ADAPTIVE SYSTEM IDENTIFICATION USING TIME-VARYING FOURIER TRANSFORM

Hadas Benisty, Yekutiel Avargel and Israel Cohen
Department of Electrical Engineering, Technion - Israel Institute of Technology
Technion City, Haifa 32000, Israel
{hadasbe@tx, kutiav@tx, icohen@ee}.technion.ac.il

Abstract

In this paper, we introduce a time-varying short-time Fourier transform (TV-STFT) for representing discrete signals. We derive an explicit condition for perfect reconstruction using time-varying analysis and synthesis windows. Based on the derived representation, we propose an adaptive algorithm that controls the length of the analysis window to achieve a lower mean-square error (MSE) at each iteration. When compared to the conventional multiplicative transfer function approach with a fixed length analysis window, the resulting algorithm achieves faster convergence without compromising for higher steady state MSE. Experimental results demonstrate the effectiveness of the proposed approach.

1. Introduction

System identification in the short-time Fourier transform (STFT) domain of linear time-invariant (LTI) systems is widely used in speech processing applications [1]-[6]. System identification in the STFT domain generally requires crossband filters between the subbands [4]. Nonetheless, when the STFT's analysis window is long and smooth in comparison to the system's impulse response, a single multiplicative term can be employed in each frequency bin. This approach is generally referred to as the multiplicative transfer function (MTF) approximation [6]. Since in real-world applications, the system to be identified is time-varying, adaptive estimation algorithms are required, using a finite amount of samples. In order to achieve a more accurate approximation and low steady-state error of the MTF approach, the analysis window should be as long and smooth as possible. However, a long window yields slower convergence than a shorter one. As a consequence, when choosing the analysis window length this trade-off has to be considered.

In this paper, a time-varying STFT (TV-STFT) is defined using a time-varying window length. In addition, a time-varying inverse STFT (TV-ISTFT) is derived, and a perfect reconstruction is achieved by enforcing a generalized form of the completeness condition. Based on the MTF approximation, an adaptive scheme for system identification, which utilizes the dynamic nature of the TV-STFT domain, is presented. Using a time-varying window length, the resulting algorithm achieves a relatively low steady state error without degrading its convergence rate.

This paper is organized as follows. In Section 2, the TV-STFT and TV-ISTFT are defined, and a perfect reconstruction condition is obtained. In Section 3, an adaptive scheme for system identification is presented in the TV-STFT domain. In Section 4, experimental results demonstrate the effectiveness of the proposed approach.

2. Time-Varying STFT

The STFT of a signal $x(n)$ is defined by [7]:

$$x_{p,k} = \sum_{m=-\infty}^{\infty} x(m)\tilde{\psi}_N(m-pL)e^{-j\frac{2\pi}{N}k(m-pL)} \quad (1)$$

where $\tilde{\psi}_N(n)$ is an analysis window of length N, L is the decimation factor, p is the frame index and k is the frequency-bin index. The ISTFT of $x_{p,k}$ can be computed using the well known overlap and add formulation:

$$ISTFT\{x_{p,k}\} = \frac{1}{N}\sum_{p=-\infty}^{\infty}\sum_{k=0}^{N-1} x_{p,k}\psi_N(n-pL)e^{j\frac{2\pi}{N}k(n-pL)} \quad (2)$$

where $\psi_N(n)$ is a synthesis window of length N. Substituting (1) into (2) and enforcing perfect reconstruction of $x(n)$ leads to the completeness condition [7]:

$$\sum_{p=-\infty}^{\infty} \psi_N(n-pL)\tilde{\psi}_N(n-pL) = 1 \quad \forall n. \quad (3)$$

Let us define a time-varying STFT (TV-STFT) of $x(n)$ by:

$$\breve{x}_{t,k} = \sum_{m=-\infty}^{\infty} x(m)\tilde{\psi}_{N(t)}(m-t)e^{-j\frac{2\pi}{N(t)}k(m-t)} \quad (4)$$

where $\tilde{\psi}_{N(t)}$, is a time-varying analysis window, whose length $N(t)$ is a piece-wise constant function:

$$N(t) = N_\nu \quad t_{\nu-1} < t \le t_\nu \quad \nu = 1,2,..,V \quad (5)$$

and V is the number of discontinuity points.
Let L_ν denote a time varying decimation factor such that a fixed overlap is preserved:

$$N_\nu / L_\nu = const \quad \forall \nu . \quad (6)$$

The windows are centered at the time instants $t \in \Gamma$ defined by:

$$\Gamma = \{ t \,|\, t = t_{\nu-1} + rL_\nu \quad , t_{\nu-1} \le t < t_\nu \\ t_0 = -\infty \quad , 0 \le r \le (t_\nu - t_{\nu-1}) / L_\nu \}. \quad (7)$$

Similarly to (2), the TV-ISTFT can be expreseed as:

$$\text{TV}-\text{ISTFT}\{\breve{x}_{t,k}\} = \frac{1}{N(t)} \sum_{t\in\Gamma} \sum_{k=0}^{N(t)-1} \breve{x}_{t,k} \psi_{N(t)}(n-t) e^{j\frac{2\pi}{N(t)}k(n-t)}. \quad (8)$$

Finally, substituting (4) into (8) and enforcing perfect reconstruction, yields the generalized form of the completeness condition:

$$\sum_{t\in\Gamma} \psi_{N(t)}(n-t)\tilde{\psi}_{N(t)}(n-t) = 1 \quad \forall n. \quad (9)$$

In order to demonstrate the structure of the analysis and synthesis windowing process, a simple case is presented, in which the analysis-window length is varied once:

$$N(t) = \begin{cases} N_1, \text{ for } t < t_1, \\ N_2, \text{ for } t \ge t_1 \end{cases}. \quad (10)$$

The decimation factor is chosen so that the overlap between consecutive windows is retained, i.e.,

$$L(t) = \begin{cases} L_1, \text{ for } t < t_1, \\ L_2, \text{ for } t \ge t_1, \end{cases} \text{ where } \frac{N_1}{L_1} = \frac{N_2}{L_2}, \quad (11)$$

and the analysis windows centers are:

$$\Gamma = \{ t \,|\, t = rL_1 \quad r \le t_1 / L_1 \\ or \quad t = t_1 + rL_2 \quad r > 0 \}. \quad (12)$$

Note that the continuity of the window functions should be preserved when switching from $\tilde{\psi}_{N_1}(n)$ to $\tilde{\psi}_{N_2}(n)$. Then, since the number of windows having non-zero values at the transition point t_1 is:

$$M = 2\left\lfloor \left[\frac{N_1}{L_1} \right] - 1 \right\rfloor + 1, \quad (13)$$

the continuity can be preserved by interlacing $\tilde{\psi}_{N_1}(n)$ and $\tilde{\psi}_{N_2}(n)$:

$$\tilde{\psi}_{N_1 \to N_2}(n-t) = \begin{cases} \tilde{\psi}_{N_1}(n-t) & n < t_1 \\ \tilde{\psi}_{N_2}(n-t) & n \ge t_1 \end{cases} \quad (14)$$
$$t_1 - \lfloor M/2 \rfloor L_1 \le t \le t_1 + \lfloor M/2 \rfloor L_2.$$

Specifically, for $t < t_1 - \lfloor M/2 \rfloor L_1$, the windowing process is emploid using $\{\tilde{\psi}_{N_1}(n), L_1\}$, and for $t > t_1 + \lfloor M/2 \rfloor L_2$, the windowing process will continue using $\{\tilde{\psi}_{N_2}(n), L_2\}$. A simple example for this windowing process is demonstrated in Figure 1, in which 75% overlap leads to $M = 3$ interlaced windows around the transition point $t_1 = 512$.

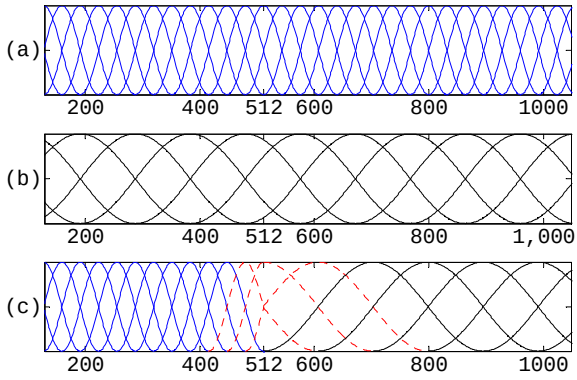

Figure 1. Time invariant windowing using a Hamming window (a) $N = 128, L = 32$ **; (b)** $N = 384, L = 96$ **; (c) a time varying windowing process switching from** $N_1 = 128, L_1 = 32$ **(blue) to** $N_2 = 384, L_2 = 96$ **(black), at the transient point** $t_1 = 512$ **using three interlaced windows (red dot).**

The synthesis windows should sustain the generalized form of the completeness condition (equation (9)):

$$\sum_{t\in\Gamma} \psi_{N_1}(n-t)\tilde{\psi}_{N_1}(n-t) = 1 \quad n < t_1 \\ \sum_{t\in\Gamma} \psi_{N_2}(n-t)\tilde{\psi}_{N_2}(n-t) = 1 \quad n \ge t_1 \quad (15)$$

Recall that the analysis windows $\tilde{\psi}_{N_1}(n), \tilde{\psi}_{N_2}(n)$ and their synthesis matches $\psi_{N_1}(n), \psi_{N_2}(n)$ satisfy (3). Since equation (3) holds for all n, it also holds particularly for $n < t_1$, which means that $\tilde{\psi}_{N_1}(n)$ and $\psi_{N_1}(n)$ sustain equation (9) in this time interval, so $\psi_{N_1}(n)$ can be used as a synthesis window for $n < t_1$. A similar claim can be made regarding $\tilde{\psi}_{N_2}(n)$ and $\psi_{N_2}(n)$, for $n \ge t_1$. Altogether, the windowing process of the TV-ISTFT divides the time domain into three parts: before, during and after the transition, that is,

$$\psi_{N(t)}(n-t) = \\ = \begin{cases} \psi_{N_1}(n-t) & t < t_1 - \lfloor M/2 \rfloor L_1 \\ \psi_{N_1 \to N_2}(n-t) & t_1 - \lfloor M/2 \rfloor L_1 \le t \le t_1 + \lfloor M/2 \rfloor L_2 \\ \psi_{N_2}(n-t) & t > t_1 + \lfloor M/2 \rfloor L_2 \end{cases} \quad (16)$$

where

$$\psi_{N_1 \to N_2}(n-t) = \begin{cases} \psi_{N_1}(n-t) & n < t_1 \\ \psi_{N_2}(n-t) & n \ge t_1 \end{cases}. \quad (17)$$

This formulation can be extended for any countable number of variation points, as long as the interlacing structure of the analysis and synthesis

windows is preserved. That is, at the non-transient frames (i.e., $t_{v-1} + \lfloor M/2 \rfloor L_v < t < t_v - \lfloor M/2 \rfloor L_v$) the TV-STFT and its inverse are computed using $\{ \breve{\psi}_{N_v}(n), \psi_{N_v}(n), L_v \}$. At each transition point t_v, M interlaced windows [see (13)] should be used to switch from N_v to N_{v+1}. The following frames should be analyzed and synthesized using $\{ \breve{\psi}_{N_{v+1}}(n), \psi_{N_v+1}(n), L_{v+1} \}$, up to the next transition point.

3. Adaptive System Identification Using TV-STFT

In this section we consider the problem of system identification using the previously defined TV-STFT. Let $x(n)$ and $y(n)$ be the input and output of an unknown LTI system, represented by its impulse response $h(n)$, with additive noise $\xi(n)$, i.e.,

$$y(n) = x(n) * h(n) + \xi(n) = \\ = d(n) + \xi(n) \tag{18}$$

The goal of system identification schemes is to estimate $h(n)$ based on samples of the input and output signals. In many practical cases (such as [4]-[6]), this estimation is performed adaptively in the STFT domain to achieve both fast convergence and low computational cost. Applying the STFT on (18) produces:

$$y_{p,k} = d_{p,k} + \xi_{p,k} \tag{19}$$

Assuming that the analysis window $\tilde{\psi}_N(n)$ is much longer and smoother in comparison to the impulse responce $h(n)$, the MTF approximation can be applied [6]:

$$d_{p,k} \approx x_{p,k} \cdot h_k^{(N)} \tag{20}$$

where

$$h_k^{(N)} \triangleq \sum_{m=-\infty}^{\infty} h(m) e^{-j\frac{2\pi}{N}km} \tag{21}$$

The effect of the analysis window length on the estimation error was explored in [6], and a trade-off between the convergence rate and the steady-state value of the estimation error was presented. In order to avoid this trade-off, the TV-STFT can be employed. Specifically, at the beginning, a short window should be used in order to achieve fast convergence. Then, as the adaptation process proceeds, the algorithm should gradually increase the window length to produce a lower steady state MSE. Using the TV-STFT with the windowing process defined in (5)-(7) and the MTF approximation, (18) becomes:

$$\breve{y}_{t_\ell,k} = \breve{d}_{t_\ell,k} + \breve{\xi}_{t_\ell,k} \approx \breve{x}_{t_\ell,k} \cdot h_k^{N(t_\ell)} + \breve{\xi}_{t_\ell,k} \tag{22}$$

where t_ℓ are the time instants $t \in \Gamma$ sorted in an ascending order so $t_{\ell-1} < t_\ell < t_{\ell+1}$. At the non-transient frames, the adaptive estimation of the system using the normalized least-mean-square (NLMS) algorithm is identical to the one described in [5], using the appropriate window lengths and decimation factors. Denote: $N(t_\ell) = N_v$ and $L(t_\ell) = L_v$, the NLMS equations are:

$$\hat{h}_k^{N_v}(t_{\ell+1}) = \hat{h}_k^{N_v}(t_\ell) + \mu \breve{e}_{t_\ell,k} \frac{\breve{x}_{t_\ell,k}^*}{\left|\breve{x}_{t_\ell,k}\right|^2} \\ \breve{e}_{t_\ell,k} = \breve{y}_{t_\ell,k} - \hat{\breve{d}}_{t_\ell,k} = \breve{y}_{t_\ell,k} - \breve{x}_{t_\ell,k} \cdot \hat{h}_k^{N_v}(t_\ell) \\ t_{v-1} + \lfloor M/2 \rfloor L_v < t_\ell < t_v - \lfloor M/2 \rfloor L_v \tag{23}$$

During the M transient frames, $\hat{h}_k^{N(t)}(t)$ is evaluated using interlaced windows defined in equation (14), which means that its frequency–bins are not evenly spaced:

$$h_k^{N(t_\ell)}(t_\ell) = \\ = \begin{cases} \sum_{m=-\infty}^{\infty} h(m) e^{-j\frac{2\pi}{N_v}km} & 0 \le k \le \left|t_v - t_l\right| + \frac{N_v}{2} \\ \sum_{m=-\infty}^{\infty} h(m) e^{-j\frac{2\pi}{N_{v+1}}km} & \left|t_v - t_\ell\right| + \frac{N_v}{2} < k < \frac{N_v + N_{v+1}}{2} + \left|t_v - t_\ell\right| \left(1 - \frac{N_{v+1}}{N_v}\right) \end{cases} \\ t_v - \lfloor M/2 \rfloor L_v \le t_\ell \le t_v + \lfloor M/2 \rfloor L_{v+1} \tag{24}$$

Equation (23) cannot be used directly in these frames, since the frequency-bins resolution and the amount of estimated system coefficients do not match. In order to resolve this problem while preserving the continuity of the system identification process, the coefficients related to mismatched frequency-bins should be adjusted before forwarded to the next frame. The estimated impulse response can be obtained by:

$$\hat{h}^{(t_\ell)}(n) = IDFT_{N_v}\left\{ h_k^{N_v}(t_\ell) \right\}. \tag{25}$$

Switching to the next frequency–bins resolution is performed by employing N_{v+1} order DFT:

$$h_k^{N(t_{\ell+1})}(t_\ell) = DFT_{N_{v+1}}\left\{ \hat{h}^{(t_\ell)}(n) \right\}. \tag{26}$$

To conclude, the required modification by the adaptive process is back and forward DFT according to the current and next window lengths. This adjustment ensures the continuity of the system identification process.

4. Experimental Results

In this section, two simulations are presented. The first one uses white Gaussian noise as input in order to validate the analysis described above. The second simulation shows the applicability to acoustic echo cancellation by using a speech signal. Hamming window was taken as the analysis window with 50% overlap, and the synthesis window was its bi-orthogonal window. The noise signal $\xi(n)$ is white, zero mean, and Gaussian with SNR of 30dB:

$$SNR = 30dB = 10\log\left(\frac{\sigma_x^2}{\sigma_\xi^2}\right) \qquad (27)$$

where σ_x^2 is the variance of the input signal $x(n)$, and σ_ξ^2 is the variance of the noise signal $\xi(n)$.

The system's impulse response was modeled as a white Gaussian noise $\beta(n)$ multiplied by an exponential decay:

$$h(n) = w(n)\beta(n)e^{-\alpha n} \qquad (28)$$

And $w(n)$ was taken as a rectangular window:

$$w(n) = \begin{cases} 1 & 0 \le n < N_h \\ 0 & otherwise \end{cases} \qquad (29)$$

The decay exponent that was used: $\alpha = 0.03$. The decimation factor was taken so a fixed overlap of 50% was sustained.

4.1. White Gaussian Noise

The input signal $x(n)$ was taken as white, zero mean, unit variance Gaussian noise, uncorrelated with $\xi(n)$. The performance of the system identification was evaluated during 24,000 samples using the normalized mean square error (MSE) in the time domain:

$$\varepsilon(n) = \frac{E\left[\left(d(n) - \hat{d}(n)\right)^2\right]}{E\left[d^2(n)\right]} \qquad (30)$$

where $\hat{d}(n)$ is the TV-ISTFT of $\hat{\hat{d}}_{t,k}$.

The mean was taken over 500 uncorrelated experiments. The impulse response was 16 samples long and was changed when convergence has been achieved – after 9,200 samples. Figure 2 shows the window lengths function $N(n)$ that was used by the TV-STFT simulation. The shorter windows were used both at the beginning of the estimation and when the impulse response was changed. Figure 3 shows the normalized error for various window lengths using the conventional STFT and the TV-STFT. A smoothed version of these curves is presented in Figure 4.

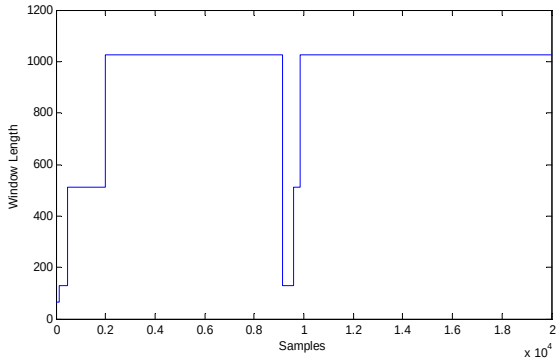

Figure 2. Window length function N(n) used for the TV-STFT in the white Gaussian noise simulation.

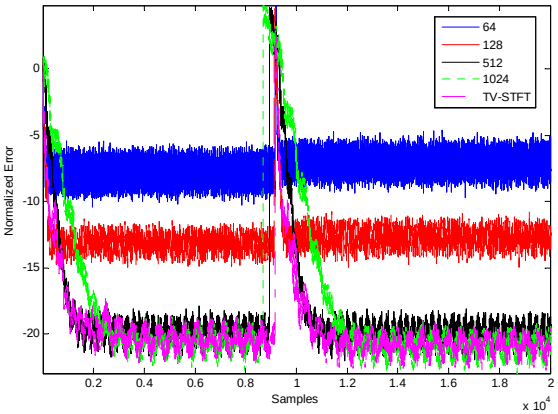

Figure 3. Normalized MSE curves using fixed window lengths (64, 128, 512, 1024 samples), and TV-STFT. The impulse response was changed after 9200 samples

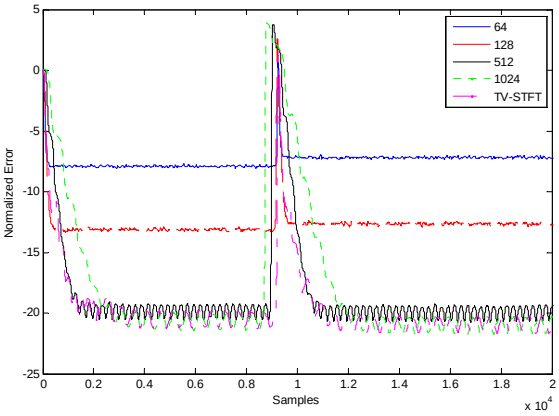

Figure 4. Smoothed (using Hamming window) noramlized MSE curves using fixed window lengths (64, 128, 512, 1024 samples), and TV_STFT. The implulse response was changed after 9200 samples.

Two methods for updating the estimated system coefficients at the transient frame were compared: One using IDFT and zero padded DFT as described above, and the second - by using cubic interpolation. Since both methods produced very similar results, the cubic interpolation was chosen in order to ease the complexity load.

The trade-off between the convergence rate and the steady-state value is well demonstrated: the shortest window (64 samples) converges very quickly, but to a relatively high value. The longest window (1024 samples) achieves the lowest steady-state value, but with a very long convergence duration. This attribute is shown both at the beginning of the estimation and when the impulse response is changed.

It is clear that the error obtained by the TV-STFT produced the lowest error during the whole simulation.

Performing STFT and ISTFT is equivalent to using analysis and synthesis filter bank [7]. The output of

a synthesis filter bank is generally cyclo-stationary [8], so this is the reason for the periodic nature of the estimation error in Figure 3. In practical applications, there is no averaging over the estimation error, so this phenomenon is not disturbing.

4.2. Acoustic Echo-Cancellation

In this simulation a practical usage for system identification using TV-STFT is demonstrated. Acoustic echo cancellation (AEC) can be modeled using equation (18): the input signal $x(n)$ is the far-end signal coming out of a loudspeaker. The echo path of the room, from the loudspeaker to an adjacent microphone, is modeled by $h(n)$. $\xi(n)$ is an additive white Gaussian noise, and the output signal $y(n)$ is the recording microphone signal. In the simulation, the input signal $x(n)$ was a speech signal sampled at 16kHz during 4 seconds. A 64 samples long impulse response was changed after 2 seconds.

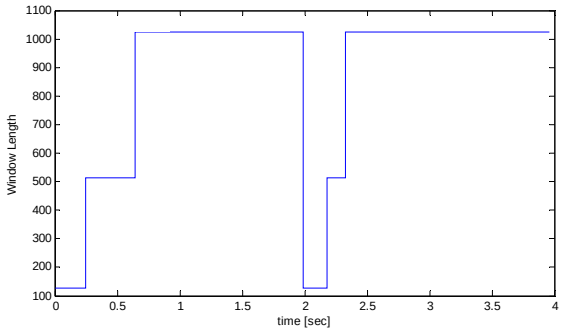

Figure 5. Window length function N(n) used for the TV-STFT in the AEC simulation

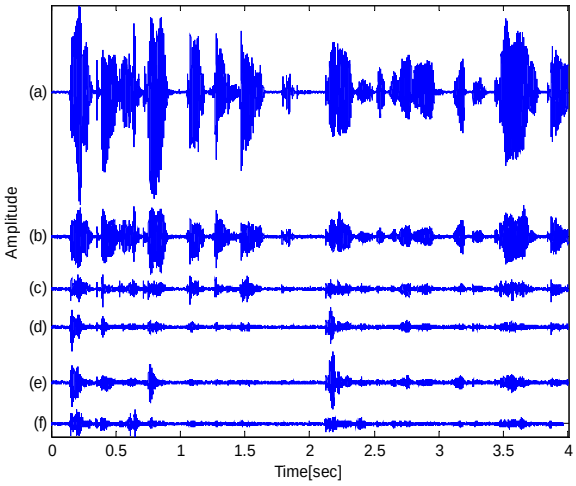

Figure 6: Speech waveforms and error signals. The impulse response was changed after 2secs. (a) Far-end Signal; (b) near-end signal; (c)-(e) error signals for N=128, N=512, N=1024 respectively; (f) TV-ISTFT using N=[128 512 1024] starting at t=0secs and t=2secs.

Figure 6 (a)-(b) shows the far-end speaker and the microphone signal respectively. The error signals Figure 6 (c)-(f) were calculated by applying TV-ISTFT on the error signals obtained by (23). A fixed window length of 128, 512, 1024 samples was used in Figure 6 (c)-(e) respectively. The TV-STFT was used in Figure 6 (f) with window length function $N(n)$ as depicted in Figure 2.

The fixed window estimation error suffers the same trade-off as before, whereas the TV-STFT achieves low error all through the estimation process.

5. Conclusions

A time-varying STFT (TV-STFT) was introduced using a time-varying window length, and its inverse transform was calculated. A perfect reconstruction condition was explored and formalized. An adaptive scheme for system identification utilizing the dynamic nature of the TV-STFT domain was presented. Performing adaptive system identification using conventional STFT leads to a trade-off between the convergence rate and steady state value of the estimation error. Using a time varying window function resolves this trade-off and produces lower estimation error, and faster convergence rate. Experimental results confirmed that the TV-STFT approach outperforms the conventional STFT approach.

Further work can be done concerning the time varying window function. In particular, derivation of the optimal values and the transition points online while considering the current stage of the estimation process and the instantaneous SNR conditions.

6. References

[1] C. Faller and J. Chen, "Suppressing Acoustic Echo in a Spectral Envelope Space", *IEEE Trans. Speech, Audio Process.*, vol. 13. No. 5, pp. 1048-1062, Sep. 2005.

[2] I. Cohen, "Relative transfer function identification using speech signals", *IEEE Trans. Speech Audio Process. (Special Issue on Multi-channel Signal Processing for Audio and Acoustics Applications)*, vol. 12, no. 5, pp. 451-459, Sep. 2004.

[3] I. Cohen, "Identification of Speech Source Coupling Between Sensors in Reverberant Noisy Environments," *IEEE Signal Process. Lett.*, vol. 11, no. 7, pp. 613-616, Jul. 2007.

[4] Y. Avargel and I. Cohen, "System Identification in the Short-Time Fourier Transform Domain With Crossband Filtering," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1305-1319, Apr. 2007.

[5] Y. Avargel and I. Cohen, “Adaptive system identification in the short-time Fourier transform domain using cross-multiplicative transfer function approximation,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 162-173, Jan 2008.
[6] Y. Avargel and I. Cohen, “On multiplicative transfer function approximation in the short-time Fourier transform domain,” *IEEE Signal Process. Lett.*, vol. 14, no. 5, pp. 337-340, May 2007.
[7] M. R. Portnoff, “Time-frequency representation of digital signals and systems based on short-time Fourier analysis,” *IEEE Trans. Signal Processing*, vol. 28, no. 2, pp. 55–69, Feb. 1980.
[8] Ohno and H. Sakai, "Optimization of' filter banks using cyclostationary spectral analysis," *IEEE Trans. Signal Process.*, vol. 44, no. 11, pp. 2718-2725, Nov. 1996.