A Framework for Quality Versus Efficiency Tradeoffs in STFT Analysis

S. Hamid Nawab and Erkan Dorken

***Abstract*—A framework is presented for increasing the computational efficiency of STFT analysis by sacrificing the quality of each signal frame's DFT in terms of SNR, frequency resolution, and frequency coverage. The resulting algorithms are dominated by a frame-adaptive vector summation process designed to ensure that the number of additions per frame does not exceed any desired limit.**

I. Introduction

High-speed evaluation of the short-time Fourier transform (STFT) [1] often involves fast algorithms for computing the discrete Fourier transform (DFT). A prominent feature of most of these algorithms is that their outputs are error-free DFT's of their inputs, except for the effect of finite-precision arithmetic. In contrast, the Poorman's transform [2] achieves computational efficiency at the expense of error in the DFT samples, even when infinite-precision arithmetic is used. The algorithm definition in this case requires the use of highly approximate values for the complex exponential coefficients multiplying the signal values in the DFT formula. Our framework for DFT-based evaluation of the STFT also yields algorithms that produce error in the DFT points in order to achieve computational efficiency. However, our approach involves the application of coarse quantization to signal values instead of multiplier coefficients. Most importantly, this framework leads to the formulation of techniques for systematically reducing output quality in order to ensure that the number of arithmetic operations per DFT does not exceed a specified limit.

Although efficient computation of the DFT plays a central role in our framework for STFT analysis, it is important to observe that certain aspects of the framework depend critically on the fact that the DFT computations are embedded within the STFT. For example, in some cases, it is possible to exploit the overlap between adjacent signal frames to gain additional computational efficiency. Another important consequence of the STFT context is that our framework can be used to design techniques for *frame-adaptive* reduction of output quality in order to keep the amount of computation per DFT below a specified threshold.

II. Basic Theory

Let us consider the discrete STFT based on a causal analysis window $w(n)$ whose length is N_w. If the temporal decimation interval is denoted by the integer L, the mth N-point DFT ($N \geq N_w$) in the discrete STFT of a signal $x(n)$ may be defined as

$$X_{mL}(k) = \sum_{n=mL-N_w+1}^{mL-N_w+N} x_{mL}(n)e^{-j(2\pi/N)kn}; \quad k = 0, 1, \cdots, N-1 \tag{1}$$

where $x_{mL}(n) = x(n)w(mL-n)$ is the mth signal frame.

Manuscript received October 23, 1993; revised October 5, 1994. This work was supported by the Rome Laboratories of the Air Force Systems Command under Contract F30602-91-C-0038 and by the Department of the Navy, Office of Naval Research, Contract N00014-93-1-0686 as part of the Advanced Research Project Agency's RASSP program. The associate editor coordinating the review of this paper and approving it for publication was Prof. Henrik V. Sorensen.

The authors are with the Electrical, Computer, and Systems Engineering Department, Boston University, Boston, MA 02215 USA.

IEEE Log Number 9409782.

Our approach to the evaluation of the STFT is based on first performing a backward differencing operation on each signal frame to obtain

$$g_{mL}(n) = \begin{cases} x_{mL}(n) - x_{mL}(n+N-1) \\ \quad \text{for } n = mL - N_w + 1 \\ x_{mL}(n) - x_{mL}(n-1) \\ \quad \text{for } mL - N_w + 1 < n \leq mL - N_w + N \end{cases} \tag{2}$$

It can then be easily shown that

$$X_{mL}(k) = \sum_{n=mL-N_w+1}^{mL-N_w+N} g_{mL}(n)W_n(k); \quad k = 1, \cdots, N/2 \tag{3}$$

where

$$W_n(k) = \left(\frac{e^{-j(2\pi/N)kn}}{1 - e^{-j(2\pi/N)k}}\right). \tag{4}$$

The expression for $X_{mL}(k)$ in (3) has the form of a DFT whose multiplicative coefficients have been modified. Because of the backward differencing operation, this expression is not valid at $k = 0$. However, the dc component is not of practical interest in many applications. We have also excluded the values of k greater than $N/2$ under the assumption that the signals of interest are real.

Suppose the samples in the signal frames $x_{mL}(n)$ undergo a Q-level quantization. The quantized frames $x^Q_{mL}(n)$ can be used to obtain an "approximate" STFT:

$$X^Q_{mL}(k) = \sum_{n=mL-N_w+1}^{mL-N_w+N} g^Q_{mL}(n)W_n(k); \quad k = 1, \cdots, N/2 \tag{5}$$

where $g^Q_{mL}(n)$ is obtained by applying the backward differencing of (2) to $x^Q_{mL}(n)$. For brevity of presentation, we will assume that the quantization is based on a *rounding* technique and that Q is an odd integer. This type of quantization results in one zero-valued level $(Q-1)/2$ negative levels and $(Q-1)/2$ positive levels. It follows that each sample in a differenced frame $g^Q_{mL}(n)$ takes on one of $2Q-1$ possible values. The expression in (5) for $X^Q_{mL}(k)$ can be viewed in terms of a *vector-summation* operation performed among column vectors. Let $\boldsymbol{X}^Q_{mL}$ be an $(N/2)$-point column vector in which the kth element ($1 \leq k \leq N/2$) is $X^Q_{mL}(k)$. In addition, let $\boldsymbol{W}_n$ denote an $(N/2)$-point column vector whose kth element ($1 \leq k \leq N/2$) is the multiplicative factor $W_n(k)$. It should be noted that there are only N unique vectors of the form $\boldsymbol{W}_n$ because $W_n(k)$ is periodic in n with a period of N. Equation (5) can now be rewritten as

$$\boldsymbol{X}^Q_{mL} = \sum_{n=mL-N_w+1}^{mL-N_w+N} g^Q_{mL}(n)\boldsymbol{W}_n. \tag{6}$$

Noting that there are only $(2Q-2)$ distinct nonzero values that can be taken up by any sample of $g^Q_{mL}(n)$, the right side of (6) may be viewed as a vector summation of *prestored* scalar multiples each of $\boldsymbol{W}_n$. This summation process is illustrated in Fig. 1 for the eight-point DFT of a four-point frame $x^Q(n)$.

The quality of the approximate STFT may be measured in terms of its error with respect to the exact STFT. When the highest magnitude within a frame is A and the number of quantization levels is Q, we divide the amplitude range from $-A$ to A into Q uniformly spaced regions. The quantization levels are located at $(2An/Q)$ for $-(Q-1)/2 \leq n \leq (Q-1)/2$. If $x(t)$ is a zero-mean white

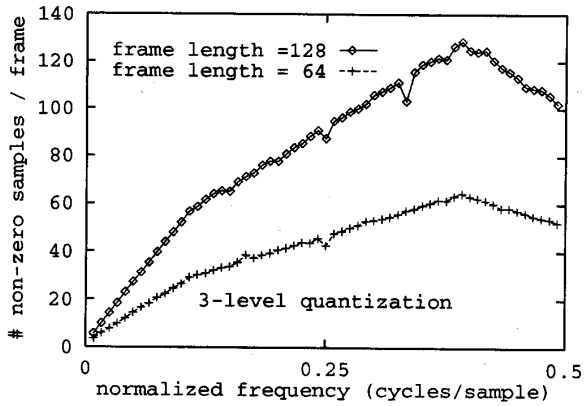

Fig. 1. Number of nonzero samples (N_v) obtained after three-level quantization and backward differencing of sinusoidal frames with different frequencies. The curve joining the points marked with ◇'s represents the results for frames of length 128 samples, whereas the curve joining the points marked with +'s depicts the results for 64-sample frames.

noise signal with a uniform distribution between $-A$ and A, the SNR for its quantized version can be shown to be $10 \log (Q^2)$. For practical implementations, it is generally desirable to keep Q very small since the storage requirement for the vectors of scaled multiplicative coefficients is $O(QN^2)$. In our experiments with the evaluation of the approximate STFT, we have used the values 3, 5, 7, and 9 for Q. They result in SNR's of 9, 14, 17, and 19 dB, respectively.

Keeping the number of quantization levels small is also advantageous for reducing the number, which is denoted by N_v, of vector additions needed for evaluating an approximate STFT frame. The value of N_v is equal to the number of nonzero samples in $g^Q_{mL}(n)$. When Q is small and $x_{mL}(n)$ is a slowly varying signal, $g^Q_{mL}(n)$ has significantly fewer nonzero samples than $x^Q_{mL}(n)$.

On the basis of an experimental study [3], we have obtained a quantitative model for estimating the number of vector additions (N_v) needed in the evaluation of $X^Q_{mL}(k)$, where the approximate STFT corresponds to a Q-level frame quantization. At a qualitative level, our model indicates that the number of vector additions is most sensitive to the highest energy regions of the frame spectrum. To examine the detailed model, let us consider a signal frame $x(n)$ obtained by windowing a linear combination of sinusoids. In particular, suppose that $x(n)$ is given by

$$x(n) = \sum_{p=1}^{M} A_p \sin(2\pi f_p n + \theta_p); \qquad 0 < n \leq N_w - 1 \quad (7)$$

Our model states that the number of vector additions needed for the mth frame in the evaluation of $X^Q_{mL}(k)$ may be approximated by the following expression:

$$N_v = \sum_{p=1}^{M} \frac{A_p^2}{\sum_{r=1}^{M} A_r^2} \alpha(f_p) \quad (8)$$

where $\alpha(f)$ is an experimentally determined function that specifies the *mean* value of N_v for frames obtained by windowing single sinusoids of normalized frequency f. The expression in (8) was obtained by first experimentally determining the function $\alpha(f)$ and then comparing it with the actual number of additions obtained for frames with arbitrarily chosen frequency content.

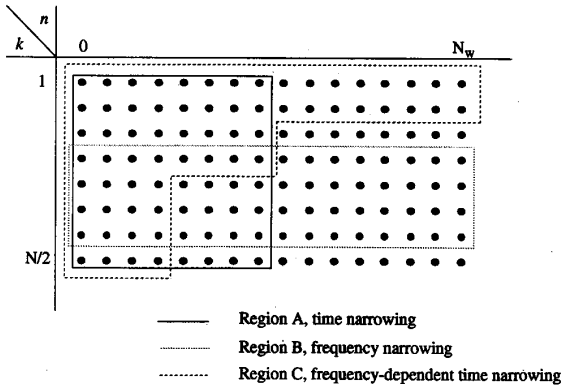

Fig. 2. Exclusion of vector elements from the vector-summation process. The columns of dots represent the $N/2$-element vectors $g^Q_{mL}(n)\boldsymbol{W}_n$, which have to be summed to obtain $X^Q_{mL}(k)$. Time-narrowing (loss of frequency resolution) takes place when the vector-summation process is restricted to region A. Restricting the vector-summation process to region B results in frequency narrowing and consequently loss of frequency coverage. The type of narrowing depicted by region C results in loss of frequency resolution but no loss of frequency coverage.

As an example of the experimental formulation of the function $\alpha(f)$, we considered the case where the frame length (N_w) is 128 and the number of quantization levels (Q) is 3. The argument of $\alpha(f)$ was considered for a range of 60 values: $0.5k/60$ for $1 \leq k \leq 60$. For each of these values of f, we generated 20 frames according to the formula

$$x(n) = \begin{cases} \sin(2\pi f n + \theta) & \text{for } 0 \leq n \leq N_w - 1 \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

where the value of θ was selected independently for each frame through the use of a random number generator. After performing three-level quantization and subsequent backward differencing on each frame, we calculated the number of nonzero samples N_v in the result. The mean of these 20 numbers is displayed as a function of frequency by the diamond (◇) marks in Fig. 2. We observe that this data can be segmented into three regions (for an explanation, see [3]), each of which may be modeled by a straight line with a different slope. The boundaries between these regions are approximately located at 0.1 and 0.4 cycles/sample. We calculated the standard deviation (averaged across frequency) of the number of nonzero samples per frame to be approximately 0.6, 4.7, and 3.8 nonzero samples for the low-, middle-, and high-frequency regions, respectively. The boundaries between these regions remain invariant at other frame lengths, as is illustrated by the experimental data plotted using + marks in Fig. 2 for the case of 64-point frames. The standard deviation of the data in each region is approximately half of that obtained for 128-point frames.

III. Frame-Adaptive Approaches

When the number of additions (as determined by the N_v measurement) required by the vector-summation in (6) exceeds the desired limit B, frame-adaptive approaches may be utilized to sacrifice frequency resolution, frequency coverage, or some combination of the two in proportion to the number of additions that have to be eliminated. Furthermore, the reduction of frequency coverage can be designed to be sensitive to the frequency content of the corresponding frame.

The loss in frequency resolution or frequency coverage results from the *exclusion* of a subset of vector elements from the vector-summation process. As an illustration, consider the situations depicted

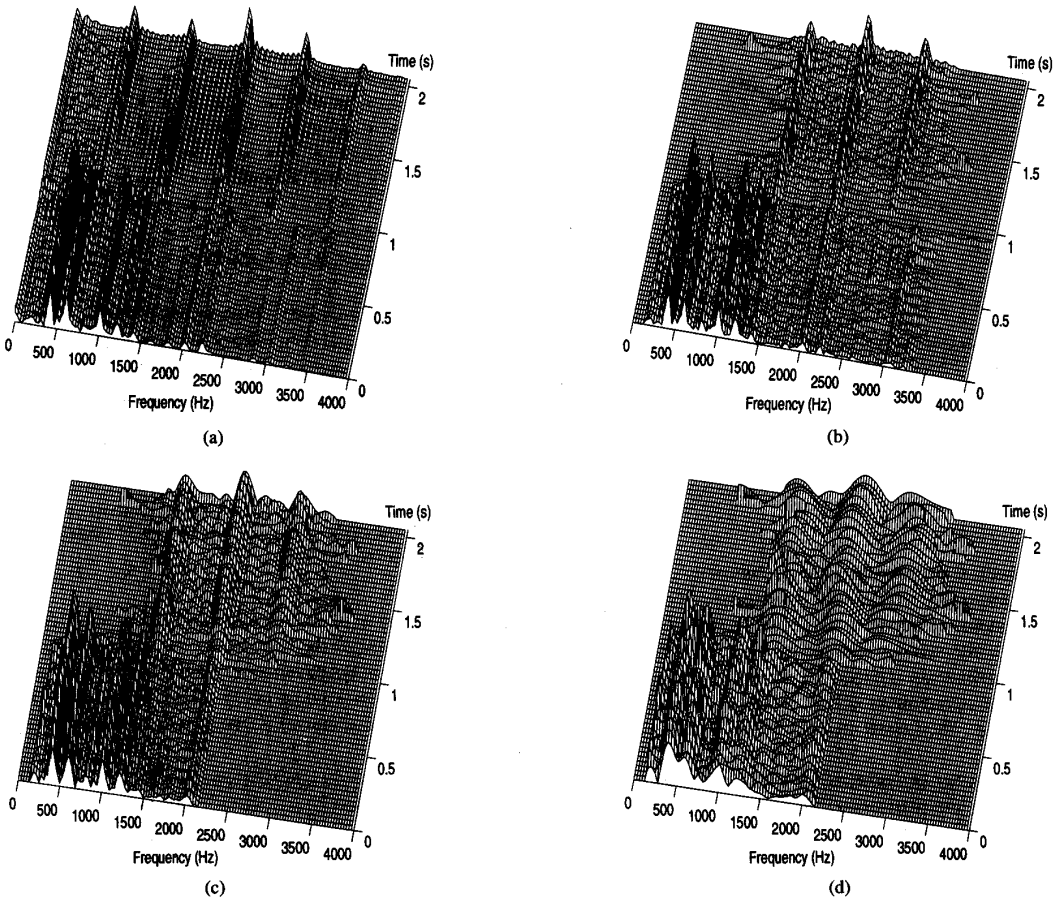

Fig. 3. Comparison of the exact and the approximate STFTs corresponding to a violin playing a sequence of two notes. Approximate STFTs were calculated using the hybrid narrowing approach with minimum frequency coverage constraint set to 2000 Hz. The plot in part (a) corresponds to the exact STFT. Plots in (b), (c), and (d) correspond to approximate STFTs with arithmetic complexity relative to that of a pruned FFT restricted to 50, 25, and 12.5%, respectively.

in Fig. 3. Each column of dots represents for a particular value of n an $N/2$-element vector $g^Q_{mL}(n)\boldsymbol{W}_n$. Such vectors are summed in accordance with (6) to obtain $\boldsymbol{X}^Q_{mL}$ for the mth signal frame. When the vector-summation process is restricted to region A in Fig. 3, the "time narrowing" (complete exclusion of vectors for particular times) results in a loss of frequency resolution. When the vector-summation process is restricted to region B in Fig. 3, we say that "frequency narrowing" (exclusion of a proper subset of each vector's elements) has taken place. In the case of region B, $\boldsymbol{X}^Q_{mL}$ is computed only for a subset of the frequency range (loss of frequency coverage) but without any loss in frequency resolution. On the other hand, region C depicts a type of frequency-dependent time narrowing, which results in loss of frequency resolution but no loss of frequency coverage.

Our approach to frame-adaptive exclusion of vector elements always excludes zero-valued vectors from the summation process. Clearly, there is no loss of output quality due to such exclusions. Let us assume that for the mth frame, there are N_v vectors remaining after the exclusion of zero-valued vectors. Since each vector is $N/2$ elements long, and in general, each element is complex, the unrestricted vector-summation process for the mth frame requires $N \times N_v$ real additions. The need for utilizing time narrowing and/or frequency narrowing arises only if $B < N \times N_v$.

In the case of frequency narrowing, it is usually desirable that the restricted frequency coverage be centered around a frequency component with significant energy. We have devised a heuristic [3], [4] for the frame-adaptive selection of the "center frequency." This involves measuring the number of nonzero samples in $g^Q_{mL}(n)$ and mapping it to a corresponding frequency in accordance with a monotonic curve used to model the data in Fig. 2. Although this technique is not guaranteed to give the best results in all situations, its practical utility is illustrated in the example to be presented in Section VI.

We have also devised a "frequency reversal" technique to help reduce the number of additions required whenever a frame is dominated by high-frequency energy. For each quantized frame $x^Q_{mL}(n)$, we compute its backward difference $g^Q_{mL}(n)$ as well as $s^Q_{mL}(n)$, which is the backward difference corresponding to $r^Q_{mL}(n) = (-1)^n x^Q_{mL}(n)$. Vector summation (with frequency and/or time narrowing) is then applied to the differenced sequence with the smaller N_v. If $s^Q_{mL}(n)$ is selected (which would be the case if $x_{mL}(n)$ is dominated by frequencies above 0.25 cycles/sample), vector summation yields $R^Q_{mL}(k)$, and $X^Q_{mL}(k)$ is obtained as $[R^Q_{mL}(N/2-k)]^*$ for $1 \leq k \leq N/2-1$.

IV. Frame Overlap

Further computational savings may be obtained in the evaluation of the approximate STFT when the analysis window is rectangular and there is overlap between consecutive signal frames. For example,

consider the case of half-window overlap. Let $x_{mL}(n) = y_{mL}(n) + h_{mL}(n)$, where

$$y_{mL}(n) = \begin{cases} x_{mL}(n) & \text{for } mL - N_w + 1 \\ & \le n \le mL - (N_w/2) \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

and

$$h_{mL}(n) = \begin{cases} x_{mL}(n) & \text{for } mL + (N_w/2) \\ & +1 \le n \le mL \\ 0 & \text{otherwise.} \end{cases} \quad (11)$$

Then, the two signals $y_{mL}(n)$ and $h_{mL}(n)$ are processed individually to form their quantized and backward differenced versions $\tilde{y}^Q_{mL}(n)$ and $\tilde{h}^Q_{mL}(n)$, respectively. The approximate transform corresponding to the mth frame then can be calculated as

$$X^Q_{mL}(k) = \sum_{n=mL-N_w+1}^{mL-(N_w/2)} \tilde{y}^Q_{mL}(n) W_n(k) + \sum_{n=mL-(N_w/2)+1}^{mL} \tilde{h}^Q_{mL}(n) W_n(k) \quad k = 1, \cdots, N/2. \quad (12)$$

Let us denote the first sum with $Y^Q_{mL}(k)$ and the second sum with $H^Q_{mL}(k)$. Since, for the next frame, $\tilde{y}^Q_{(m+1)L}(n)$ is equal to $\tilde{h}^Q_{mL}(n)$, the short-time transform for this frame can be calculated as

$$X^Q_{(m+1)L}(k) = H^Q_{mL}(k) + H^Q_{(m+1)L}(k). \quad (13)$$

By saving $H^Q_{mL}(k)$ for every frame, the short-time transform for the next frame can be calculated by performing operations only on the second half of that frame. This reduces the number of additions by approximately a factor of 2. The number of extra addition operations required to evaluate (13) is negligible compared with the number of additions saved on the basis of half-frame overlap between consecutive frames.

V. Quantization Overhead

We now consider the computational costs associated with frame quantization. In particular, we illustrate a procedure for frame quantization that is dominated by arithmetic operations. Furthermore, in comparison with the FFT-based evaluation of a frame's DFT, the number of arithmetic operations in our quantization procedure is smaller by more than an order of magnitude. In contrast, the number of additions in the *unrestricted* vector summation process can be as large as or, in some cases, greater than the number of arithmetic operations involved in an FFT-based evaluation of a frame's DFT. Consequently, as compared with frame quantization, the vector summation process provides a much larger playing field for improving the computational efficiency of the approximate STFT. However, the cost of frame quantization becomes a more significant factor as the bound B begins to approach values that are an order of magnitude smaller than the FFT's arithmetic cost.

We have used the *rounding* technique for frame quantization into Q uniform levels. This involves the assignment of each frame sample to a quantization level of value $2A_m n/Q$, where A_m is a frame-dependent parameter, and n is the *level number* that can take on one of Q different integer values. We calculate the value of A_m in terms of the *rms* value of the samples in a frame, that is

$$A_m = \sqrt{2}\left[\frac{1}{N_w}\sum_{n=mL-N_w+1}^{mL} x^2_{mL}(n)\right]^{1/2}. \quad (14)$$

The $\sqrt{2}$ scale factor ensures that for a pure sinusoid, the value of A_m is the same as the peak amplitude of the sinusoid. Each sample value may be divided by $2A_m/Q$ and rounded to the nearest integer in order to determine the corresponding level number. To ascertain the computational cost of this quantization procedure, we note that the calculation of A_m requires N_w squaring operations and N_w addition operations. The assignments of level numbers to all the frame samples requires a total of N_w divisions and N_w rounding operations. We conclude that the complexity of the quantization procedure is on the order of $4N_w$ arithmetic operations. In Table I, we have tabulated the ratio of the arithmetic complexity for the quantization procedure to the arithmetic complexity $(N/2)\log(N_w)$ of a pruned N-point FFT algorithm. It is observed from the table that for various values of N_w and N, this ratio is significantly lower than 0.1.

TABLE I
Ratio of the Arithmetic Complexity for the Quantization Procedure ($4N_w$ Operations) to the Arithmetic Complexity of a Pruned N-Point FFT Algorithm ($N/2\log(N_w)$ Operations). The Ratios are Calculated for (N_w, N) Pairs Such that $N \ge 2N_w$.

	N=128	N=256	N=512	N=1024
N_w=64	0.06	0.03	0.015	0.0075
N_w=128	*	0.05	0.025	0.0125
N_w=256	*	*	0.05	0.025
N_w=512	*	*	*	0.044

VI. Example

Our example corresponds to a 16 000-point signal obtained by sampling at 8 KHz the sound of a violin playing a sequence of two notes. We performed exact and approximate STFT analysis with a 128-point rectangular window, a DFT size of 256, and a decimation interval of 64. Using a time-pruned FFT algorithm [5], the exact STFT (displayed in Fig. 3(a)) required 8960 real arithmetic operations[1] (40% multiplications) per frame. The rest of Fig. 3 corresponds to various STFT approximations based on three-level quantization and utilizing time and frequency narrowing with minimum frequency coverage constrained at 2000 Hz. For parts (a)–(c), the number of additions per frame relative to the total arithmetic complexity (including the number of multiplications) of the FFT were restricted to 50, 25, and 12.5%, respectively. Observe that the highest energy harmonics are always captured (a result of the "frequency centering" technique), and frequency resolution begins to decrease once the 2000-Hz constraint on frequency coverage is reached.

References

[1] S. H. Nawab and T. Quatieri, "Short-time Fourier transform," in *Advanced Topics in Signal Processing*, J. S. Lim and A. V. Oppenheim, Eds. Englewood Cliffs, NJ: Prentice-Hall, 1988.
[2] M. P. Lamoureux, "The Poorman's transform: Approximating the Fourier transform without multiplication," *IEEE Trans. Signal Processing*, vol. 41, no. 3, pp. 1413–1415, Mar. 1993.
[3] E. Dorken, "Approximate processing and knowledge based reprocessing of nonstationary signals," Ph.D. dissertation, Boston Univ., Sept. 1993.
[4] E. Dorken and S. H. Nawab, "Frame-adaptive techniques for quality versus efficiency tradeoffs in STFT analysis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Procesing*, Apr. 1994, pp. 449–452, vol. 3.
[5] D. P. Skinner, "Pruning the decimation in-time FFT algorithm," *IEEE Trans. Acoust., Speech, Signal Processing*, pp. 193–194, Apr. 1976.
[6] H. V. Sorensen and C. S. Burrus, "Efficient computation of the short-time fast Fourier transform," in *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Procesing*, Apr. 1988, pp. 1894–1897, vol. 3.

[1] This number is reduced by approximately 15% if algorithm two in [6] is used for the exact STFT.