A LOW DELAY, VARIABLE RESOLUTION, PERFECT RECONSTRUCTION SPECTRAL ANALYSIS-SYNTHESIS SYSTEM FOR SPEECH ENHANCEMENT

Dirk Mauler and Rainer Martin

Institute of Communication Acoustics (IKA), Ruhr-Universität Bochum, 44780 Bochum, Germany
email: {Dirk.Mauler, Rainer.Martin}@rub.de

ABSTRACT

The choice of the window function and window length in short time analysis-synthesis (AS) systems based on the discrete Fourier transform (DFT) has to balance conflicting requirements: Long windows provide high spectral resolution while short windows allow for high temporal resolution. Furthermore, for many applications a low algorithmic delay is desirable. Therefore, long standard windows such as the Hann or Hamming windows cannot be used. In this contribution we propose a novel AS system which achieves perfect reconstruction (PR) and a low delay by using a variable length analysis window and a relatively short synthesis window. The variable length analysis windows allow a spectral analysis that is adapted to the signals span of stationarity. The AS windows are designed such that they can be switched at any time instant without violating PR. We show that the spectral representation of typical speech data is improved as compared to AS systems with standard windows.

1. INTRODUCTION

Motivated by computational efficiency and good decorrelation properties many digital signal algorithms like noise reduction operate in the frequency domain. Frequently, the fast Fourier transform (FFT) is used for the efficient transformation from the time to the frequency domain. In the short-time discrete Fourier transform (DFT) technique the time domain signal is partitioned in possibly overlapping frames of data. Each frame is transformed to the frequency domain where the frequency domain algorithm possibly modifies the data. The time domain signal is then obtained via the inverse transformation and appropriate reassembling of the data [5].
Two conflicting requirements rule the choice of a proper frame length of a DFT based analysis-synthesis system: On the one hand, a long window is required to achieve a frequency resolution that satisfies the requirements of the frequency domain algorithm. On the other hand, to obtain a useful spectral representation the window length must not exceed the span of stationarity of the time domain signal. Additionally, in an overlap-add synthesis system the latency of the system is a function of the window length. Therefore, for low delay applications like hearing aids, long synthesis window lengths are not admissible.

Researchers have published ideas that aim at improving either a better exploitation of stationary signal sections [7], at increasing the spectral resolution [4] or at reducing the delay [3].

In this paper we present windows for a perfect reconstruction DFT based AS system. The main idea is to provide a switchable set of pairs of AS windows that allows to adapt the window length to the span of stationarity of the data and to use a short synthesis window to achieve short delay. By this, a variable spectral and temporal resolution can be achieved at a constant low delay.

Starting from the DFT based spectral AS system in Section 2 we present the AS window set in general (Section 3) and give an example for an AS window set in Section 4. The window set is evaluated and compared with standard windows in Section 5. Conclusions summarize the most important results.

2. DFT BASED ANALYSIS-SYNTHESIS SYSTEM

We consider a block-based analysis-system with K frequency bins and a frame advance of R samples. If we restrict the system to uniform frequency resolution the DFT can be used and efficiently implemented by means of an FFT algorithm. Then, the spectral coefficients, $X_k(m)$, of the sampled time domain data $x(n)$ are obtained as

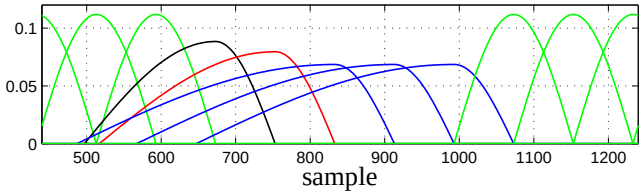
$$X_k(m) = \sum_{n=0}^{K-1} x(mR+n)\, h(n)\, e^{-2\pi kn/K}, \qquad (1)$$

where K is the DFT length, $h(n)$ denotes an analysis window, m is the subsampled (frame) index and $k = 0..K-1$ is the discrete frequency bin index.

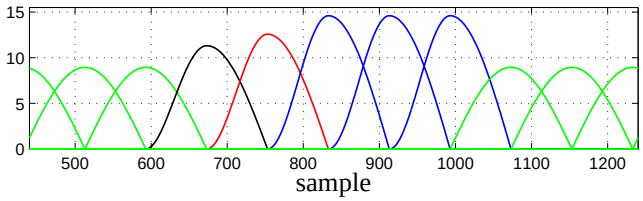
Given the spectral coefficients, $X_k(m)$, that may be weighted with a spectral gain, $G_k(m)$, the signal synthesis is performed via IDFT, multiplication with a synthesis window, $f(n)$, and a subsequent overlap-add operation [8].

3. A SWITCHABLE ANALYSIS-SYNTHESIS WINDOW SET

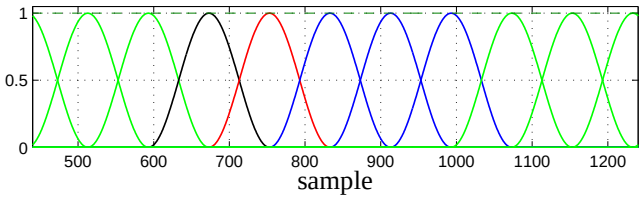
We aim to provide several pairs of analysis and synthesis windows that show the same low latency but differ in their spectral and temporal resolution. The DFT length shall always be K, requiring zero padding when shorter windows are to be used. If no spectral modification is performed we want the system to reproduce the (delayed) input signal, thus achieving perfect reconstruction (PR). Furthermore, arbitrary switching between AS window pairs shall be possible without disturbing the PR property of the system. This idea is illustrated in Figure 1. A sequence of overlapping analysis windows is shown in Figure 1(a) with the corresponding synthesis windows in Figure 1(b) such that overlapping Hann windows result that add up to identity (Figure 1(c)) and therefore form a PR system. Note that the analysis window amplitudes are normalized on the window energy. The corresponding synthesis window amplitudes are chosen to achieve PR. As also the shape of the synthesis window depends on



(a) Analysis windows.



(b) Synthesis windows.



(c) Product of analysis-synthesis windows and overlap-add result (dashed).

Figure 1: Example for switching between AS window pairs. The frame advance in this example is $R = 80$ samples. Perfect reconstruction is achieved at all times.

the analysis window, analysis and synthesis windows used in frame m form a unique pair.

3.1 Perfect reconstruction (PR)

To simplify the subsequent discussion we describe each window as a concatenation of four segments

$$h^i(n) = \{h_1^i(n), h_2^i(n), h_3^i(n), h_4^i(n)\} \tag{2}$$

where i denotes the index of an AS window pair. We introduce M, the length of segments $h_3^i(n)$ and $h_4^i(n)$, and $d \leq K - 2M$, the length of segment $h_1^i(n)$. Segment $h_2^i(n)$ is of length $K - 2M - d$ samples. The synthesis windows $f^i(n)$ shall be partitioned in the same fashion. In general, these windows will not be symmetric.
PR is achieved, for example, in a system with frame advance $R = M$ if the effective window after analysis and synthesis, $h(n)f(n)$, is a Hann window of length $2M$. We therefore require the samplewise product of every AS window pair, i, to be a periodic Hann window of length $2M$

$$\begin{aligned} & h^i(n) f^i(n) \\ = & \begin{cases} 0, & 0 \leq n < K - 2M \\ Hann_{2M}(n - K + 2M), & K - 2M \leq n < K, \end{cases} \end{aligned} \tag{3}$$

where the periodic Hann window of length L is defined as

$$Hann_L(n) = 0.5\left(1 - \cos(2\pi\frac{n}{L})\right), n = 0, ..., L-1. \tag{4}$$

If this condition is fulfilled and without spectral modifications the AS window pairs can be switched arbitrarily without violating the PR property. Furthermore, since the DFT length is always equal to K, the same number of frequency bins is generated and no explicit interpolation is required when windows are switched.
PR is still preserved for an integer frame advance $R = M/2^p, p = 0, 1, 2, ...$ (requiring appropriate scaling of the synthesis window). The advantage of a smaller frame advance is the smaller decimation and hence the higher aliasing and imaging attenuation of the analysis and the synthesis filters [8].
Other PR AS window combinations exist with possible different ratio K/M. In this paper we restrict ourselves to the Hann window because of its good reconstruction properties.

3.2 Reconstruction after spectral modification

If the spectral representation is multiplicatively modified with a spectral gain, $G_k(m)$, spectral aliasing and residual imaging components of the AS system may no longer cancel out each other as in the unmodified case. Time domain aliasing occurs if the convolutional product of windowed input sequence and the impulse response of the spectral gain function is longer than the DFT length K. If a symmetric analysis window, like the square-root Hann, is used and the impulse response of the spectral gain function has a linear phase and has its maximum concentrated around lag 0, the convolution of both shows temporal aliasing at the beginning and at the end of the frame. A tapered synthesis window (square-root Hann) can alleviate these effects to a large extend so that aliasing is hardly or not perceived [6].

If, as in this proposal, an asymmetric synthesis window is used, that keeps the $2M$ most recent samples of the convolution product, care has to be taken, that cyclic aliasing in those samples is prevent or is at least at a low level. The amount of aliasing is basically a function of the effective length of the gain filter impulse response. In many applications, like noise reduction, the major energy of the total impulse response is concentrated within the first few taps. We therefore make the simplifying assumption that the impulse response is non zero only in the first d taps. Then, in order to prevent aliasing in the $2M$ most recent samples of the convolution product, section $h_1^i(n)$ of the analysis window has to contain zeros. Note however, that even for less than d zeros aliasing might still not be perceivable due to the tapered shape of the synthesis window.

In both window approaches, symmetric and asymmetric, temporal aliasing can be eliminated by sufficient zero padding, i.e by increasing the window and transformation length K. This, however, would considerably increase the computational complexity.

3.3 System delay

The system latency introduced by the proposed block-based processing overlap-add structure is given by

$$\tau = \tau_{ola} + \tau_{group} + \tau_{process} \tag{5}$$

where $\tau_{ola} = 2M/f_s$ is the effective length of the windows to be overlapped and added, f_s is the sampling frequency, τ_{group} is the group delay introduced by the spectral gain, and

$\tau_{process}$ is the time that is necessary for processing the data of a single frame. In a block-based system, where the IO buffers are synchronously accessed, the processing delay $\tau_{process}$ is proportional to an integer multiple of the frame advance, R. Assuming that the processing can be accomplished during a single block capture the total latency for the proposed window set amounts to $\tau = (2M+R)/f_s + \tau_{group}$.

3.4 Computational complexity

The computational complexity can be significantly reduced by using a pruned IFFT for synthesis and a pruned FFT for analysis whenever synthesis or analysis windows are used whose effective lengths are shorter than the DFT length K. If the DFT length is a power of two, i.e. $K = 2^q$, and if the effective length of a synthesis or an analysis window is also a power of two, i.e. $2M = 2^l$, then a simple modification to the radix-2 decimation in-time algorithm allows a time-saving of approximately $(q-l)/q$ with respect to the unpruned case [2].

4. WINDOW SET EXAMPLE

In the following, we give an example for an AS PR window set that fulfils the constraint given in (3) and realizes a delay of $\tau_{ola} + \tau_{process} = 10$ ms at $f_s = 16$ kHz ($K = 512$, $M = 64$, $R = 32$). We describe here only two exemplary AS window pairs, $i = I$ and $i = II$, the first realizing a high temporal resolution and the second a high frequency resolution at the same low latency.
A square-root Hann window with zero padding is obtained if we set

$$\begin{aligned} h_1^I(n) &= f_1^I(n) = 0, \\ h_2^I(n) &= f_2^I(n) = 0, \\ h_3^I(n) &= f_3^I(n) = \sqrt{Hann_{2M}(n)}, \\ h_4^I(n) &= f_4^I(n) = \sqrt{Hann_{2M}(n+M)}. \end{aligned}$$

The analysis and the synthesis window are plotted in the upper half of Figure 2. Since zero padding corresponds to an interpolation of the spectrum the frequency resolution of analysis window $h^I(n)$ is basically $f_s/(2M) = 125$ Hz for a sampling frequency of $f_s = 16$ kHz. The advantage of the spectral interpolation is that both, the transform based on the analysis window $h^I(n)$ and the transform based on the analysis window $h^{II}(n)$ (see below) result in the same number of spectral coefficients, K, so that spectral parameters can be easily updated even if the window set is switched.

A second window pair with a higher spectral resolution of $f_s/(K-d) \approx 35.7$ Hz ($d = 64$) is defined by the analysis window

$$\begin{aligned} h_1^{II}(n) &= 0, \\ h_2^{II}(n) &= \sqrt{Hann_{2(K-M-d)}(n)}, \\ h_3^{II}(n) &= \sqrt{Hann_{2(K-M-d)}(n+K-2M-d)}, \\ h_4^{II}(n) &= \sqrt{Hann_{2M}(n+M)} \end{aligned}$$

and the respective synthesis window

$$\begin{aligned} f_1^{II}(n) &= f_2^{II}(n) = 0, \\ f_3^{II}(n) &= \frac{Hann_{2M}(n)}{\sqrt{Hann_{2(K-M-d)}(n+K-2M-d)}}, \\ f_4^{II}(n) &= \sqrt{Hann_{2M}(n+M)}. \end{aligned}$$

Window pair II is plotted in the lower half of Figure 2. Note that the amplitudes of the analysis windows are to be normalized to the square-root of the total window power for reasons of power preservation in the transformed domain [1]. As a consequence, spectral parameters obtained before and after a switching of the window pairs are interchangeable and can be updated as if they had been originated from an analysis with one and the same window. For PR, the synthesis windows are to be scaled by the inverse of the normalization factor.

The phase response of the asymmetric analysis window $h^{II}(n)$ is non-linear. If linear phase is required, e.g. a symmetric flat-top window can be used instead.

5. DISCUSSION

In this section we compare the window proposal from Section 4 with square-root Hann windows for the analysis and synthesis with window and transformation length 512 or 128 samples. The former is comparable with the long asymmetric window in terms of spectral resolution, the latter is comparable with the proposed window set in terms of latency.

5.1 Comparison at same spectral resolution

In terms of spectral resolution a square-root Hann window of length 512 is comparable with the analysis window $h^{II}(n)$. However, because of a length 512 synthesis window the delay $\tau_{ola} + \tau_{process}$ introduced with the symmetric square-root Hann windows is 34 ms and is therefore 3.4 times larger than with the window pair II.

Another effect can be observed in Figure 3 where the spectral mean powers of the undisturbed German word "sollte" are plotted after analysis with different analysis windows. This word contains unvoiced, voiced and transition speech sounds. We notice that the mean power obtained with the square-root Hann analysis (dashed red line) is lagging behind the mean powers of the windows $h^I(n)$ and $h^{II}(n)$ (thin blue and thick green lines). Especially at stationarity boundaries, as e.g. at 0.1 s or 0.34 s the power in the original time domain signal is less well reflected in the mean power of the square-root Hann window as compared to the proposed windows. This blurring is a consequence of the fact that the long square-root Hann window spans across instationary data. Interestingly, the window $h^{II}(n)$, despite of being nearly as long as the square-root Hann window, shows a steep response to the speech onsets at 0.1 s or 0.34 s. This behavior is due to the asymmetry of the window which emphasizes the most recent samples.

A short window, like the proposed $h^I(n)$, should be used during transitions. The advantage is a better temporal resolution which can be observed e.g. at 0.09 s. In stationary

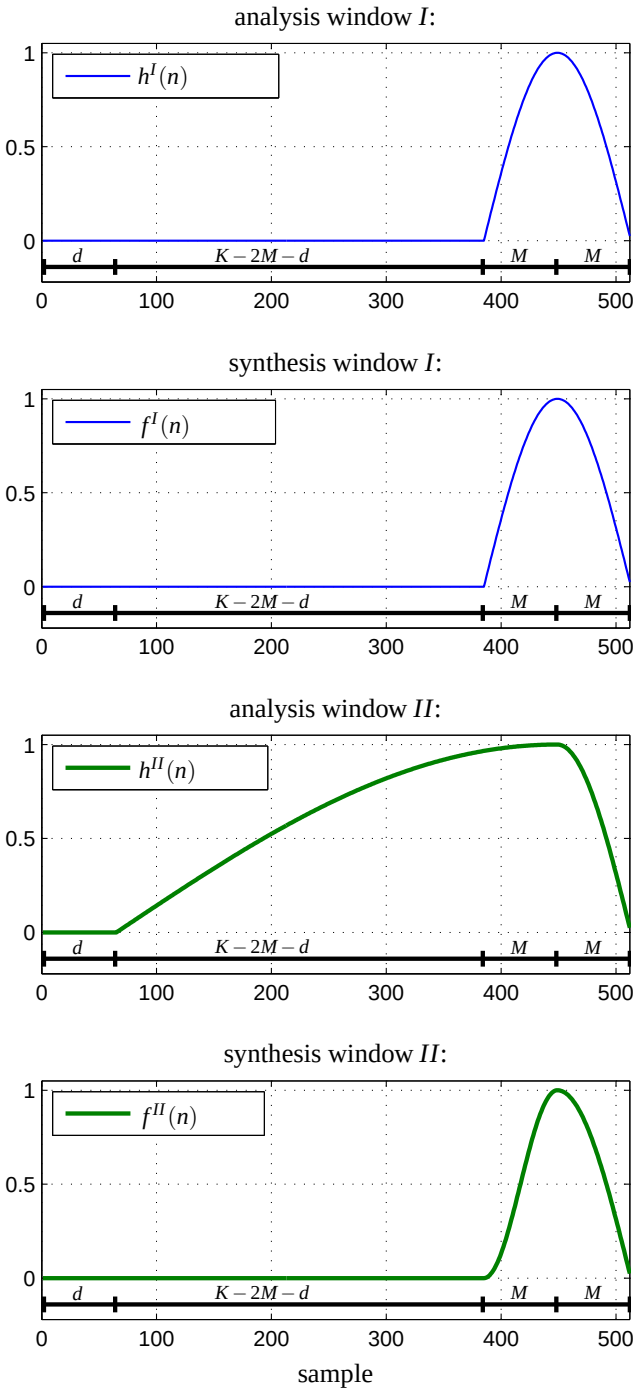


Figure 2: Example of analysis and synthesis window pairs with high temporal (window pair I) and high spectral resolution (window pair II) for window length $K = 512$, $M = 64$, and $d = 64$ samples.

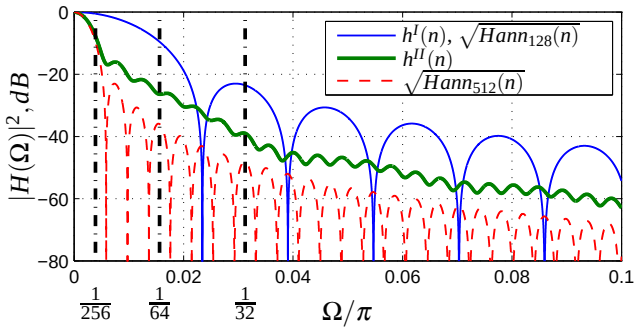


Figure 4: Frequency response of analysis windows ($\Omega = 2\pi f/f_s$, $H(\Omega) = FFT\{h(n)\}$).

sections (e.g. 0.12 s - 0.26 s) a longer window has the advantage of yielding a higher frequency resolution and being less sensitive to the temporal pitch structure.

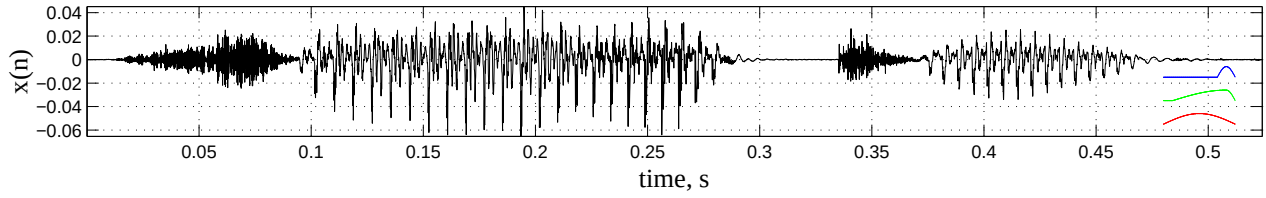
5.2 Comparison at same delay

A square-root Hann window of length 128 *without* zero padding for analysis and synthesis produces the same delay as the window sets I and II. However, using the long window, $h^{II}(n)$, the spectral resolution is considerably larger than for a square-root Hann analysis window of length 128. The price to pay for the improved spectral resolution is a higher computational complexity.

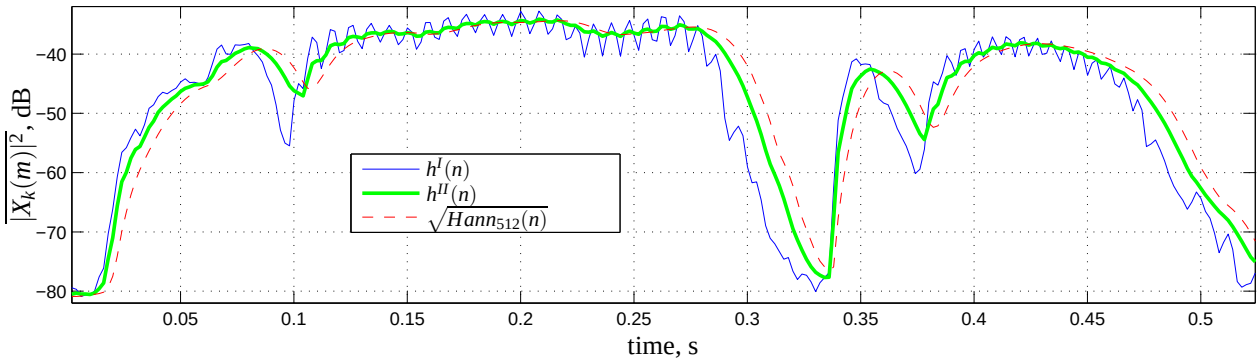
In Figure 4 we compare the frequency responses of the long asymmetric window (thick green line) with those of the square-root Hann window of length 128 which is identical with the frequency response of the zero padded square-root Hann window, $h_I(n)$ (thin blue line). For convenience, the frequency response of the square-root Hann window of length 512 is also plotted (dashed red line). We observe that comparing the two windows of equal delay the long asymmetric window has a much stronger side lobe attenuation than the square-root Hann window of length 128. Furthermore, the main lobe is considerably narrower than for the square-root Hann window of length 128. The asymmetric window nearly achieves the same narrow main lobe width of the long square-root Hann window but yields roughly 10 - 15 dB less side lobe attenuation than the 512-tap square-root Hann window. However, the larger attenuation of the long square-root Hann window is payed with a delay 3.4 times larger than the delay of the asymmetric window.

The dash-dotted black lines at $\Omega/\pi = 1/256$ and $\Omega/\pi = 1/64$ mark half of the channel bandwidth for a 512 point transform and for a 128 point transform, respectively, and therefore indicate the spectral position of the channel adjacent to the channel at $\Omega = 0$. While the 128-tap square-root Hann window attenuates channel crosstalk by 9.5 dB, the long asymmetric window attenuates crosstalk by 7.7 dB.

With a subsampling factor $R = 32$, the spectral components at $\Omega/\pi > 1/32$ produce spectral aliasing if they are not sufficiently attenuated by the analysis window. Thus, for the same algorithmic delay, Figure 4 shows that the asymmetric window produces significantly less aliasing than the 128-tap square-root Hann window.



(a) Time domain signal.



(b) Spectral mean power (125 Hz - 7500 Hz).

Figure 3: Time domain samples (a) and spectral mean power (b) using different analysis windows for the example of the undisturbed german word “sollte”. In the lower right corner in (a) the window lengths of the analysis windows used under (b) are plotted to scale. The frame advance is $R = 32$ samples (2 ms).

5.3 Computational complexity

The DFT length $K = 512$, as well as the effective length of the short windows, $h^I(n)$, $f^I(n)$, and $f^{II}(n)$, are powers of two ($q = 9$, $l = 7$). As argued in Section 3.4, by using a pruned FFT and a pruned IFFT, approximately $(q-l)/q =$ 75% of computation time with respect to the unpruned DFT can be saved [2].

6. CONCLUSIONS

We presented a short-time DFT based analysis-synthesis system with a set of perfect reconstruction analysis-synthesis windows. Each pair of analysis-synthesis windows produce the same latency and are interchangeable during operation without violating the PR property. The main motivation is to offer a greater flexibility in the window choice and therefore better exploitation of the advantages of a long analysis window, which is high spectral resolution and insensitivity to temporal pitch structures, and the advantages of a short analysis window, which is the applicability to short-time stationary data, like plosives or stops and a good temporal resolution. The increased flexibility of the proposed system is payed for with an increased computational complexity which, in turn, can be reduced by using pruned FFTs. Future work will focus on strategies for the automatic selection of window pairs for improved time-frequency analysis of speech signals.

Acknowledgment

This work was in part supported by grants from the European Union FP6, Project 004171 HEARCOM. We thank G. Enzner for stimulating discussions on this subject.

REFERENCES

[1] Daniel W. Griffin and Jae S. Lim. Signal Estimation from Modified Short-Time Fourier Transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-32(2):236–243, April 1984.

[2] David P. Skinner. Pruning the Decimation In-Time FFT Algorithm. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 24:193–194, April 1976.

[3] Heinrich W. Löllmann and Peter Vary. A Warped Low Delay Filter for Speech Enhancement. In *Proc. Intl. Workshop Acoustic Echo and Noise Control (IWAENC)*, September 2006.

[4] Ismo Kauppinen and Kari Roth. Improved Noise Reduction in Audio Signals Using Spectral Resolution Enhancement With Time-Domain Signal Extrapolation. *IEEE Trans. on Speech and Audio Processing*, 13(6):1210–1216, November 2005.

[5] Jont B. Allen, Lawrence R. Rabiner. A Unified Approach to Short-Time Fourier Analysis and Synthesis. *Proceedings of the IEEE*, 65(11):1558–1564, November 1977.

[6] Rainer Martin and Richard V. Cox. New Speech Enhancement Techniques for Low Bit Rate Speech Coding. In *IEEE Workshop on Speech Coding*, pages 165–167, Juni 1999.

[7] Richard C. Hendriks, Richard Heusdens, and Jesper Jensen. Adaptive Time Segmentation for Improved Speech Enhancement. *IEEE Trans. on Audio, Speech, and Language Processing*, 14(6):2064–2074, November 2006.

[8] Ronald E. Crochiere. *Multirate Digital Signal Processing*. Prentice-Hall Inc., 1^{st} edition, 1983.