Time-frequency Masking

EECS 352: Machine Perception of Music & Audio
• The **Short-Time Fourier Transform (STFT)** is a succession of local Fourier Transforms (FT)
STFT

• If we used a window of $N$ samples, the FT has $N$ values, from 0 to $N-1$; e.g., if $N = 8$...
STFT

- Frequency index 0 is the **DC component**; it is always real (it is the sum of the time values!)
STFT

• Frequency indices from 1 to floor(N/2) are the “unique” complex values \((a + j*b)\)
**STFT**

- Frequency indices from floor(N/2) to N-1 are the “mirrored” complex conjugates (a - j*b)

![Time signal](image1)

FT

![Real spectrum](image2)

+ j*

![Imaginary spectrum](image3)

Windows i

Time

frequency

Frame i
STFT

• If N is even, there is a **pivot component** at frequency index N/2; it is always real!
STFT

- Summary of the frequency indices and values in the STFT (in colors!)

N frequency values = frequency 0 to N-1

Frequency 0 = DC component (always real)

Frequency 1 to floor(N/2) = “unique” complex values

Frequency N/2 = “pivot” component (always real)

Frequency floor(N/2) to N-1 = “mirrored” complex conjugates
Spectrogram

- The (magnitude) **spectrogram** is the magnitude (absolute value) of the STFT.
• For a complex number $a + j* b$, the absolute value is $|a + j* b| = \sqrt{a^2 + b^2}$

Zafar Rafii, Winter 2014
• All the N frequency values (frequency indices from 0 to N-1) are **real and positive** (abs!)}
Spectrogram

- Frequency indices from 0 to floor(N/2) are the unique frequency values (with DC and pivot)

Zafar Rafii, Winter 2014
• Frequency indices from floor(N/2)+1 to N-1 are the **mirrored frequency values**

Spectrogram
• Since they are redundant, we can discard the frequency values from floor(N/2)+1 to N-1.
The spectrogram has therefore \( \text{floor}(N/2)+1 \) unique frequency values (with DC and pivot).

Spectrogram

Real spectrogram

Imaginary spectrogram

Magnitude spectrogram

frame \( i \)

time

frequency

Zafar Rafii, Winter 2014
Spectrogram

• Why the magnitude spectrogram?
  – Easy to visualize (compare with the STFT)
  – Magnitude information more important
  – Human ear less sensitive to phase
Spectrogram

• When you display a spectrogram in Matlab...
  – *imagesc*: data is scaled to use the full colormap
  – $10\times\log_{10}(V)$: magnitude spectrogram in dB
  – *set(gca,’YDir’,’normal’)*: y-axis from bottom to top
The signal **cannot be reconstructed** from the spectrogram (phase information is missing!)

- Magnitude spectrogram
- Imaginary spectrogram
- Real spectrogram
- Time signal

STFT

Zafar Rafii, Winter 2014
Time-frequency Masking

• Suppose we have a mixture of two sources: a music signal and a voice signal
Time-frequency Masking

- We assume that the sources are **sparse** = most of the time-frequency bins have null energy.

Music signal + Voice signal = Mixture signal

Zafar Rafii, Winter 2014
• We assume that the sources are **sparse** = most of the time-frequency bins have null energy.

Music spectrogram + Voice spectrogram = Mixture spectrogram

Music signal: Mostly low energy bins
Voice signal: Mostly low energy bins
Mixture signal: Mostly mixed energy bins

Zafar Rafii, Winter 2014
Time-frequency Masking

- We assume that the sources are **disjoint** = their time-frequency bins do not overlap
Time-frequency Masking

- We assume that the sources are **disjoint** = their time-frequency bins do not overlap

Zafar Rafii, Winter 2014
Time-frequency Masking

• Assuming sparseness and disjointness, we can discriminate the bins between mixed sources
Time-frequency Masking

• Assuming sparseness and disjointness, we can **discriminate** the bins between mixed sources.
Time-frequency Masking

- Bins that are likely to belong to one source are assigned to 1, the rest to 0 = binary masking!

Music spectrogram
Voice spectrogram
Mixture spectrogram

Source of interest
Interfering source

Binary mask

Music signal
Voice signal

Zafar Rafii, Winter 2014
Time-frequency Masking

- By multiplying the binary mask to the mixture spectrogram, we can “preview” the estimate.
• However, we cannot derive the estimate itself because we cannot invert a spectrogram!

Binary mask \( \times \) Mixture spectrogram

Masked spectrogram

Music estimate
Time-frequency Masking

• We mirror the redundant frequencies from the unique frequencies (without DC and pivot)
We then apply this full binary mask to the STFT using an element-wise multiplication.
Time-frequency Masking

- The estimate signal can now be reconstructed via inverse STFT
Time-frequency Masking

- Sources are not really sparse or disjoint in time-frequency in the mixture

Music signal $+$ Voice signal $=$ Mixture signal

Music spectrogram

Voice spectrogram

Mixture spectrogram

Zafar Rafii, Winter 2014
Time-frequency Masking

- Bins that are likely to belong to one source are close to 1, the rest close to 0 = soft masking!

Music spectrogram
Voice spectrogram
Mixture spectrogram

Music signal
Voice signal
Source of interest
Interfering source

Soft mask

Zafar Rafii, Winter 2014
Time-frequency Masking

• Let’s listen to the results!
Question

• How can we efficiently model a binary/soft time-frequency mask for source separation?...

• To be continued...

Mixture spectrogram

Soft mask

Zafar Rafii, Winter 2014