ADAPTIVE SHORT-TIME ANALYSIS-SYNTHESIS FOR SPEECH ENHANCEMENT

Daniel Rudoy [*†], *Prabahan Basu* [*], *Thomas F. Quatieri* [†], *Bob Dunn* [†], *and Patrick J. Wolfe* [*]

[*] Harvard University
School of Engineering and Applied Sciences
33 Oxford Street Cambridge, MA 02138
{rudoy, basu, patrick}@seas.harvard.edu

[†] MIT Lincoln Laboratory
244 Wood Street, Lexington, MA 02420
{quatieri, rbd}@ll.mit.edu

ABSTRACT

In this paper we present a new adaptive short-time Fourier analysis-synthesis scheme and demonstrate its efficacy in speech enhancement. While a number of adaptive analyses have previously been proposed to overcome the limitations of fixed-resolution schemes, we propose here a modified overlap-add procedure that enables efficient resynthesis. Our adaptation scheme extends earlier work using local measures of time-frequency concentration, and is applicable to power spectral density estimation for the case of noisy speech. We provide evidence of increased gains in signal-to-noise ratios for synthetic signals as well as empirical evidence of reduced musical noise based on expert listening tests for voiced and phonetically balanced utterances observed in noise, relative to a standard baseline speech enhancement system whose time-frequency resolution is fixed.

Index Terms— Analysis-synthesis, adaptive segmentation, speech enhancement, time-frequency concentration, filterbanks

1. INTRODUCTION

Most short-time Fourier analysis-synthesis schemes used in speech applications such as time-scale modification and enhancement employ a fixed-resolution decomposition of the time-frequency plane. In the traditional short-time Fourier transform (STFT), the trade-off between time and frequency resolution is controlled by a (typically smooth, symmetric) window function whose support is ordinarily chosen in the range 15–30 ms. It is known, however, that certain speech sounds such as vowels are well modeled as stationary processes over 40–80 ms segments, whereas transient events such as plosives occur on a much shorter time scale. Controlling the time-frequency resolution of STFT analysis in a signal-adaptive way is therefore desirable in order to avoid smearing transients while at the same time maximally preserving steady-state harmonic content.

In the general context of time-frequency analysis, a number of adaptive analyses have been proposed with this goal in mind. Adaptive STFT schemes such as those described in [1, 2] and references therein, however, are designed for analysis only, and lack efficient reconstruction operators. More recent work has attempted to overcome this problem: for instance, an adaptive local trigonometric transform was incorporated into a speech enhancement scheme in [3], and [4] proposed hypothesis tests to discriminate between transient and stationary regions. The promise of such methods for speech enhancement is predicated on the notion that estimates of both the local speech and noise spectra can be made more robust through adaptive methods that match the time-frequency structure of the underlying speech signal [1]. This in turn may lead to lower-variance estimates of the local speech spectrum, thus contributing to a reduction in the well-known "musical noise" artifact [5].

In this paper, we extend the analysis method of [2] to provide a new adaptive short-time Fourier analysis-synthesis scheme that in turn yields a signal-dependent speech enhancement method. As described in Section 2, our approach is based on a local measure of time-frequency concentration introduced in [2], but with the addition of a new modified overlap-add procedure that enables efficient resynthesis. In Section 3 we provide empirical results demonstrating the applicability of this adaptive analysis-synthesis system to speech enhancement. We then conclude with a brief discussion in Section 4.

2. ADAPTIVE ANALYSIS-SYNTHESIS

2.1. Time-frequency concentration

Our signal-dependent short-time analysis approach is based on an efficient adaptive scheme first proposed in [2] where spectral kurtosis is used as a measure of the local time-frequency concentration for each member of a set of competing STFTs generated using different window lengths. Specifically, for a signal with STFT, $X_p(t,\omega)$, where the parameter p indexes the length of the underlying analysis window, the local spectral kurtosis as a function of time is given by:

$$C(t,p) = \frac{\iint |X_p(\tau,\omega)w(\tau-t)|^4 d\tau d\omega}{\left(\iint |X_p(\tau,\omega)w(\tau-t)|^2 d\tau d\omega\right)^2}, \qquad (1)$$

where $w(t)$ is a window centered at 0 that localizes the measure. Maximizing time-frequency concentration as per (1) favors short-time segments that place most of the energy in the smallest region of the time-frequency plane. In particular, shorter windows are chosen around time-localized transients such as plosives, since this will produce the most concentrated energy distribution of the STFT coefficients. On the other hand, vowels and voiced consonants, which are oscillatory in nature, will tend to be spread over time but localized in frequency. Thus, energy concentration will be maximized when spectrally concentrated, temporally broad windows are used. When spectral peaks are rapidly varying across time (due for instance to rapid vocal tract variation in the case of speech), we have observed that the spectral kurtosis measure tends to adapt window length to the motion of the formant tracks—shorter windows are selected if formants are rapidly changing.

In [2] an adaptive STFT is constructed by optimizing the instantaneous window length in order to maximize the time-frequency concentration computed in (1). In particular, to construct an adaptive STFT from M fixed-resolution STFTs, M spectrograms are computed and interpolated onto the finest time-frequency lattice. Next, around each time coordinate in this lattice, a discretized version of (1) is calculated for each STFT and the window length corresponding to the STFT with the maximal local concentration is se-

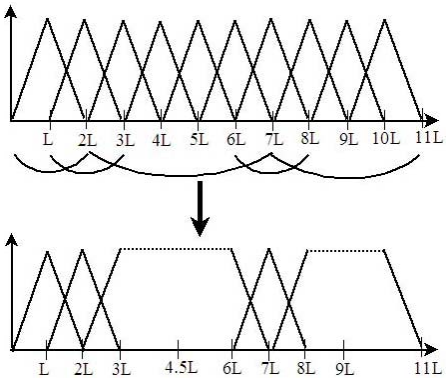

Fig. 1. *An example of how a fixed-resolution scheme (top) is modified to achieve an adaptive time segmentation (bottom).*

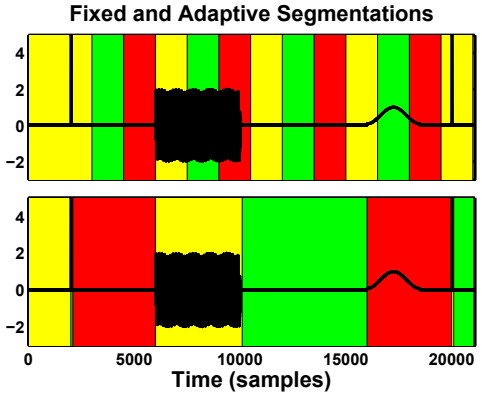

Fig. 2. *Fixed resolution (top) and adaptive (bottom) short-time analysis of a simple synthetic waveform.*

lected. Since the computation is done at the finest time-frequency lattice, the resultant representation is highly redundant. As described in Sections 2.2 and 2.3 below, we modify and extend this scheme to enable efficient synthesis.

2.2. Iterative Clustering for Adaptive Analysis

We first describe a new adaptation scheme based on iterative clustering. After computing a fixed-resolution STFT using the shortest desired window $w[n]$ of length $2L$ with frame step size of L samples, the method iteratively merges neighboring windows based on a modified version of the time-frequency concentration measure (1). At each iteration in the adaptation scheme, a decision is made to either merge two adjacent short-time segments and analyze them using a single longer window or to analyze them using distinct shorter windows. In particular, consider two neighboring short-time segments $x_l = x[n]w_l[pL - n]$ and $x_r = x[n]w_l[(p + 1)L - n]$, where x_l and x_r denote segments centered at pL and $(p+1)L$ respectively. We define a new, merged, window w_m centered at $(p + 1/2)L$ used to compute the short-time segment $x_m = x[n]w_m[pL - n]$ as follows:

$$w_m[n] = w_l[pL - n] + w_r[(p + 1)L - n]. \quad (2)$$

An example of this is shown in Figure 1. The time-frequency concentration is then computed for each of the three resulting short-time sections using:

$$C(x_w) = \frac{\sum_k \left|\sum_n x_w[n]e^{-j2\pi kn/N}\right|^4}{\left(\sum_k \left|\sum_n x_w[n]e^{-j2\pi kn/N}\right|^2\right)^2}, \quad (3)$$

where $x_w[n]$ indexes the short-time section $x_l[n]$, $x_r[n]$ or $x_m[n]$. If the time-frequency concentration of the concatenated short-time segment denoted by $C(x_m)$ exceeds the maximum of the concentrations of the individual short-time segments, denoted by $C(x_l)$ and $C(x_r)$ respectively, then we use the merged window w_m instead of w_l and w_r. The complete algorithm proceeds by growing each of the individual windows in the finest-resolution scheme according to the above criterion and is summarized in Algorithm 1.

Following [2], we use a waveform comprised of a sum of sinusoids, two impulses and a bump function to illustrate our analysis scheme. Figure 2 shows a fixed (top) and adaptive (bottom) segmentation, with varying widths of the superimposed rectangles corresponding to the temporal extent of the underlying analysis windows and different colors providing a visual contrast.

Algorithm 1 Adaptive STFT

1. Set $w_l[n] = w_r[n] = w[n]$. Set $p = q = 1$
2. Compute the spectral kurtosis of the current left short-time segment $x_l = x[n]w_l[pL - n]$ by (3)
3. Compute the spectral kurtosis of the current right short-time segment $x_r = x[n]w_r[(p + q)L - n]$ using (3)
4. Set $w_m[n] = \sum_{k=p}^{p+q} w[kL - n]$ and compute the spectral kurtosis for the combined frame $x_m = x[n]w_m[pL - n]$
5. If $C(x_m) > \max(C(x_l), C(x_r))$ then set $w_l[n] = w_m[n]$ and q=q+1, otherwise set p=p+q, $q = 1$, and set $w_l[n]$=$w[n]$

2.3. Synthesis

Our second contribution is an efficient scheme for resynthesis. Even though synthesis is theoretically possible through the inversion of the underlying (Gabor) frame operator, to our knowledge no practical method is known. On the other hand, the iterative approach we have taken allows for an efficient synthesis procedure based on the well-known overlap-add method [1]. Given a set of STFT coefficients $X(pL, k)$, where L is a time-decimation factor, the synthesis of a sequence $y[n]$ through the overlap-add method (OLA) is given by:

$$y[n] = \frac{L}{W(0)} \sum_{p=-\infty}^{\infty} \left(\frac{1}{N} \sum_{k=0}^{N-1} X(pL, k)e^{j\frac{2\pi}{N}kn} \right), \quad (4)$$

where $W(0) = \sum_{n=-\infty}^{\infty} w[n]$. Perfect reconstruction (i.e., when $y[n] = x[n]$) is achieved if the following OLA constraint is met:

$$\sum_{p=-\infty}^{\infty} w[pL - n] = \frac{W(0)}{L}. \quad (5)$$

This constraint requires that all the analysis windows form a partition of unity; see, for example, Figure 1.

In our adaptive scheme, synthesis relies on the fact that all longer windows were constructed using (2). Hence, at the end of the adaptive analysis, each selected window can be decomposed into a summation of windows that were used to compute the initial fixed-resolution STFT. Thus, if the windows chosen for the initial STFT satisfy the OLA constraint (5) then so does the set of variable length windows derived through the application of Algorithm 1. Consequently, efficient synthesis is possible by using the OLA procedure (4).

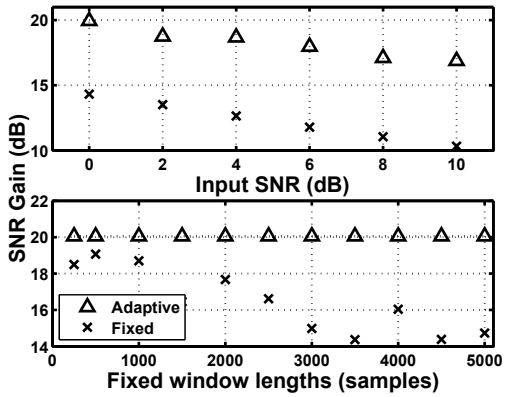

Fig. 3. *Performance comparison of the fixed and adaptive-resolution systems in enhancement of a synthetic waveform as input SNR is varied (top) or STFT analysis window lengths of are varied (bottom).*

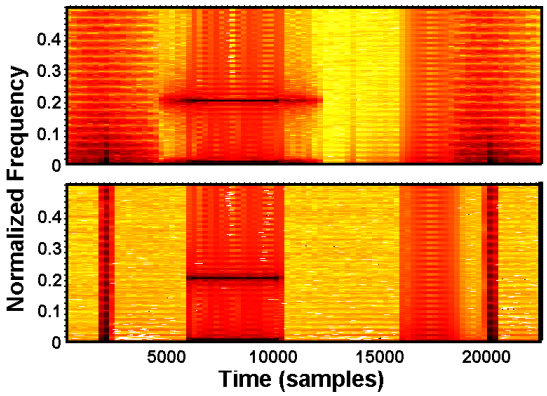

Fig. 4. *Spectrograms of the synthetic waveform denoised by the fixed (top) and adaptive (bottom) schemes using* 512*-sample Hamming windows with* 50% *overlap.*

2.4. Synthetic Denoising Example

The synthesis module enables the evaluation of the adaptive short-time scheme in a denoising setting. Here, white noise is added to the synthetic waveform shown in Figure 2 and the standard Wiener rule, with the spectra of the clean and noise signals given, is used for enhancement, so that the only differences in performance are due to the adaptive segmentation. The fixed-resolution and adaptive schemes utilize the windows shown in Figure 2. Figure 3 shows the SNR gains for each scheme computed over a range of input SNRs (top) and as the length of the analysis window for the STFT is varied in increments of 500 samples (bottom) while input SNR was fixed at 2 dB. It is clear that the adaptive scheme not only achieves higher gains than the fixed scheme for a range of input SNRs, but also outperforms it regardless of what fixed window length is used. The difference between the two schemes is underscored in the spectrograms of the denoised waveform shown in Figure 4 where we see that the fixed-resolution scheme smears the signal components in time.

3. ADAPTIVE SPEECH ENHANCEMENT

As an example of the applicability of this new adaptive analysis-synthesis scheme to real data, we demonstrate its performance in the context of speech enhancement. Here, noise reduction is typically achieved through the spectral attenuation of each short-time segment and so selecting the appropriate temporal resolution is crucial. For instance, the importance of using short windows for transients is underscored in the spectrograms of the word "piecemeal" shown in Figure 5. The onset of the plosive "p" is preserved when our adaptive-resolution scheme is employed for enhancement, in the manner discussed in Section 3.1, and smeared when the fixed-resolution STFT is used. However, if we were to use the same short window in the voiced segments, then their harmonic structure would be smeared (noise would be suppressed in formant rather than harmonic nulls). Indeed, the adaptive scheme uses longer windows in voiced segments; we see in Figure 5 that the harmonic structure of the voiced portions is preserved as well as in the fixed-resolution scheme.

Another example of our adaptive analysis scheme is shown in Figure 6 which depicts the fixed (top) and adaptive (bottom) time-segmentations of the phrase "and amazed" taken from a TIMIT utterance. The fixed-resolution scheme uses 20-ms triangular windows, while the base window size of the adaptive scheme was set to 10 ms; the windows overlap by 50% in both schemes. It is evident that longer windows are chosen for the voiced segments while shorter segments are chosen for transients such as the voiced plosive "d" at the end of the word "amazed".

In order to appropriately evaluate our adaptive analysis-synthesis scheme in the context of speech enhancement, we compare its performance against that of a fixed-resolution scheme using a baseline system described in Section 3.1. We report differences in SNR gain together with results of informal listening tests aimed at assessing whether the amount of musical noise is reduced in the signal enhanced by the adaptive system. As has been observed in [3] and [5], if the adaptation is properly segmenting stationary and transient regions, then the variance of the estimates of the speech spectrum is reduced, thereby reducing the amount of musical noise present. In particular, we show that the adaptive scheme can reduce the amount of musical noise without relying on inter-frame smoothing, which also implies great potential to preserve transients.

3.1. Enhancement System

We assume the standard additive observation model: $y[n] = x[n] + w[n]$ where $x[n]$ is the clean speech signal, $w[n]$ is white Gaussian noise and $y[n]$ is the resultant noisy signal. The adaptive enhancement system makes use of a 10-ms triangular window with a 5-ms frame rate, as a base tiling which satisfies the OLA constraint (5), and uses the iterative scheme of Section 2.2 to adapt the lengths of the analysis windows. Subsequently, the following baseline enhancement scheme, is applied. The speech magnitude spectrum $|\tilde{X}(\omega)|$ is first estimated by a magnitude suppression rule given by: $|\tilde{X}(\omega)| = \left(\frac{|Y(\omega)|}{|Y(\omega)|+\sigma_w}\right)|Y(\omega)|$ where $\tilde{X}(\omega)$ is subsequently smoothed by an 11 sample Hanning window. The noise variance σ_w^2 is provided to the enhancement scheme since, in practice, it is typical to estimate it from regions of speech absence. Finally, the enhanced coefficients are obtained by an application of the following magnitude suppression rule: $\hat{X}(\omega) = \left(\frac{|\tilde{X}(\omega)|}{|\tilde{X}(\omega)|+\sigma_w}\right)Y(\omega)$.

3.2. Experimental Results and Expert Listening Tests

Using the enhancement scheme described above, we evaluate the performance of the adaptive analysis-synthesis system and compare

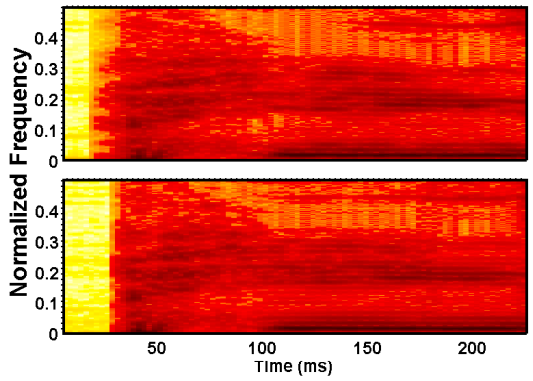

Fig. 5. *Spectrograms of the utterance "piecemeal" observed in white Gaussian noise at* 0 *dB SNR and enhanced using the adaptive (top) and fixed resolution (bottom) schemes.*

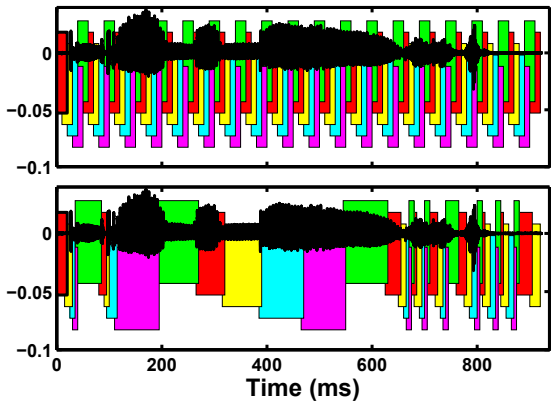

Fig. 6. *Fixed resolution and adaptive short-time analyses of the phrase "and amazed" extracted from a TIMIT utterance.*

it to that of a fixed-resolution STFT that uses 20 ms windows with 50% overlap using a small corpus of voiced speech data collected at MIT Lincoln Laboratory consisting of 5 male and 5 female speakers each uttering the sentences: "Why were you away a year Roy?" and "Nanny may know my meaning." In addition, four phonetically balanced TIMIT utterances (2 male speakers, 2 females speakers) were added. First, we measured SNR gain obtained by both schemes for a range of input SNRs (0–10 dB) and found that the adaptive scheme consistently measured better (0.5–1.5 dB gain).

	Adaptive System	No Preference	STFT
Voiced	74.3%	21.3%	4.4%
TIMIT	33.8%	53.7%	12.5%

Table 1. Summary of listener preferences from tests for the enhancement scheme resulting in least musical noise. The results are averaged over ten listeners and all voiced and TIMIT utterances.

We also conducted a series of informal listening tests with ten trained listeners at MIT Lincoln Laboratory to evaluate whether using the adaptive system reduced the amount of musical noise in the enhanced waveform as compared to a fixed resolution system. Each listener was presented with the two voiced sentences previously described, each spoken by two male and two female speakers, at 0 dB and 5 dB SNR, for a total of 16 utterances. Each listener was also presented with four TIMIT sentences (2 male speakers, 2 female speakers) for a total of 8 utterances. For each presented waveform the listener heard the clean and the noisy samples followed by two repetitions of the enhanced sentences by both algorithms in a random order. At the end of the presentation the listener was asked which of the enhanced waveforms, if any, had musical noise. The results are summarized in Table 1. Among the voiced data, a large majority of responses (74.4%) indicated that less musical noise was present in the waveforms enhanced by the adaptive system—less than 5% preferred the fixed resolution system. The results using the phonetically balanced TIMIT utterances are more modest, but nonetheless promising. We suspect that the performance gap between the two cases can be closed through the incorporation of a probabilistic model of speech presence. Overall, however, the results show that the adaptation can help reduce the musical noise artifact.

4. DISCUSSION

In this work, we have proposed an iterative adaptive short-time analysis scheme based on a measure of time-frequency concentration together with an efficient procedure for resynthesis. We have evaluated our scheme in the context of enhancement of synthetic data and have explored its applicability in the context of speech enhancement. Results indicate that the analysis scheme adapts well to the time-frequency structure of speech and consequently allows for improved enhancement as measured both by SNR and informal listening tests. We anticipate that further improvements may be brought about by incorporating a probabilistic model of speech presence.

Acknowledgements: Lincoln Laboratory authors were supported by the Department of Defense under Air Force contract FA8721-05-C-0002. The opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government. Harvard authors were supported by the Defense Advanced Research Projects Agency under Grant No. HR0011-07-1-0007, by the National Science Foundation under Grant No. DMS-0652743. The first author is supported by a National Defense Science and Engineering Graduate Fellowship.

5. REFERENCES

[1] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*, Prentice Hall, 2002.

[2] D. L. Jones and R. G. Baraniuk, "A simple scheme for adapting time-frequency representations," *IEEE Transactions on Signal Processing*, vol. 42, no. 12, pp. 3530–3535, 1994.

[3] S. Srinivasan and W. B. Kleijn, "Speech enhancement using adaptive time-domain segmentation," *Proceedings of the International Conference Spoken Language Processing (ICSLP)*, pp. 869–872, 2004.

[4] R. C. Hendriks, R. Heusdens, and J. Jensen, "Adaptive time segmentation for improved speech enhancement," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 6, pp. 2064–2074, 2006.

[5] Y. Hu and P. C. Loizou, "Speech enhancement based on wavelet thresholding the multitaper spectrum," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 1, pp. 59–67, 2004.