Musical Sound Source Separation
based on Computation Auditory Scene Analysis

Jinyu Han

Interactive Audio Lab
Northwestern University
Outline of presentation

- Cocktail party problem
- Overlapped Harmonics
- Least Square Estimation
Cocktail Party Problem

Fig. 1. A cocktail party (Image from Breakfast at Tiffany’s: Paramount Pictures)
Cocktail Party Problem

**Ensemble**: pick one instrument
Fig. 2. Bach Chorale: Ach Gott und Herr
Audio Source Separation

• Separating out the individual sounds in an audio mixture
Practical Applications

- Hearing Aids
- Automated transcription of speech and music
- Automated sound source identification
- Speech recognition systems
Interesting Questions

• How do humans separate sounds?

• Can we build a machine to do this?

• What cues in the sound are important to separate one sound from background noise?
Approaches to Audio Separation

• **Blind Source Separation (BSS)**
  - Few assumptions about the sound source itself
  - Usually works on mixture of at least two channels
  - Methods include: ICA, NMF, Beamforming

• **Computational Auditory Scene Analysis (CASA)**
  - Use heuristic grouping cues based upon psychological observation
  - Typically deal with single channel mixture
Outline of presentation

- Cocktail party problem
- Overlapped Harmonics
- Least Square Estimation
It’s NOT easy

Violin

Bassoon

Time (s)
It’s NOT easy

Lay each source on top of each other
It’s NOT easy (Time Domain)

Mixture of violin and bassoon
Time-frequency domain
Time-frequency domain

Violin
Overlap
Overlap
Overlap
Overlap
Bassoon
Time-frequency domain

Bassoon

Violin

Overlap

Overlap

Overlap

Overlap

Bassoon
Time-frequency domain

Violin

Overlap

Overlap

Overlap

Overlap

Bassoon
Time-frequency domain

Mixture

Violin
Overlap
Overlap
Overlap
Overlap
Bassoon

0 0.417959 0.835918 1.25388 1.67184 2.0898

2414 1389 345
Approach #1

Bassoon
Violin
discard
discard
Bassoon
discard
Bassoon
discard
Bassoon
discard
Bassoon

0 0.417959 0.835918 1.25388 1.67184 2.0898
Approach #1

• Find the un-overlapped parts (belonging to a single source) in the mixture

• Rebuild the sources from the un-overlapped parts
Approach #1

• Find the un-overlapped (ie single-source) parts of the mixture
  – Use a multi pitch tracker to track the pitch of each source

• Rebuild the sources from the un-overlapped parts
  – Based on the pitch, find the harmonics for each source
  – Rebuild only using the un-overlapped harmonics
Just Take the un-overlapped Part
Just Take the un-overlapped Part
DO SOMETHING with the overlap!!!
Take a close look at the first 10 harmonics
A 3-D plot

The Amplitude of the first 10 harmonics over time

1st Harmonic

5th Harmonic
Un-overlapped harmonic amplitude of bassoon from the mixture
Un-overlapped harmonic amplitude of bassoon from the mixture
Harmonic amplitude of bassoon in the mixture

4th and 8th harmonics are overlapped by violin in the mixture
Spectral Smoothness

• The Amplitude of a harmonic partial is usually close to the amplitudes of the nearby partials of the same sound.
Approach #2

• Find the un-overlapped (ie single-source) parts of the mixture
• Rebuild the sources from the un-overlapped parts
• Rebuild the overlapped parts by interpolating from the un-overlapped parts adjacent to the overlapped parts
Spectral Smoothness

Any problem???

Original

Reconstruction: 4th and 8th harmonics are interpolated from the neighboring harmonics
Harmonic amplitude envelope

Divide the amplitude of the harmonic at time $t$ by the amplitude of the harmonic at time $t=0$.

Harmonic amplitude

Harmonic amplitude envelope
Harmonic amplitude envelope

Harmonic amplitude envelope (in a log scale)
Common Amplitude Modulation (CAM)

- The amplitude envelopes of different harmonics of the same source exhibit similar temporal dynamics
Common Amplitude Modulation (CAM)

Amplitude of un-overlapped harmonics

Harmonic amplitude envelope (normalized by the first frame)
Common Amplitude Modulation

• For the overlapped harmonics, assume we know the amplitude at $t = 1$.

• Reconstruct the harmonic amplitude ($t = 2, 3, \ldots$) using the amplitude of the first frame ($t=1$) and the envelope of the neighboring harmonic envelope.
Approach #3

• Find the un-overlapped (ie single-source) parts of the mixture

• Estimate the amplitude of un-overlapped harmonics at $t = 1$;

• Rebuild the overlapped harmonics using the envelope of un-overlapped harmonics
Approach #3

- Let $H_4(t)$ indicate the amplitude of the $4^{th}$ harmonic at time $t$.
- The estimation:
  \[ H_4'(t) = \frac{H_4(1) \cdot H_5(t)}{H_5(1)} \]
Common Amplitude Modulation

Original

Reconstruction

The Amplitude of the first 10 harmonics over time.

Reconstructed using neighboring harmonics.
Outline of presentation

- Cocktail party problem
- Overlapped harmonics
- Least Square Estimation
Time-frequency domain

Bassoon
Violin
Overlap
Overlap
Overlap
Overlap
Bassoon
Estimated envelope of bassoon scaled by the initial value

A and B indicate the amplitude of harmonic at the begging time \( t = 1 \)

Estimated envelope of violin scaled by the initial value

Observed amplitude in the mixture
Estimated envelope of bassoon scaled by the initial value

A and B indicate the amplitude of harmonic at the beginning time $t = 1$

Estimated envelope of violin scaled by the initial value

Observed amplitude in the mixture
Estimate the initial amplitude

- Identify the overlapped parts from the mixture
- Estimate the amplitude envelope for the overlapped harmonics based on un-overlapped harmonics
- Do a least square estimation
Audio Source Separation

• Separating out the individual sounds in an audio mixture
Approach #1 Bassoon
Approach #3 Bassoon
Original Bassoon
Approach #1 Violin
Approach #3 Violin
Original Violin
Audio Source Separation

- Harmonic Masking
- Spectral Smoothness
- Common Amplitude Modulation
END

- Cocktail party problem
- Overlapped harmonics
- Least Square Estimation
Auditory Scene Analysis

- Listeners parse the complex mixture of sounds arriving at the ears in order to form a mental representation of each sound source.

- This perceptual process is called *auditory scene analysis*.

- Two conceptual processes of *auditory scene analysis (ASA)*:
  - **Segmentation.** Decompose the acoustic mixture into sensory elements (segments).
  - **Grouping.** Combine segments into groups, so that segments in the same group likely originate from the same environmental source.
Computational auditory scene analysis

- Computational auditory scene analysis (CASA) approaches sound separation based on ASA principles
- Pitch continuity
- Harmonic partials
- Spectral shape
- Harmonic temporal envelope
- ……
Pitch and Harmonics
Timbre and Spectral shape

- **Harmonic structure feature**
  - Normalized relative amplitudes of harmonics
Timbre and Spectral shape

PCA of Harmonic Structures

2nd Component

1st Component

Violin
Clarinet
Saxophone
Bassoon
Group pitches into streams
Harmonic temporal Envelope

Fig: Amplitude envelopes of a clarinet playing a G#
Common Amplitude Modulation

- Harmonics of same source have correlated envelope
- Harmonics with strong energy are more correlated
Outline of presentation

- Cocktail party problem
- Computational Auditory Scene Analysis (CASA)
- Harmonic instrument separation based upon CASA
Sinusoids

\[ \text{Amp. } 10 \sin(2\pi \cdot 2.5 \cdot t + \pi/4) \]

- **A Sin[ϕ]**
- **Zero Crossings**
- **Amplitude Peaks**
- **Period** \( P = \frac{2\pi}{\omega} = \frac{1}{f} \)
- **Peak-to-Peak Amplitude = 2A**
Fourier Transform

• Fourier Transform break a signal into sum of sinusoids
Sinusoid Model

- Each Harmonic:
  \[ h_n(t) = \alpha_n(m) \cos(2\pi f_n(m)t + \phi_n(m)) \]  \hspace{1cm} (2)

- Harmonic Sound:
  \[ x_m(t) = \sum_{n} \sum_{h_n=1}^{H_n} \alpha_n^{h_n}(m) \cos(2\pi f_n^{h_n}(m)t + \phi_n^{h_n}(m)) \]  \hspace{1cm} (3)

- \( \alpha_n^{h_n}(m) \), \( f_n^{h_n}(m) \) and \( \phi_n^{h_n}(m) \) are the amplitude, frequency and phase of sinusoidal component \( h_n(t) \).
Phase Change

- Phase change of a harmonic is a function of Pitch.
  \[ \phi_n^{h_n}(m + 1) - \phi_n^{h_n}(m) = 2\pi h_n F_n(m) T \]
- Predicted change is accurate in lower-numbered harmonics
Phase Change
Sinusoidal Model

Spectral value of $x_m(t)$ at frequency bin $k$:

$$X(m, k) = \sum_n S_n^{h_n}(m) W(kf_b - h_n F_n(m))$$

- $W$: the DFT of the analysis window.
- $f_b$: frequency resolution of the DFT.
- $F_n(m)$: pitch.

Sinusoidal parameter:

$$S_n^{h_n}(m) = \frac{\alpha_n^{h_n}(m)}{2} e^{i\phi_n^{h_n}(m)}$$
Sinusoidal Model

- Common Amplitude Modulation implies that:
  - $\gamma_{n}^{h}$ could be estimated from another harmonic of source $n$.
  - $\gamma_{m_{0}\rightarrow m}^{h}$ $\approx$ $\gamma_{m_{0}\rightarrow m}^{h^{*}}$ $= \frac{\alpha_{n}^{h^{*}}(m)}{\alpha_{n}^{h}(m_{0})}$.
  - $h_{n}^{*}$ is non-overlapped harmonic with strong energy.
  - Phase Change $\Delta \phi_{n}^{h}$ is only dependent of pitch.
Sinusoidal Model

- Revisit:
  - Energy of source $n$ in spectrogram $(m, k)$ is

$$X_n(m, k) = S_n^{hn}(m) W(kf_b - h_n F_n(m))$$

where sinusoidal parameter $S_n^{hn}(m) = \frac{\alpha_n^{hn}(m)}{2} e^{i\phi_n^{hn}(m)}$.

- Sinusoidal parameter change:

$$S_n^{hn}(m) = S_n^{hn}(m_0) \gamma_{m_0 \rightarrow m}^{hn} e^{i \sum_{\ell = m_0}^{m} \Delta \phi_n^{hn}(\ell)}$$

- Amplitude scaling:

$$\gamma_{m_0 \rightarrow m}^{hn} = \frac{\alpha_n^{hn}(m)}{\alpha_n^{hn}(m_0)}$$
Sinusoidal Model

- Review:
  - Observed spectral value of the mixture

\[ X(m, k) = \sum_n S_n^{h_n}(m) W(kf_b - h_n F_n(m)) \]

- Sinusoidal parameter of harmonic \( h_n \).

\[ S_n^{h_n}(m) = S_n^{h_n}(m_0) \gamma_{m_0 \rightarrow m}^{h_n} e^{i \sum_{\nu = m_0}^m \Delta \phi_n(\nu)} \]

- Rewrite:

\[ X(m, k) = \sum_n S_n^{h_n}(m_0) R_n(m, k) \]

- \( R_n \) is known

\[ R_n(m, k) = W(kf_b - h_n F_n(m)) \gamma_{m_0 \rightarrow m}^{h_n} e^{i \sum_{\nu = m_0}^m \Delta \phi_n(\nu)} \]

- \( X(m, k) \) is observed.
Sinusoidal Model

Rewrite equation (11) in overlapping region \( D(m_0, m_i; k_0, k_i) \):

\[
\begin{pmatrix}
R_1(m_0, k_0) & \ldots & R_N(m_0, k_0) \\
\vdots & \ddots & \vdots \\
R_1(m_i, k_i) & \ldots & R_N(m_i, k_i)
\end{pmatrix}
\begin{pmatrix}
S_1^{h_1}(m_0) \\
\vdots \\
S_N^{h_i}(m_0)
\end{pmatrix}
= 
\begin{pmatrix}
X(m_0, k_0) \\
\vdots \\
X(m_i, k_i)
\end{pmatrix}
\tag{13}
\]

\[
\downarrow \downarrow
\]

\[RS = X\]
\tag{14}

Least-squares estimation of \( S \) is given by:

\[
S = (R^HR)^{-1}R^HX
\]
\tag{15}