NEW FAST WAVELET PACKET TRANSFORM ALGORITHMS FOR FRAME SYNCHRONIZED SPEECH PROCESSING

Andrzej Drygajlo

Signal Processing Laboratory, Swiss Federal Institute of Technology Lausanne
CH-1015 Lausanne, Switzerland, E-mail: andrzej.drygajlo@lts.de.epfl.ch

ABSTRACT

In this paper we present orthogonal overlapped block transforms as a frame synchronized signal analysis tool with the capability of arbitrary multiresolution time-spectral decomposition of speech signals. Our prime interest is in the representation of nonstationary discrete-time signals in terms of wavelet packets, and we concentrate on their fast transform algorithms. Wavelet packet representations provide a local time-spectral description which reveals the nonstationary nature of a signal. They allow the speech signal to be accurately parameterised for such applications as speech and speaker recognition, where a front-end is responsible for the frame synchronized feature extraction. In this case the fast overlapped block transform algorithms represent an elegant and efficient solution to the implementation of wavelet packet transforms, since their FFT-like lattice block structure provides all possible multiresolution time-spectral coefficients. The frame synchronization is also preserved in subbands which allows a new subband-based approach for speech recognition.

1. INTRODUCTION

Time-frequency transforms, including discrete wavelet packet transforms (DWPT) [13, 4], are generally acknowledged to be useful for studying non-stationary signals and, in particular, have been shown to be of value in the perceptually tuned analysis and parameterization of speech signals. Although there have been several studies reported in the literature, there is still considerable work to be done investigating the utility of wavelet packet time-frequency transforms for speech processing applications, e.g. for speech and speaker recognition, where a front-end processing is responsible for the feature extraction from the incoming digitized speech. This front-end produces a stream of vectors that may represent the time-spectral characteristics of speech over short, possibly overlapping intervals. Recent literature in this area reports that the use of the DWPT coefficients for input to classical recognisers such as those based on Hidden Markov Models (HMM) is difficult because they are designed to accept frame synchronous data, and the wavelet packet transforms realizations represent a filter bank approach with different sampling rates in subbands. Many researchers in the speech processing domain believe that frame duration is not a parameter of the wavelet packet transforms, and the only acceptable solution is to use the sampled continuous wavelet packet transforms [9]. Unfortunately, this produces frame synchronous data in a redundant and inefficient fashion and does not retain all the features that are offered by the DWPT.

This paper contributes to this ongoing investigation through the development of frame synchronized wavelet packet transform algorithms. Our prime interest is in the representation of speech signals in terms of discrete orthogonal wavelet packets used as a signal analysis system with the capability of arbitrary multiresolution time-spectral decomposition, and we concentrate on the fast transform algorithms for such systems.

2. FRAME SYNCHRONIZED SPEECH ANALYSIS

Most parameters in a speech or speaker recognition system are computed on a frame-by-frame basis [10]. Frame duration is defined as the length of time (or number of samples) over which a set of parameters is valid. Frame period is a similarly used term that denotes the length of time between successive parameter calculations.

Equally important is the interval over which the frame parameters are computed. The number of samples used to do this is known as the window duration. Window duration controls the amount of averaging used in the parameters calculation. The frame duration and window duration together control the rate at which the parameters track the dynamics of the signal.

Much of our current knowledge and intuition of speech is derived from analysis involving assumptions of short-time stationarity (e.g. Short-Time Fourier Transform (STFT) based analysis). The STFT-based algorithms use overlapped windows of a fixed duration to perform the spectral estimation through an FFT algorithm. The time and frequency resolu-

tions are fixed by the length of the window and the FFT, and are the same across the entire time/frequency range. This type of analysis is referred to as an overlapping analysis, because with each new frame, only a fraction of the signal data changes. The amount of overlap to some extent controls how quickly parameters can change from frame to frame. Frame duration and window duration are normally adjusted as a pair. Generally, since a shorter frame duration is used to capture rapid dynamics of the spectrum, the window duration should also be correspondingly shorter so that the detail in the time-frequency representation is not excessively smoothed. Such methods are, by their very nature, incapable of revealing the true nonstationary nature of speech.

The motivation for using the DWPT is to obtain more discriminative information by providing a different resolution at different parts of the time-frequency plane. Instead of using a single analysis window like the STFT, the DWPT uses different windows at different frequency subbands. In this case, the subband window duration corresponds to the chosen subband frequency resolution and subband filter impulse response length. In particular, the wavelet transform uses short windows at high frequencies and long windows at low frequencies.

3. CASCADE LATTICE STRUCTURES AND FILTER BANKS

It is well known that orthogonal wavelet packet transforms can be designed by hierarchical association of perfect reconstruction paraunitary filter banks (tree structured filter bancs) [12]. Filter banks, however, were studied from the viewpoint of subband decomposition, which masks the time decomposition properties. These properties should be transparent in frame synchronized wavelet packet transform applications.

The main characteristic of paraunitary filter bank lattice structure is that it is composed of a cascade of orthogonal operators, delays, downsamplers for analysis filters and upsamplers for synthesis filters. The lattice structure implements, among others, two-channel maximally-downsampled paraunitary FIR filter banks of even length L, having the perfect reconstruction property, e.g. as in Fig. 1 ($L = 4$) where a criss-cross or butterfly operation is as presented in Fig. 2.

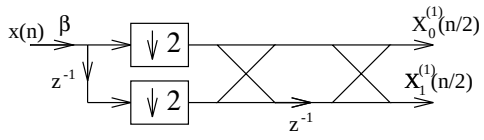

Figure 1: Two-channel maximally-downsampled paraunitary filter bank.

It also has a hierarchical property, i.e., higher order paraunitary filter banks can be obtained from those of lower order by adding more lattice sections. Another important property of the lattice structure is that by changing the rotation parameters γ_i we can generate all paraunitary filter banks. They depend on the desirable characteristics of the prototype lowpass filter. All these properties make them particularly important with reference to orthogonal wavelet packets. The rotation parameters γ_i can be directly computed by an optimization procedure in order to obtain a variety of orthogonal prototype filters with balanced regularity, frequency selectivity, number of orthogonal operators, and phase. Such a unified filter design procedure has been proposed in [11]. It can be employed to compute different types of filters such as Daubechies filters, binomial filters, Malvar MLT and ELT filters, those developed for orthogonal wavelet transforms, and other paraunitary filter sets. Thus, if we constructed the tree structured (binary or M-ary) filter bank using lattices, we could generate all orthogonal wavelet packet bases by manipulating the lattice parameters. To keep the illustration simple an example of four-channel ($N = 4$) tree-structured filter bank was chosen as in Fig. 3 ($L = 4$). As each level of the tree decomposition is calculated there is a decrease in temporal resolution and a corresponding increase in frequency resolution in each stage of the p-stage filter bank ($N = 2^p, p = 1, 2, ..., \log_2 N$).

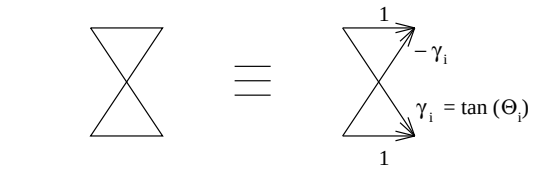

Figure 2: Butterfly operation.

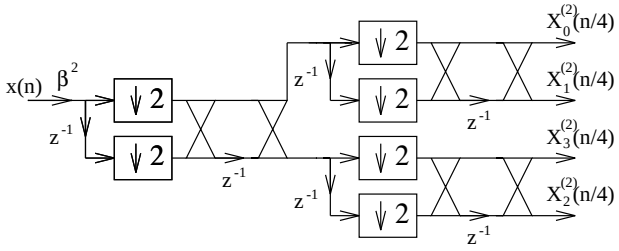

Figure 3: Four-channel tree-structured filter bank.

4. FAST OVERLAPPED BLOCK TRANSFORM ALGORITHMS

The two-channel cascade lattice can be re-arranged in an overlapped block transform lattice as depicted in Fig. 4 ($L = 4$) [6].

In this polyphase structure all downsamplers and delays can be moved to the left side, and the rotation parameters γ_i are the same as for the cascade structure. In a similar way, we can build a polyphase structure for the filter bank of Fig. 3.

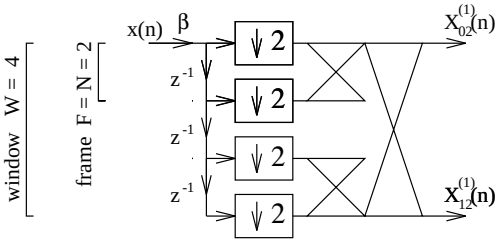

Figure 4: Overlapped block transform lattice for $N = 2$.

The result appears in Fig. 5.

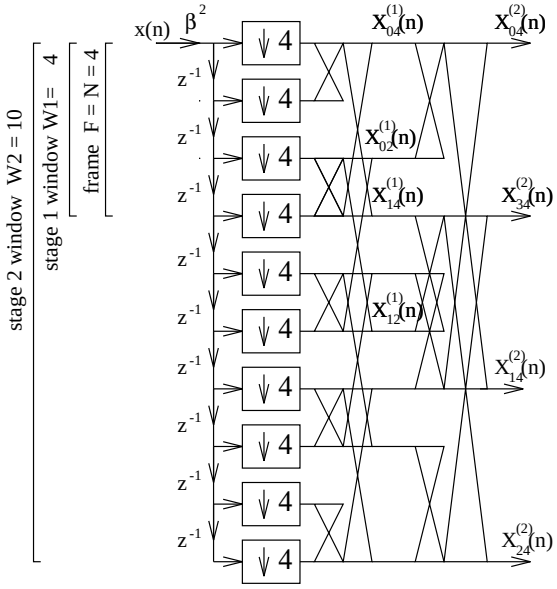

Figure 5: Overlapped block transform lattice for $N = 4$.

The polyphase lattice block structure characterized by a memoryless polyphase matrix represent an elegant and efficient solution to the implementation of wavelet packet transforms [6]. This approach leads directly to fast overlapped block transform algorithms for orthogonal short-time DWPTs which are very similar to the fast short-time Fourier transform (STFT) algorithms based on in-place butterfly operations. In the case of our algorithms, butterfly operations are characterized by lattice rotation and delay parameters. The direct and inverse transform algorithms possess the perfect reconstruction property. The complete lattice block structure provides all possible multiresolution time-spectral coefficients, since the DWPTs based on the same prototype filter correspond to its sub-structures. In this case, starting from a complete uniform tree or a polyphase block structure a sub-tree or a polyphase block sub-structure is selected, that matches at best the signal time-spectral characteristics following a given criterion. Such sub-structures allow the partitioning of the time-spectral domain into nonuniform tiles in connection with the time-spectral contents of the signal to be analyzed. The following four examples in Fig. 6 show all possible orthogonal decompositions of the time-frequency plane using the structures from Figs. 3 and 5. The corresponding tree structures are presented in Fig. 7.

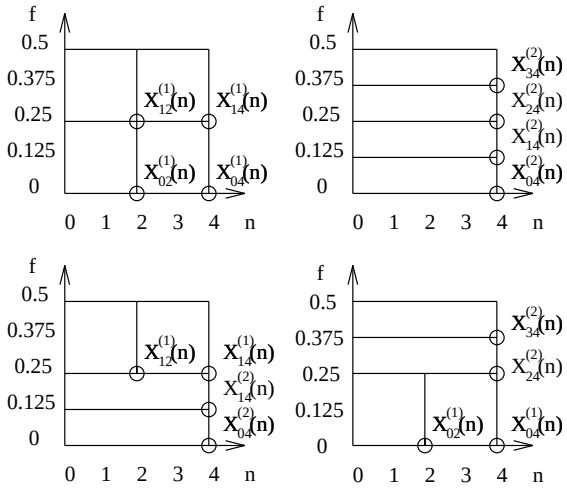

Figure 6: Orthogonal decompositions of time-frequency plane for $N = 4$.

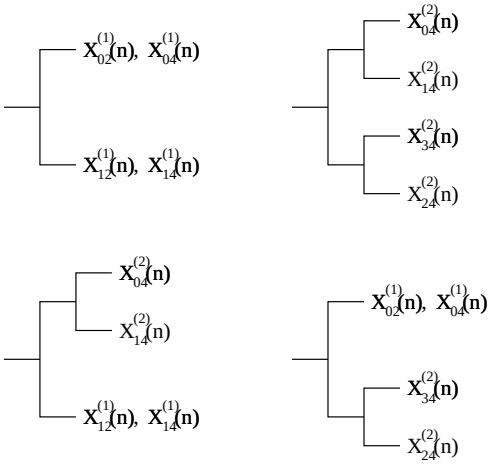

Figure 7: Tree structures for $N = 4$.

The overlapped block lattice structure provides also all types of easily programmable solutions from single-rate (linear time-invariant) systems to maximally downsampled systems, through all intermediary multirate systems with all admissible integer sampling rate alterations [2].

In all these cases, only the non-overlapping factor (lattice block shift), related to the downsampling factor and frame advance, is to be changed. The frame duration of the short-time DWPT depends on the lattice block shift and the choice of the DWPT. Such lattice block structures are computationally efficient since, to calculate each coming block of multiresolution time-spectral coefficients, it is only neces-

sary to realize non-overlapping lattices which do not belong to the previous block (Fig. 8).

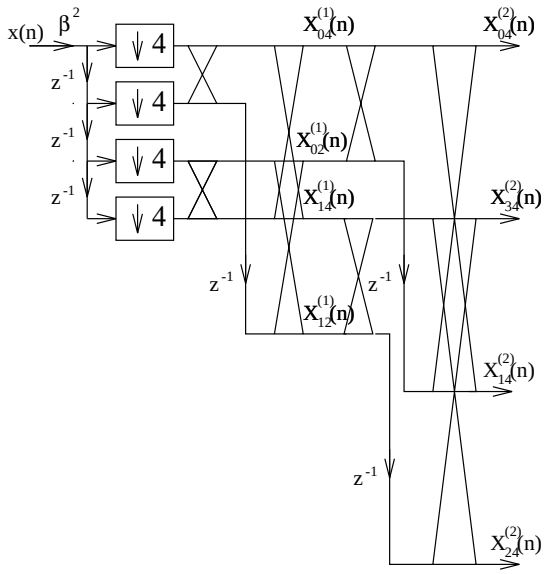

Figure 8: Frame-synchronized lattice structure.

5. CONCLUSIONS

Generalization of this new frame synchronous transparent formulation allows for an uncountable number of short-time DWPTs to be generated by choosing a type of criss-cross lattice operator, a lattice block structure, a subsampling factor (lattice block shift), and the rotation parameters of the lattices. It also provides a simple procedure for designing time-varying frame-synchronous wavelet packet transforms [7, 5]. From the presented formulation we see that analysis structures based on the developed principle can provide an extremely effective way to extract the multiresolution time-spectral characteristics of speech signals for frame synchronous processing systems. It is also worth mentioning that the analysis/synthesis scheme presented in this paper can be applied to design other efficient multiresolution speech processing systems such as perceptually tuned speech coders [3] and speech enhancement systems for robust speech recognition in noise based on the principle presented in [8]. The fast wavelet packet transform algorithms proposed in this paper can be directly used in the framework of HMM or hybrid HMM/Artificial Neural Network (ANN) systems for speech recognition where the idea is to split the whole frequency band (represented in terms of critical bands) into a few subbands on which different recognizers are independently applied and then recombined at a certain speech unit level to yield global scores and a global recognition decision [1].

6. REFERENCES

1. H. et al. Bourlard. Towards subband-based speech recognition. In *Proc. of the 8th European Signal Processing Conf. (EUSIPCO-96)*, Trieste, 1996. to be published.
2. B. Carnero and A. Drygajlo. Fast short-time orthogonal wavelet packet transform algorithms. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'95)*, pages 1161–11164, Detroit, 1995.
3. B. Carnero and A. Drygajlo. Perceptual coding of speech using a fast wavelet packet algorithm. In *Proc. of the 8th European Signal Processing Conf. (EUSIPCO-96)*, Trieste, 1996. to be published.
4. A. Drygajlo. Fast orthogonal transform algorithms for multiresolution time-sequency signal decomposition and processing. In *SPIE Conf. Mathematical Imaging: Wavelet Applications in Signal and Image Processing*, volume 2034, pages 349–358, San Diego, 1993.
5. A. Drygajlo. Overlapped block structure for fast time-varying orthogonal wavelet packet transform algorithms. In *Proc. of the 7th European Signal Processing Conf. (EUSIPCO-94)*, pages 668–671, Edinburgh, 1996.
6. A. Drygajlo and B. Carnero. Lattice structures for orthogonal wavelet packet implementations. In *Proc. of the European Conf. on Circuit Theory and Design (ECCTD'93)*, pages 1379–1384, Davos, 1993.
7. A. Drygajlo and N. Thevoz. Multiresolution speech analysis using fast time-varying orthogonal wavelet packet transform algorithms. In *Proc. 4th European Conf. on Speech Communication and Technology (EUROSPEECH'95)*, pages 255–258, Madrid, 1995.
8. A. Drygajlo, N. Virag, and G. Cosendai. Robust speech recognition in noise using speech enhancement based on masking properties of the auditory system and adaptive hmm. In *Proc. 4th European Conf. on Speech Communication and Technology (EUROSPEECH'95)*, pages 473–476, Madrid, 1995.
9. R.F Favero and R.W. King. Wavelet parameterization for speech recognition. *Int. Conf. on Signal Processing Applications and Technology*, 36:1444–1449, 1988.
10. J.W. Picone. Signal modeling techniques in speech recognition. *Proc. IEEE*, 81:1215–1247, 1993.
11. O. Rioul and P. Duhamel. A remez exchange algorithm for orthonormal wavelets. *IEEE Trans. Circuits and Syst.-II*, 41:550–560, 1994.
12. A.K. Soman and P.P. Vaidyanathan. On orthonormal wavelets and paraunitary filter banks. *IEEE Trans. Signal Processing*, 41:1170–1183, 1993.
13. M.V. Wickerhauser. *Adapted Wavelet Analysis from Theory to Software*. A K Peters, Wellesley, 1994.