Lecture 4: Short-time Fourier Transform (STFT)

Mark Hasegawa-Johnson

ECE 537: Speech Processing

Outline

Review: DTFT of Windowed Periodic Signals

- The Fourier transform of an impulse train is an impulse train:

$$\sum_{m=-\infty}^{\infty} \delta[n-mT_0] \leftrightarrow \frac{2\pi}{T_0} \sum_{k=0}^{T_0-1} \delta\left(\omega - \frac{2\pi k}{T_0}\right)$$

- Voiced speech is created by sending a periodic voice source through a resonant pipe:

$$x[n] = w[n]\left(h[n] * \sum_{m=-\infty}^{\infty} \delta[n-mT_0]\right)$$

$$X(\omega) = \frac{1}{T_0} \sum_{k=0}^{T_0-1} H_k W\left(\omega - \frac{2\pi k}{T_0}\right)$$

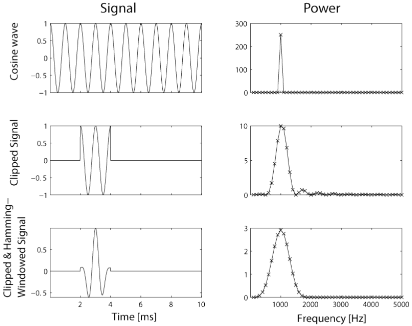
Outline

Definition of Short-Time Fourier Transform

The STFT is a function of **both time,** n, **and frequency**, ω. Time specifies the window offset from which the DTFT is computed.

$$X_n(\omega) \equiv \sum_{m=-\infty}^{\infty} w[n-m]x[m]e^{-j\omega m}$$

1. Clipping the signal (rectangular window) blurs the spectrum, and adds sidelobes
2. Windowing (e.g., Hamming) increases the blur, but decreases the sidelobes

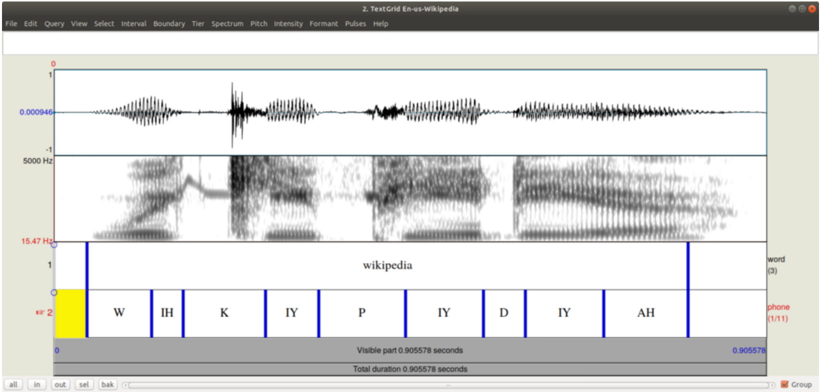


Definition of Spectrogram

$$S(n,\omega) = 20\log_{10}|X_n(\omega)| = 10\log_{10}|X_n(\omega)|^2$$

- Spectrogram = log periodogram. Periodogram is a (noisy) estimate of the power spectrum, so spectrogram is useful for both random and deterministic signals.
- One "Bell" ($\log_{10}|X_n(\omega)|^2$) is the amount of power lost by a telephone signal in one mile of copper cable.
- Most humans can hear a **percentage difference** in loudness corresponding to about one decibel ($10\log_{10}|X_n(\omega)|^2$).

Example Spectrogram



GTK3, `https://commons.wikimedia.org/wiki/File:Waveform_spectrogram_and_transcription_of_wikipedia_in_praat.png`

Inverse STFT

- The inverse DTFT of $X_n(\omega)$ is $w[n-m]x[m]$:

$$w[n-m]x[m] = \frac{1}{2\pi}\int_0^{2\pi} X_n(\omega)e^{j\omega n}d\omega$$

- If there is no noise, either formula below is fine. If there is noise, the second formula is better, because averaging over multiple inverse transforms can reduce noise. Notice that the denominator is a constant, e.g., for a Hamming, $\sum_{n=-\infty}^{\infty} w[n-m] = 0.54N$, while for a rectangle, $\sum_{n=-\infty}^{\infty} w[n-m] = N$.

$$x[m] = \frac{w[n-m]x[m]}{w[n-m]} = \frac{\sum_{n=-\infty}^{\infty} w[n-m]x[m]}{\sum_{n=-\infty}^{\infty} w[n-m]}$$

Outline

Shifting each frame so it starts at $n=0$

- Suppose the window is nonzero for $m \in [0, N)$. Then the frame, $w[n-m]x[m]$, is nonzero for $m \in (n-N, n]$.
- `numpy.fft.fft` wants the frame to be nonzero for $m \in [0, N)$. We can do this with the variable substitution $m' = n-(N-1)+m$, resulting in

$$X_n(\omega) = e^{-j\omega(n-N+1)} \sum_{m'=0}^{N-1} w[(N-1)-m']x[n+m']e^{-j\omega m'}$$

- If the window is symmetric, $w[(N-1)-m'] = w[m']$.
- The phase shift corresponds to time shift by $n-N+1$.

DFT Implementation of the STFT

1. Chop out an N-sample frame, and move it to $m \in [0, N)$.
2. Window it.
3. Take its DFT.
4. When you compute the inverse STFT, remember to time-shift the frame back to its original place – this is equivalent to adding $\omega(n - N + 1)$ back to the phase.

Outline

Multiplication in Time = Convolution in Frequency

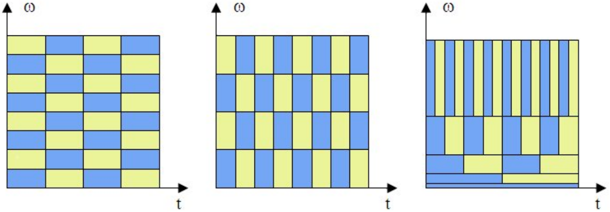
STFT is the transform of $w[n-m]x[m]$:

$$X_n(\omega) = \sum_{m=-\infty}^{\infty} w[n-m]w[m]e^{-j\omega m}$$

...which is equivalent to $X(\omega)$ convolved with a phase-shifted window:

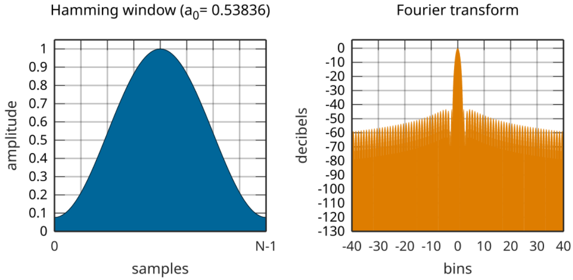
$$X_n(\omega) = \frac{1}{2\pi}\left(X(\omega) * e^{-j\omega n}W(-\omega)\right)$$

Point to ponder: The difference between $X_1(\omega)$ and $X_2(\omega)$ is that the infinite-length DTFT, $X(\omega)$, has been blurred by differently phase-shifted (i.e., differently time-shifted) window spectra.



- $w[n]$ can have a narrow mainlobe (good frequency resolution) if it's longer in the time domain (poor temporal resolution).
- A shorter $w[n]$ (good temporal resolution) has a wider mainlobe (worse frequency resolution).
- Wavelet methods use different windows at different frequencies.

- Hamming window is designed to have small sidelobes. The DTFT of a Hamming is **very close** to a triangle in frequency with halfwidth of $4\pi/N$ radians/sample – we can ignore the sidelobes, down to a noise floor of about -40dB.
- Rectangular window has a narrower center-lobe (halfwidth $2\pi/N$), but some spectral energy is splattered to other frequencies by the sidelobes.



Outline

Lowpass filter interpretation of STFT

The STFT $X_n(\hat{\omega})$ can be interpreted as a frequency-shift operator applied to $x[n]$ (shifting it by $\hat{\omega}m$ radians/sample, so that the part of the spectrum that used to be centered at $\omega = \hat{\omega}$ is now centered at $\omega = 0$), followed by a lowpass filter (using $w[n]$ as the lowpass filter):

$$X_n(\hat{\omega}) = \sum_{m=-\infty}^{\infty} w[n-m]x[m]e^{-j\hat{\omega}m} = w[n] * \left(x[n]e^{-j\hat{\omega}n}\right)$$

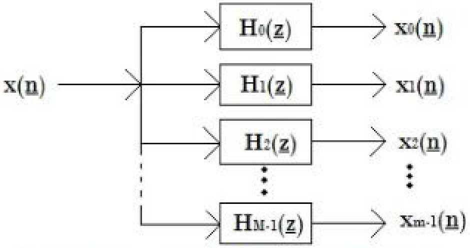
Bandpass filter interpretation of STFT

$X_n(\hat{\omega})$ can also be interpreted as $x[n]$ bandpass filtered by $h_{\hat{\omega}}[n] = e^{j\hat{\omega}n}w[n]$, followed by downmodulation:

$$X_n(\omega) = \sum_{m=-\infty}^{\infty} w[n-m]x[m]e^{-j\omega m} = e^{-j\hat{\omega}n}\left(h_{\hat{\omega}}[n] * x[n]\right)$$

Bandpass filter interpretation of STFT and ISTFT

Many audio coding algorithms are based on the interpretation of the STFT as a bank of bandpass filters, centered at the DFT bin frequencies $\hat{\omega}_k = \frac{2\pi k}{N}$:



Multidimensional Analysis Filter Banks

Outline

Examples

1. Using the approximation that a Hamming window is roughly triangular in frequency, figure out how much we can downsample the STFT, in the time domain, without causing aliasing.
2. It's possible to downsample the STFT by more than the factor we found in part (1), and, nevertheless, to recover $x[m]$ exactly, with no errors. Why?

Outline

Summary

$$
\begin{aligned}
X_n(\hat{\omega}) &\equiv \sum_{m=-\infty}^{\infty} w[n-m]x[m]e^{-j\hat{\omega}m} \\
&= e^{-j\hat{\omega}(n-N+1)} \sum_{m'=0}^{N-1} w[(N-1)-m']x[n+m']e{-}j\hat{\omega}m' \\
&= w[n] * \left(x[n]e^{-j\hat{\omega}n}\right) \\
&= e^{-j\hat{\omega}n}\left(h_{\hat{\omega}}[n] * x[n]\right)
\end{aligned}
$$

$$
x[m] = \frac{w[n-m]x[m]}{w[n-m]} = \frac{\sum_{n=-\infty}^{\infty} w[n-m]x[m]}{\sum_{n=-\infty}^{\infty} w[n-m]}
$$