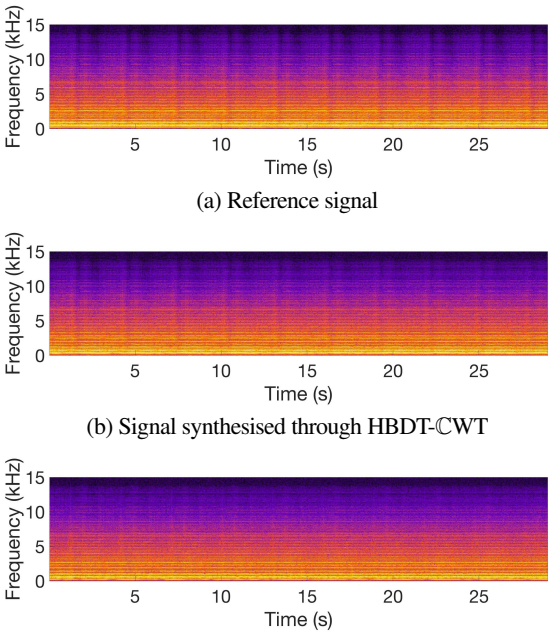
$$\hat{x}(m)=-\sum_{i=1}^{v}\hat{\alpha}(i)x(m-i)+\hat{u}(m). \tag{10}$$

The approach of solving the Yule-Walker equations to obtain AR parameter estimates directly as in [1] has been shown to deliver less reliable estimates and have potential for model instability [17]. Burg's method is an alternative approach in which reflection coefficients estimates are calculated first and subsequently AR parameters are obtained by Levinson recursion. This method has been shown to provide more reliable results and guarantee model stability [15][17] and is thus employed in our implementation.

A scheme for extrapolation of 1D signals by linear prediction discussed in [18] and [19] is facilitated by an initialised, purely recursive, transposed direct form II IIR filter. (10) with $\hat{u}(m)=0,\forall m$ is assumed to predict process $x(m)$ as $\hat{x}(m)$ (in our case predicting coefficients as $\hat{b}^c_{p,n}$ and $\hat{a}^c_{p,N}$) and the filter is initialised with all known samples $\{x(m-1),x(m-2),\ldots,x(m-v)\}$ (in our case the known $b^c_{p,n}$ and $a^c_{p,N}$ coefficients) as well as estimated AR coefficients $\{\hat{\alpha}(1),\hat{\alpha}(2),\ldots,\hat{\alpha}(v)\}$. An inherent shortcoming of this procedure however, is that due to the necessity for stability, the attenuation of the all-pole IIR filter leads to amplitude decay in the filter's output over time. This approach is therefore not employed and input $\hat{u}(m)$ is provided as a white Gaussian noise process scaled by its variance $\hat{\sigma^2}$ which is also estimated by Burg's method [15]. Filter initialisation increases the amount of information available to predictors and was not previously considered in [1].

3. EXPERIMENTAL RESULTS

The experimental results have been derived from a set of pseudo-harmonic, approximately pitch-stable, vibrato-free sounds produced by the recorded orchestral sections, ensembles and individual instruments listed in Table 1. This set is representative of the class of sounds for which the disclosed analysis and synthesis method in its current form is particularly suited. All input signals are monophonic, have a pitch of A3 (220 Hz) and a sample rate of 44.1 kHz. Sound files were rendered from a number of contemporary orchestral sample libraries in the software sampler environment Kontakt®. Individual recordings played by Kontakt® are repeated for the extent of each of our rendered files and amplitude-faded between repetitions to create the illusion of the orchestral instruments sustaining a note for the duration of each rendered file. As can be seen in Figure 4 (a) this results in noticeable but not substantial amplitude variation. This is also present in Figure 4 (b) but almost absent in (c).



(a) Reference signal

(b) Signal synthesised through HBDT-$\mathbb{C}$WT

(b) Signal synthesised through HBWT

Fig. 4. Spectrograms of (a) the Horn Section (see Table 1) and (b)(c) the signals synthesised based on (a) through linear prediction of the coefficients of the HBDT-$\mathbb{C}$WT and HBWT.

To quantitatively examine the merit of autoregressive HBDT-$\mathbb{C}$WT synthesis over its HBWT counterpart a decomposition and autoregressive analysis of transform coefficients of 1,280,200 samples of each sound is performed. Coefficient synthesis and subsequent inverse transform reconstruction then produces an output signal of the same length as the input signal. The synthesised sounds are compared in Table 1 by RMSE and magnitude squared coherence averaged across the entire spectrum ($\overline{\mathbf{C_{xy}}}$) to the 1,280,200 samples of each input sound immediately following the samples on which analysis was performed. The decomposition level for both transforms is set to $N=4$. As can be seen from Table 1, autoregressive HBDT-$\mathbb{C}$WT synthesis exhibits superiority for all input signals in both the time domain and the spectral domain. Qualitatively, the signals synthesised through the HBDT-$\mathbb{C}$WT are almost indistiguishable from the reference signals. The signals synthesised through the HBWT counterpart however, bear less perceptual resemblance to the reference signals, with audible spectral distortion.

4. CONCLUSION

An improvement to the harmonic-band wavelet transform based autoregressive synthesis model has been presented through the introduction of the dual-tree complex wavelet transform. The results demonstrate substantial qualitative and quantitative enhancement of the existing technique.

5. REFERENCES

[1] P. Polotti, *Fractal Additive Synthesis: Spectral Modelling of Sound for Low Rate Coding of Quality Audio*, Ph.D. thesis, Lausanne, 2003.

[2] P. Polotti and G. Evangelista, "Fractal additive synthesis," *IEEE Signal Processing Magazine*, vol. 24, no. 2, pp. 105–115, March 2007.

[3] X. Amatriain, J. Bonada, A. Loscos, and X. Serra, "Spectral processing," in *DAFX: digital audio effects*, chapter 10, pp. 393–445. John Wiley & Sons, 2011.

[4] J. P. Princen and A. B. Bradley, "Analysis/synthesis filter bank design based on time domain aliasing cancellation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 5, pp. 1153–1161, 1986.

[5] J. P. Princen, A. W. Johnson, and A. B. Bradley, "Subband/transform coding using filter bank designs based on time domain aliasing cancellation," in *ICASSP'87. IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 1987, vol. 12, pp. 2161–2164.

[6] X. Shao and S. G. Johnson, "Type-iv dct, dst, and mdct algorithms with reduced numbers of arithmetic operations," *Signal Process.*, vol. 88, no. 6, pp. 1313–1326, June 2008.

[7] H. S. Malvar, *Signal Processing with Lapped Transforms*, Artech House, 1992.

[8] T. Q. Nguyen and R. D. Koilpillai, "The theory and design of arbitrary-length cosine-modulated filter banks and wavelets, satisfying perfect reconstruction," *IEEE Transactions on Signal Processing*, vol. 44, no. 3, pp. 473–483, Mar 1996.

[9] I. W. Selesnick, R. G. Baraniuk, and N. C. Kingsbury, "The dual-tree complex wavelet transform," *IEEE Signal Processing Magazine*, vol. 22, no. 6, pp. 123–151, Nov 2005.

[10] C. Gallagher and A. Kokaram, "Nonparametric wavelet based texture synthesis," *IEEE International Conference on Image Processing 2005*, vol. 2, pp. II–462, Sept 2005.

[11] D. O'Regan and A. Kokaram, "Multi-resolution sound texture synthesis using the dual-tree complex wavelet transform," *2007 15th European Signal Processing Conference*, pp. 350–354, Sept 2007.

[12] F. C. A. Fernandes, R. L. C. van Spaendonck, and C. S. Burrus, "A new framework for complex wavelet transforms," *IEEE Transactions on Signal Processing*, vol. 51, no. 7, pp. 1825–1837, July 2003.

[13] J. Alhava, *LT-toolbox v0.99*, Tampere, Finland, 2013.

[14] N. G. Kingsbury, "The dual-tree complex wavelet transform: a new technique for shift invariance and directional filters," *IEEE Digital Signal Processing Workshop*, vol. 86, pp. 120–131, 1998.

[15] S. M. Kay, *Modern Spectral Estimation: Theory & Application*, Prentice-Hall, 1988.

[16] S. Wang, *Enhancing brain-computer interfacing through advanced independent component analysis techniques*, Ph.D. thesis, University of Southampton, 2009.

[17] M.J.L. de Hoon, T.H.J.J. van der Hagen, H. Schoonewelle, and H. van Dam, "Why yule-walker should not be used for autoregressive modelling," *Annals of nuclear energy*, vol. 23, no. 15, pp. 1219–1228, 1996.

[18] I. Kauppinen and K. Roth, "Audio signal extrapolation: Theory and applications," in *Proc. DAFx*, 2002, pp. 105–110.

[19] M. Fink, M. Holters, and U. Zölzer, "Comparison of various predictors for audio extrapolation," in *Proc. of the 16th Int. Conference on Digital Audio Effects (DAFx-13)*, 2013, pp. 1–7.