AN ADAPTIVE TIME-FREQUENCY ANALYSIS SCHEME FOR IMPROVED REAL-TIME SPEECH ENHANCEMENT

Kristian Timm Andersen[1,2] *Marc Moonen*[1]

[1]KU Leuven, ESAT/STADIUS, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium
[2]Widex A/S, Nymøllevej 6, DK-3540 Lynge, Denmark

ABSTRACT

An adaptive time-frequency analysis scheme is proposed for improved real-time speech enhancement. The proposed scheme uses a filtering of the short-time Fourier transform (STFT) to obtain an adaptive resolution and is computationally efficient, since it uses fast Fourier transforms (FFTs) in the analysis and synthesis filters. Unlike previously suggested methods, the proposed method allows the time-frequency resolution to be chosen independently for each frequency bin. Perfect reconstruction is achieved by calculating the speech enhancement gains based on the adaptive analysis and then applying these to a STFT which is then filtered to allow the causal part of the gains to pass through the synthesis filter. The proposed method is shown to have superior performance compared to a fixed resolution STFT scheme with an equal time delay of 10 ms.

Index Terms— Adaptive short-time Fourier transform, real-time systems, speech enhancement.

1. INTRODUCTION

Traditional speech enhancement techniques in real-time sound processing devices split the input signal into a number of frequency bands, process each band according to a selected strategy, and combine the bands into a broadband output signal. The width and sharpness of the filters effectively determines the resolution in time and frequency. However, in a speech signal some segments consist of very specific frequency components that are stationary over long periods (e.g., vowels) while other signal segments have a very short duration but span a wide frequency range (e.g., many consonants). In speech analysis, frames of 30-50ms are common. Such a time resolution is far from optimal when processing transients, onsets or plosives which are known to have a duration of less than 5ms [1]. If the analysis is not adapted to the signal components, it is obviously hard to find an appropriate trade-off between resolution in time and resolution in frequency. Additionally, it is instrumental in real-time processing that the time delay introduced by the processing is kept very low, in the order or 10 ms [2]. The choice of filter bank is consequently a fundamental decision for real-time speech enhancement as it is indeed bound to limit some aspects of the performance.

To overcome the problems with a fixed filter bank, an adaptive STFT scheme has previously been suggested [3]. The method allows the time-frequency resolution to be chosen freely at a given time, but forces all frequency bins to have the same resolution and is not suitable for low delay implementation. The method proposed in this paper allows the time-frequency resolution to be chosen independently for each frequency bin while still allowing perfect reconstruction of the enhanced signal. In addition, it is suitable for low delay implementation and has a low computational complexity. Previous work that is suitable for low-delay implementation considered only window switching for a predefined set of windows that must be equal for all frequency bins [4][5]. The proposed method uses filtering of the STFT in the frequency domain and utilizes a novel synthesis stage that is appropriate for low delay processing.

The paper is organized as follows. The proposed analysis and synthesis scheme is described in section 2 and 3 respectively. Section 4 through 6 contains a description of a complete speech enhancement scheme. Experimental results are found in section 7 and the conclusion is given in section 8.

2. ADAPTIVE ANALYSIS

The adaptive time-frequency analysis uses an STFT with a suitable short window such as a periodic *generalized cosine window*. A longer window, with better frequency resolution, is obtained by summing this window with the succeeding windows with a specified hop-size. An example of a suitable window is the periodic Hann window of length N with a hop-size of $R = N/4$ given by:

$$h(n) = 0.5 \cdot \left(1 - \cos\left(\frac{2\pi n}{N}\right)\right), \quad 0 \leq n < N \qquad (1)$$

The window is grown until a non-stationarity is detected and then the window is reduced down to the original short window.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

Given a sum of M windows:

$$g_M(n) = \sum_{m=0}^{M-1} h(N - L + mR + n) \quad (2)$$

The frequency analysis of $x(n)$ is calculated using the STFT:

$$X_M(k,i) = \sum_{n=0}^{L-1} g_M(n)x(n+iR)e^{-2\pi jnk/L} \quad (3)$$

where $L >> N$ is the DFT size, k is the frequency index and i is the decimated time index. For each new decimated time index, the sum of windows is either reset to the short window $g_1(n)$ or grown by one short window into g_{M+1}. In the latter case, the frequency analysis is calculated as:

$$\begin{aligned} X_{M+1}(k,i) &= \sum_{n=0}^{L-1} g_{M+1}(n)x(n+iR)e^{-2\pi jnk/L} \\ &= \sum_{n=0}^{L-1} (g_1(n) + g_M(n+R))x(n+iR)e^{-2\pi jnk/L} \\ &= X_1(k,i) + X_M(k,i-1)e^{2\pi jRk/L} \end{aligned} \quad (4)$$

i.e. the new short window frequency analysis is added to the previous sum of windows frequency analysis shifted by $W_R = e^{2\pi jRk/L}$, which is equivalent to a time-shift of R in the time domain. It is noted that each frequency bin can be updated independently, so it is possible to reset the window for one frequency bin by calculating $X_1(k,i)$, while growing the sum of windows for another bin. This means that each frequency bin can have its own window, i.e. its own value of M and therefore its own time-frequency resolution. Also, due to the periodicity of the DFT basis functions, M can effectively be arbitrarily large so that L no longer defines the DFT-size, only the frequency bins for the frequency analysis. In this case, $g_M(n)$ should be bounded (BIBO stability) [6]. The filter in (4) does not satisfy this constraint, so in the following only BIBO stable filters are considered:

$$\widetilde{X}(k,i) = \sum_{p=0}^{P-1} b_p X_1(k,i-p)W_R^p + \sum_{p=1}^{P-1} a_p \widetilde{X}(k,i-p)W_R^p \quad (5)$$

where a_p and b_p are the real filter coefficients of a P'th order BIBO stable filter. An example of a 1st-order auto-regressive (AR) filter is seen in Figure 1. As long as the signal is stationary in frequency bin k, $\widetilde{X}(k,i)$ is updated using (5). When a non-stationarity is detected, the filter output is reset to $X_1(k,i)$ and all filter nodes are reset to 0. In this way, the analysis window gradually grows from $X_1(k,i)$ to the full window with the improved frequency resolution. As the window grows, the underlying function $\widetilde{g}_M(n)$ also grows and for analysis purposes $|\widetilde{X}_M(k,i)|^2$ indicates the energy of the current adaptive time-frequency bin normalized by the precomputed energy of the underlying window function $\widetilde{g}_M(n)$.

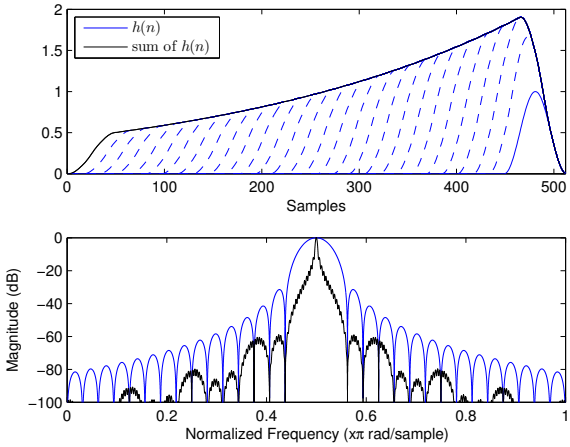

Fig. 1. *Top: A sum of Hann windows for the filter* $\widetilde{X}(k,i) = X_1(k,i) + 0.95 \cdot \widetilde{X}(k,i-1)W_R$ *and* $R = N/4$. *Bottom: Magnitude spectrum of one frequency band (energy normalized) for the first window and the sum of windows for* $L = 512$.

An example of a spectrogram of a speech signal is seen in Figure 2. It is seen that the adaptive window resets to the short window at all onsets and offsets while growing to the long window when the signal is stationary. It is also seen that only frequency bins with a significant energy change are reset, for instance at 0.89 seconds for the frequency bins below 1 kHz.

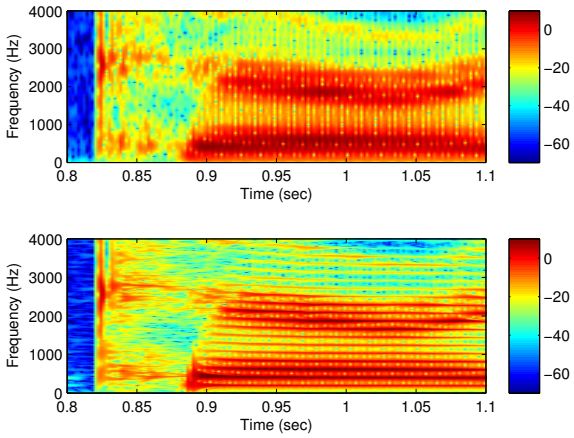

Fig. 2. *Spectrogram of the start of the word 'carrying' using* $|X_M(k,i)|^2$ *(top) and* $|\widetilde{X}_M(k,i)|^2$ *(bottom). It is seen that the adaptive analysis clearly shows both the transient at 0.82 seconds and the harmonics of the following voiced sound.*

3. SYNTHESIS WITH LOW DELAY

In the proposed analysis scheme each frequency bin has its own time-frequency resolution. This means that synthesis through the inverse DFT (IDFT) is not possible since this assumes that the underlying window function is the same for all frequency bins at a given time. Therefore, it is proposed to calculate a gain $G_{\widetilde{X}_M}(k,i)$ from $\widetilde{X}_M(k,i)$, see e.g. section 5, and then apply this gain to $X_1(k,i)$: $Y_1(k,i) = G_{\widetilde{X}_M}(k,i)X_1(k,i)$. Since $X_1(k,i)$ corresponds to a standard STFT it is possible to invert and synthesize $Y_1(k,i)$ using Overlap-Add (OA) [6]. However, to accommodate a real-time implementation with a low delay, a synthesis window $w(n)$ [7] must be applied to the inverse of $Y_1(k,i)$, denoted $y_i(n) = \text{IDFT}[Y_1(k,i)]$. $w(n)$ is usually chosen to be shorter or equal to the length of $h(n)$ and since $G_{\widetilde{X}_M}(k,i)$ is calculated based on the adaptive analysis, it is a potentially much longer filter than $w(n)$ which means that the synthesis window severely attenuates the filter ringing from $G_{\widetilde{X}_M}(k,i)$. As delay restrictions prevent calculating the filter ringing due to later frames, it is proposed to reconstruct part of the attenuated filter ringing in a similar way as in (5):

$$\widetilde{Y}(k,i) = \sum_{p=0}^{P-1} c_p Y_1(k,i-p)W_R^p + \sum_{p=1}^{P-1} d_p \widetilde{Y}(k,i-p)W_R^p \tag{6}$$

By transforming the filter to the time-domain it is seen that (6) corresponds to a filtered version of OA using only the previous frames of $y_i(n)$ and that standard OA can be considered an acausal filtering of $Y(k,i)$[1]. It is noted that standard OA can be made causal by accepting the increased delay from processing the signal in the full frame of length L instead of only in the analysis window of length N.

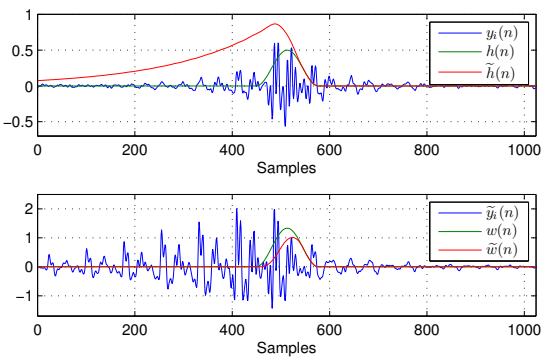

Fig. 3. *An output frame before (top) and after (bottom) applying a 1st order AR filter in equation (6) along with analysis and synthesis windows.*

[1]There is a technical difference between the time- and frequency domain methods as the frequency domain method is subject to circular convolution.

To provide perfect reconstruction (PR), a modified synthesis window $\widetilde{w}(n)$ must also take (6) into account:

$$\widetilde{w}(n) = w(n)\frac{h(n)}{\widetilde{h}(n)} \tag{7}$$

where $\widetilde{h}(n)$ is calculated by using $h(n)$ as input to the corresponding time-domain filter of (6) and $w(n)$ is chosen to provide PR for $X(k,i)$. An example is shown in Figure 3.

4. DETECTING SIGNAL NON-STATIONARITY

There are many ways to detect the duration of a stationary signal. Here the Likelihood Ratio Test from [4] is used. To determine if a new $X_1(k,i)$ belongs to the same statistical process as $\widetilde{X}(k,i-1)$, the following test statistic is used:

$$\text{LR} = \frac{|X_{1,M}(k,i)|}{|\widetilde{X}_M(k,i-1)|}\exp\left(-0.5\left(\frac{|X_{1,M}(k,i)|^2}{|\widetilde{X}_M(k,i-1)|^2} - 1\right)\right) \tag{8}$$

which is compared to a threshold value λ. If $LR > \lambda$, the adaptive analysis is updated using (5). Otherwise, the filter is reset to $\widetilde{X}(k,i) = X_1(k,i)$ and all filter nodes are reset to 0. To make (8) less susceptible to random fluctuations in the energy, the filter is only reset if a certain number V of adjacent bins $X_1(k,i), ...X_1(k+\Delta,i)$ fail the test statistic. Since that can leave bins 'hanging' if less than V adjacent bins exist that are not reset, these 'hanging' bins are also reset.

5. SPEECH ENHANCEMENT BASED ON ADAPTIVE ANALYSIS

The adaptive analysis $|\widetilde{X}_M(k,i)|^2$ is used as input to a speech enhancement algorithm. The gain is calculated as the MMSE-log gain [8]:

$$G_{\widetilde{X}_M}(k,i) = \frac{\xi(k,i)}{1+\xi(k,i)}\exp\left(\frac{1}{2}\int_{\nu(k,i)}^{\infty}\frac{e^{-t}}{t}dt\right) \tag{9}$$

where

$$\nu(k,i) = \frac{\xi(k,i)}{1+\xi(k,i)}\gamma(k,i), \quad \gamma(k,i) = \frac{|\widetilde{X}_M(k,i)|^2}{\lambda_N(k,i)}$$

and $\xi(k,i)$ is the *a posteriori* signal-to-noise ratio (SNR) [9], which is calculated recursively:

$$\xi(k,i) = \alpha \cdot \frac{G_{\widetilde{X}_M}(k,i-1)^2|\widetilde{X}_M(k,i-1)|^2}{\lambda_N(k,i-1)} + (1-\alpha)\cdot\max(\gamma(k,i)-1,0) \tag{10}$$

where $\lambda_N(k,i)$ is the noise estimate and $\alpha = 0.98$ is the smoothing parameter. The gain is applied to $X_1(k,i)$ as described in section 3:

$$Y_1(k,i) = G_{\widetilde{X}_M}(k,i)X_1(k,i) \tag{11}$$

6. COMPUTATIONAL COMPLEXITY AND DELAY

Only the computational complexity of the adaptive analysis and synthesis is evaluated here, since any speech enhancement algorithm and non-stationarity criteria can be applied in the overall processing scheme. Both the analysis and synthesis operations use an (I)DFT of length L, which can be efficiently computed using the FFT algorithm. Also, since L usually is a power of 2 and $N << L$, most of the samples in the FFT are zero, which enables the use of a pruned FFT [10] to further reduce the complexity. The adaptive analysis and synthesis uses two filters in each frequency bin, which can be chosen as 1st-order AR filters.

The delay is determined by the combined analysis and synthesis window and the hopsize R. Considering first a fixed filter bank, if the analysis window $h(n)$ is given by (1) and $R = N/4$ then PR is ensured if $w(n) = \frac{2}{3}h(n)$. The total delay for such a filter bank is $N + R$, which for a 8ms window $h(n)$ gives a total delay of 10ms. Changing to the adaptive analysis does not alter the delay, as the adaptation is only used to calculate a modified gain function and if the synthesis window is calculated using (7), then applying (6) does not change the delay, since the synthesis window perfectly cancels the phase shift of (6). Therefore the total delay of the adaptive analysis is the same as for a standard STFT.

7. EXPERIMENTAL RESULTS

5 male and 5 female speech samples sampled at 16 kHz from the TIMIT database have been mixed with speech-shaped noise at various SNRs and subjected to speech enhancement as in section 5 using i) the fixed-resolution STFT analysis (**FA**) and ii) the proposed adaptive analysis (**AA**) with $L = 1024$, $N = 128$, $R = N/4$, $h(n)$ and $w(n)$ are Hann windows scaled to give PR and the synthesis window for the adaptive analysis is found using (7). For reference, **AA** is also compared to iii) the adaptive analysis where the analysis filter is forced to always update using (5) (**LA**); i.e. non-stationarities are never detected, iv) the adaptive analysis where the synthesis window $w(n)$ is used instead of (6) (**AAW**) and v) the adaptive analysis synthesized using standard OA (**AAOA**). In all cases (5) and (6) are 1st-order AR filters with $a_1 = 0.92$ and $d_1 = 0.85$. To determine the time-frequency resolution, the test statistic from section 4 is used on the clean speech signal with $\lambda = 0.65$ and $V = 20$. The noise $\lambda_N(k, i)$ is assumed known and is precomputed as an average across all time frames. The speech output is evaluated using PESQ. PESQ is an objective measure of speech quality that has been used to evaluate speech enhancement algorithms and has a good correlation with subjective listening tests [11] [12].

The results from the experimental analysis is shown in Figure 4. The plot shows the improvement in PESQ compared to the noisy input signal $x(n)$. The proposed method (**AA**)

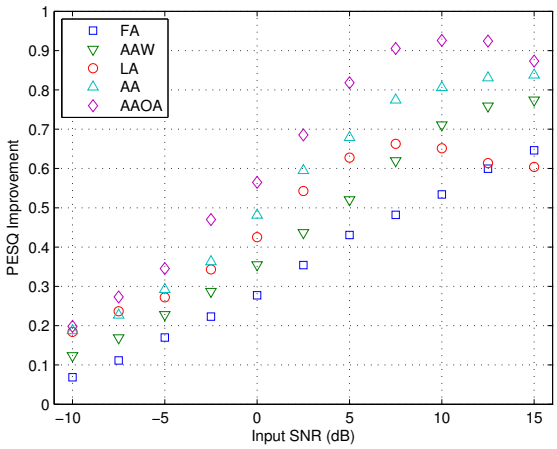

Fig. 4. *Improvements in PESQ compared to the noisy signal.*

consistently exhibits an improvement in PESQ compared to the fixed-resolution method (**FA**). For positive SNRs the improvement is above 0.2 and even for negative SNRs, **AA** is better than **FA**. Further improvement is achieved by relaxing the time delay constraint and synthesizing $y(n)$ using OA in **AAOA**, which is because OA allows the full filter ringing of the gain function. Using the normal synthesis window $w(n)$, it is seen that **AAW** performs significantly worse than **AA** which indicates the successful reconstruction of the causal part of the filter ringing using (6). The lower values for **LA** indicate that the adaptive analysis indeed improves the speech enhancement compared to just using a longer analysis frame. It has been verified through informal listening tests that the proposed processing scheme results in reduced musical noise and a perceptually clearer signal compared to the alternative method.

8. CONCLUSION

An adaptive time-frequency analysis scheme has been proposed, which allows independent time-frequency resolutions for each frequency bin. The proposed method perfectly reconstructs the signal, has a low computational complexity and low time delay which makes it suitable for real-time implementation. It has been shown that the proposed method results in superior speech quality compared to a fixed resolution STFT processing scheme with equivalent time delay. In the used example, the time delay was 10 ms. If larger delays can be accepted, it has also shown that the adaptive analysis can be combined with standard Overlap-Add synthesis, resulting in a slight improvement of the enhanced signal in terms of PESQ.

9. REFERENCES

[1] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*, Piscataway, NJ: IEEE Press, 2000.

[2] J. Agnew and J. M. Thornton, "Just noticeable and objectionable group delays in digital hearing aids," *Journal of the American Academy of Audiology*, vol. 11, no. 6, pp. 330–336, 2000.

[3] D. Rudoy, P. Basu, T.F. Quatieri, B. Dunn, and P.J. Wolfe, "Adaptive short-time analysis-synthesis for speech enhancement," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 2008, pp. 4905–4908.

[4] Dirk Mauler and Rainer Martin, "Improved reproduction of stops in noise reduction systems with adaptive windows and nonstationarity detection," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, pp. 2:1–2:17, 2009.

[5] Dirk Mauler and Rainer Martin, "A low delay, variable resolution, perfect reconstruction spectral analysis-synthesis system for speech enhancement," *15th European Signal Processing Conference (EUSIPCO 2007), Poznan, Poland*, pp. 222–226, 2007.

[6] J.G. Proakis and D.G. Manolakis, *Digital Signal Processing*, Prentice-Hall, Upper Saddle River, New Jersey, 1996.

[7] R.E. Crochiere, *Multirate Digital Signal Processing*, Prentice-Hall, 1983.

[8] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 33, no. 2, pp. 443–445, 1985.

[9] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 32, no. 6, pp. 1109–1121, 1984.

[10] D.P. Skinner, "Pruning the decimation in-time fft algorithm," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 24, no. 2, pp. 193–194, 1976.

[11] Yi Hu and P.C. Loizou, "Evaluation of objective quality measures for speech enhancement," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 1, pp. 229–238, 2008.

[12] Thomas Rohdenburg, Volker Hohmann, and Birger Kollmeier, "Objective perceptual quality measures for the evaluation of noise reduction schemes," *9th International Workshop on Acoustic Echo and Noise Control*, pp. 169–172, 2005.