Short-Time Fourier Transform with Applications to Speech Enhancement and Speech Recognition

Igor Fedorov

Dec. 3, 2015
Motivation

Figure: Cocktail party problem

Figure: Noise cancelling headphones

Figure: Siri, Ok Google, etc.
So far, you have studied signals which vary in frequency.

500 Hz sine wave, sampled at 8 kHz
What about signals that vary in frequency and time?

500 Hz sine wave followed by 2 kHz sine wave, sampled at 8 kHz
Most real-world signals have time-varying spectral characteristics.

Example: Speech
Instead of performing a DFT on the entire signal, let’s extract windows

Let $x^i$ be the $i$’th window of $x$:

$$x^i[n] = w[i \times m - n]x[(i - 1)m + n]$$

where $m$ is the window length

$$X_i(\omega) = DFT(x^i)$$

**Time Frequency Tradeoff**

As $m \uparrow$, frequency resolution $\uparrow$, but time resolution $\downarrow$
**Question:** Can we recover $x$ from $\{X_i\}_{i=1}^{N}$?

**Figure:** 1024 point Hanning window

- In effect, we are convolving each segment of $x$ with a low-pass filter
Better Time-Frequency Analysis

No more issues at window boundaries!
Overlap-Add Reconstruction

Question: Under what conditions does this method perfectly reconstruct $x$?

Answer: When the overlap is 75% [Allen, et. al. 1977]
Spectrogram

Stack the $X_i$ into a matrix, called a spectrogram:

$$X = \begin{bmatrix} X_1 & X_2 & \cdots & X_N \end{bmatrix}$$

- $X$ is a complex matrix. To visualize it, we can look at $|X|$.
Spectrogram for Speech Recognition
Given $x = x_{speech} + x_{noise}$, we would like to extract $x_{speech}$.
Speech and Noise Have Different Spectrograms
Denoising

- There are many speech denoising/speech enhancement/source separation algorithms
- Most approaches have a training phase, followed by a denoising phase

Microphone Array Based Approach
- Instead of a single microphone, we have access to \( L \) microphones
- **Spatial** filtering vs. **temporal** filtering

Dictionary Based Approach
- Use machine learning tools to learn what constitutes speech and noise
Array Approach

- Plane wave assumption
- Let $\mathbf{v}_i$ be the $i$'th microphone signal
- For now, assume a single source, $x_{speech}$ is present and $\theta$ is the angle of arrival, then

$$
\begin{bmatrix}
    v_1[n] \\
    \vdots \\
    v_L[n]
\end{bmatrix}
= x_{speech}[n]
\begin{bmatrix}
e^{j2\pi f_0 \tau_1} \\
\vdots \\
e^{j2\pi f_0 \tau_L}
\end{bmatrix}
$$

- Simplest approach: average the signals
- Can you use extra degrees of freedom to do more
Demo

- Mixed at -5dB SNR
Dictionary Learning Approach: Non-negative Matrix Factorization

- We wish to **learn** the **parts** of a speech/noise signal.
- Specifically, we will try to decompose $|X_{speech}|$:

\[ |X_{speech}| = WH, \quad W \geq 0, \quad H \geq 0 \]
(a) NMF bases.