2. SHORT-TIME FOURIER TRANSFORM

CONTENTS

SHORT-TIME FOURIER TRANSFORM (STFT) ANALYSIS
SHORT-TIME FOURIER TRANSFORM (STFT) SYNTHESIS
STFT MAGNITUDE ANALYSIS/SYNTHESIS
MODIFIED STFT/STFTM SYNTHESIS
SPEECH APPLICATIONS

2.1 Short-Time Fourier Transform (STFT) Analysis

- Given time-series \( x[n] \), the STFT at time \( n \) is given as:

\[
X(n, \omega) = \sum_{m=-\infty}^{\infty} x[m]w[n - m]e^{-j\omega m},
\]

where \( w[n] \) is the analysis window, which is assumed to be non-zero only in the interval \([0, N_w - 1]\).

- The discrete STFT is obtained by sampling \( X(n, \omega) \) over the unit circle:

\[
X(n, k) = X(n, \omega)\bigg|_{\omega = \frac{2\pi k}{N}} = \sum_{m=-\infty}^{\infty} x[m]w[n - m]e^{-j\frac{2\pi}{N} km},
\]

where \( N \) is the frequency sampling factor and \( 2\pi/N \) is the frequency sampling interval.
An alternate (filtering) view of the discrete STFT is:

\[
X(n, \omega_0) = \sum_{m=-\infty}^{\infty} (x[m]e^{-j\omega_0 m})w[n-m] = (x[n]e^{-j\omega_0 n}) * w[n].
\]  

(2.3)

That is, the signal \(x[n]\) is first modulated with \(e^{-j\omega_0 n}\), and then passed through a filter with impulsive response \(w[n]\).

Fig. 7.3 Filtering view of STFT analysis at frequency \(\omega_0\): (a) block diagram of complex exponential modulation followed by a lowpass filter; (b) operations in the frequency domain.
An equivalent representation of (2.3) is:

\[ X(n, \omega_0) = e^{-j\omega_0 n} (x[n] * w[n] e^{j\omega_0 n}). \]  \hspace{1cm} (2.4)

That is, the sequence \( x[n] \) is first passed through the filter \( w[n] \) with a linear phase factor. The output is then modulated by \( e^{-j\omega_0 n} \).

Fig. 7.4 Alternative filtering view of STFT analysis at frequency \( \omega_0 \): (a) block diagram of bandpass filtering followed by complex exponential modulation; (b) operations in the frequency domain.
Analysis and synthesis with the discrete STFT

Fig. 7.5 the filtering view of analysis and synthesis with the discrete STFT: (a) the discrete STFT analysis as the output of a filter bank consisting of bandpass filters; (b) filter bank summation procedure for signal synthesis from the discrete STFT.
2.1.3 Time-Frequency Resolution Tradeoffs

The STFT can be also written as

\[
X(n, \omega) = \frac{1}{2\pi} \int_{-\pi}^{\pi} W(\theta) e^{j\theta n} X(\omega + \theta) d\theta,
\]

(2.5)

\(X(\omega)\) is the Fourier transform of \(x[m]\) and \(W(-\omega)e^{j\omega n}\) as the Fourier transform of \(w[n-m]\) with respect to \(m\).

The size of \(w[n]\) affects the time-frequency resolution of STFT:

<table>
<thead>
<tr>
<th>Window size of (w[n])</th>
<th>Bandwidth of (W(\omega))</th>
<th>Time resolution</th>
<th>Frequency resolution</th>
<th>Good for:</th>
</tr>
</thead>
<tbody>
<tr>
<td>long</td>
<td>narrow</td>
<td>bad</td>
<td>good</td>
<td>sinusoidal components, (harmonic)</td>
</tr>
<tr>
<td>short</td>
<td>wide</td>
<td>good</td>
<td>bad</td>
<td>fast time-varying components, (rapid conversational speech)</td>
</tr>
</tbody>
</table>

A fundamental problem of STFT and other time-frequency analysis techniques is the selection of the windows to achieve a good tradeoff between time and frequency resolution.
Example:

Fig. 7.8 Effect of the length of the analysis window on the discrete Fourier Transform of linearity frequency-modulated sinusoid of 25 ms whose frequency decreases from 1250 Hz to 625 Hz. The Fourier transform uses a rectangular window centered at 12.5 ms, as illustrated in (a). Transform are shown for different window lengths: (b) 5 ms [solid in (a)]; (c) 10 ms [dashed in (a)]; (d) 20 ms [dotted in (a)].
2.2 Short-Time Synthesis

- The STFT is highly redundant and it can be inversed.

- For each \( n \), we take the inverse Fourier transform of \( X(n, \omega) \) from the STFT. We then obtain \( f_n[m] = x[m]w[n - m] \). Evaluating \( f_n[m] \) at \( m = n \), we obtain \( x[n]w[0] \). Assuming \( w[0] \neq 0 \), we have \( x[n] = f_n[n]/w[0] \). This gives to the following synthesis equation:

\[
x[n] = \frac{1}{2\pi w[0]} \int_{-\pi}^{\pi} X(n, \omega)e^{j\omega n} d\omega. \tag{2.6}
\]

- To reduce the computational complexity, the STFT is not computed at every time sample, but rather at a certain time-decimation rate. In some cases, the discrete STFT may not be invertible, i.e. there are certain constraints on the frequency-sampling and time-decimation rates.
Two common methods for STFT synthesis are the Filter Bank Summation (FBS) method and the Overlap-Add (OLA) method. Due to time limitation, the details are omitted.

**FBS Method**

\[
y[n] = \left[ \frac{1}{Nw[0]} \right] \sum_{k=0}^{N-1} X(n, k) e^{\frac{-2\pi i kn}{N}}
\]

Adding Frequency Components For Each \( n \)

FBS Constraint: \( \sum_{k=0}^{N-1} W(\omega - \frac{2\pi}{N} k) = Nw[0] \)

For \( N_w < N \) \( \rightarrow y[n] = x[n] \)

**OLA Method**

\[
y[n] = \left[ \frac{L}{W(0)} \right] \sum_{p = -\infty}^{\infty} x[pL - n] w[n]
\]

Adding Time Components For Each \( n \)

OLA Constraint: \( \sum_{p = -\infty}^{\infty} w[pL - n] = \frac{W(0)}{L} \)

For \( \omega_c < \frac{2\pi}{L} \) \( \rightarrow y[n] = x[n] \)

**Fig. 7.13 Duality between the FBS and OLA constraints.**
2.3 Short-Time Fourier Transform Magnitude

- The squared STFT magnitude is called the spectrogram or short time Fourier transform magnitude (STFTM).
- Although the phase information is removed in STFT, many signals can still be uniquely synthesized from STFTM.
- The relation between STFTM and the short-time autocorrelation $r[n,m]$ is:

$$r[n,m] = \frac{1}{2\pi} \int_{-\pi}^{\pi} |X(n,\omega)|^2 e^{j\omega m} d\omega,$$

(2.7)

$$|X(n,\omega)|^2 = \sum_{m=-\infty}^{\infty} r[n,m] e^{-j\omega m},$$

(2.8)

where $m$ is the autocorrelation “lag”, $r[n,m] = f_n[m]*f_n[-m]$ and $f_n[m]x[m]w[m-n]$. 
2.4 Signal Estimation from Modified STFT and STFTM

Assume that the STFT $X(n, \omega)$ has been modified by a function $H(n, \omega)$ to:

$$Y(n, \omega) = X(n, \omega)H(n, \omega).$$ (2.9)

This corresponds to a new short-time segment $g_n[m] = f_n[m] \ast h[n, m]$, where $h[n, m]$ can be seen as a time-varying system impulse response.

Heuristic synthesis methods: Signal estimation from the modified STFT is performed by the FBA and OLA synthesis, and estimation from the modified STFTM is performed by the sequential extrapolation method.

In the least-squared-Error Signal estimation approach, we minimize the mean-squared-error (MSE) between $Y(n, \omega)$ and the STFT $X_e(n, \omega)$ of the signal $x_e(n)$ to be estimated.

$$X_e(n, \omega) = \sum_{m=-\infty}^{\infty} w[n - m] x_e[m] e^{-j\omega m}. \quad (2.10)$$
The optimization results in the following solution for $x_e(n)$:

$$x_e[n] = \frac{\sum_{m=-\infty}^{\infty} w[m-n] f_m[n]}{\sum_{m=-\infty}^{\infty} w^2[m-n]}, \quad (2.11)$$

where $f_m[n]$ is the inverse Fourier transform of the short-time segment at time $m$, corresponding to the modified STFT, $Y(m, \omega)$. 


2.5 Time-Scale Modification and Enhancement of Speech

Time-Scale Modification – articulation rate change

![Diagram showing time-scale modification](image)

Fig. 7.24 Model for uniform change in articulation rate. The articulation rate is increased by a factor of two, so the source and speech waveforms are reduced to half their original length.

- Note the speech waveform is not simply compressed or expanded. The waveform maintains its shape, by decreasing or increasing the number of pitch periods during voicing and making unvoiced sounds last shorter or longer.
The STFT of the speech signal

\[ X(n, \omega) \approx U(n, \omega)H(n, \omega), \tag{2.12} \]

where \( U(n, \omega) \) and \( H(n, \omega) \) are the STFTs of the source and the time-varying vocal tract system, respectively. For a time scale modification factor of \( \rho = 1 / L \), we first form \( X(nL, \omega) \approx U(nL, \omega)H(nL, \omega) \) and then discarding or repeating spectral slices to form a new STFT \( Y(nL, \omega) \). The source (pitch and voicing state) and system (vocal tract spectrum) characterization are roughly preserved.

Fig. 7.25 Time-scale compression by the Fairbanks technique. Alternating short-time segments are discarded and the remaining ones are overlapped and added.
In the synchronized overlap add (SOLA) method, successive frames to be overlapped are cross-correlated. The peak of the cross-correlation function gives the time shift to make the two overlapping frames synchronize and thus add coherently.

Fig. 7.27 The problem of “pitch synchrony” in the OLA and LSE synthesis approaches.
Fig. 7.29 Time-scale modification with iterative LSE estimation: (a) original STFTM with $L=128$; (b) modified STFTM with $L=64$; (c) STFTM of LSE estimate.
Assume that the noise-corrupted signal $y[n]$ is given by

$$y[n] = x[n] + b[n], \quad (2.13)$$

where $x[n]$ is the clean speech signal and $b[n]$ is the additive noise sequence. The approach is to modify each spectral slice of the STFT of $y[n]$ to remove noise, and from the resulting modified STFT, construct an enhanced speech waveform. Certain model or prior information of the speech signals and noise is required in order to determine which frequency components should be attenuated. E.g. speech is low-pass, while noise is highpass above a certain frequency.

Similarly, STFT- and STFTM-based synthesis methods can both be used to address the problem.