Sound separation and enhancement

- A common problem: Separate or enhance sounds
- Tools are applicable to other forms of data as well.
- A popular approach: can be done with principal component analysis
- A common problem: separate music components
- Suppress “bleed” in music recordings
- Speech from noise
- A popular approach: can be done with probabilistic latent component analysis
- Sounds in general is composed of such “building blocks” or “themes”. Sound components sounds that we perceive.
- Notes from music
- Phoneme-like structures combine in utterances.
- The individual component sounds “combine” to form the
  final complex sounds that we perceive.
- Notes from music
- Phoneme-like structures combine in utterances.
- The individual component sounds “combine” to form the
  final complex sounds that we perceive.
- Notes from music
- Phoneme-like structures combine in utterances.
- The individual component sounds “combine” to form the
  final complex sounds that we perceive.
- Notes from music
- Phoneme-like structures combine in utterances.
- The individual component sounds “combine” to form the
  final complex sounds that we perceive.
- Notes from music
- Phoneme-like structures combine in utterances.
- The individual component sounds “combine” to form the
  final complex sounds that we perceive.
- Notes from music
- Phoneme-like structures combine in utterances.
- The individual component sounds “combine” to form the
  final complex sounds that we perceive.
- Notes from music
- Phoneme-like structures combine in utterances.
- The individual component sounds “combine” to form the
  final complex sounds that we perceive.
- Notes from music
- Phoneme-like structures combine in utterances.
- The individual component sounds “combine” to form the
  final complex sounds that we perceive.
- Notes from music
- Phoneme-like structures combine in utterances.
- The individual component sounds “combine” to form the
  final complex sounds that we perceive.
- Notes from music
- Phoneme-like structures combine in utterances.
- The individual component sounds “combine” to form the
  final complex sounds that we perceive.
- Notes from music
- Phoneme-like structures combine in utterances.
- The individual component sounds “combine” to form the
  final complex sounds that we perceive.
- Notes from music
- Phoneme-like structures combine in utterances.
- The individual component sounds “combine” to form the
  final complex sounds that we perceive.
- Notes from music
- Phoneme-like structures combine in utterances.
- The individual component sounds “combine” to form the
  final complex sounds that we perceive.
- Notes from music
- Phoneme-like structures combine in utterances.
- The individual component sounds “combine” to form the
  final complex sounds that we perceive.
- Notes from music
- Phoneme-like structures combine in utterances.
- The individual component sounds “combine” to form the
  final complex sounds that we perceive.
- Notes from music
- Phoneme-like structures combine in utterances.
- The individual component sounds “combine” to form the
  final complex sounds that we perceive.
- Notes from music
- Phoneme-like structures combine in utterances.
- The individual component sounds “combine” to form the
  final complex sounds that we perceive.
- Notes from music
- Phoneme-like structures combine in utterances.
- The individual component sounds “combine” to form the
  final complex sounds that we perceive.
- Notes from music
- Phoneme-like structures combine in utterances.
- The individual component sounds “combine” to form the
  final complex sounds that we perceive.
- Notes from music
- Phoneme-like structures combine in utterances.
- The individual component sounds “combine” to form the
  final complex sounds that we perceive.
- Notes from music
- Phoneme-like structures combine in utterances.
- The individual component sounds “combine” to form the
  final complex sounds that we perceive.
- Notes from music
- Phoneme-like structures combine in utterances.
- The individual component sounds “combine” to form the
  final complex sounds that we perceive.
- Notes from music
- Phoneme-like structures combine in utterances.
- The individual component sounds “combine” to form the
  final complex sounds that we perceive.
- Notes from music
- Phoneme-like structures combine in utterances.
- The individual component sounds “combine” to form the
  final complex sounds that we perceive.
- Notes from music
- Phoneme-like structures combine in utterances.
- The individual component sounds “combine” to form the
  final complex sounds that we perceive.
- Notes from music
- Phoneme-like structures combine in utterances.
- The individual component sounds “combine” to form the
  final complex sounds that we perceive.
- Notes from music
- Phoneme-like structures combine in utterances.
- The individual component sounds “combine” to form the
  final complex sounds that we perceive.
- Notes from music
- Phoneme-like structures combine in utterances.
- The individual component sounds “combine” to form the
  final complex sounds that we perceive.
- Notes from music
- Phoneme-like structures combine in utterances.
- The individual component sounds “combine” to form the
  final complex sounds that we perceive.
- Notes from music
- Phoneme-like structures combine in utterances.
- The individual component sounds “combine” to form the
  final complex sounds that we perceive.
- Notes from music
- Phoneme-like structures combine in utterances.
- The individual component sounds “combine” to form the
  final complex sounds that we perceive.
- Notes from music
- Phoneme-like structures combine in utterances.
- The individual component sounds “combine” to form the
  final complex sounds that we perceive.
- Notes from music
- Phoneme-like structures combine in utterances.
- The individual component sounds “combine” to form the
  final complex sounds that we perceive.
- Notes from music
- Phoneme-like structures combine in utterances.
- The individual component sounds “combine” to form the
  final complex sounds that we perceive.
- Notes from music
- Phoneme-like structures combine in utterances.
- The individual component sounds “combine” to form the
  final complex sounds that we perceive.
- Notes from music
- Phoneme-like structures combine in utterances.
- The individual component sounds “combine” to form the
  final complex sounds that we perceive.
- Notes from music
- Phoneme-like structures combine in utterances.
- The individual component sounds “combine” to form the
  final complex sounds that we perceive.
- Notes from music
- Phoneme-like structures combine in utterances.
- The individual component sounds “combine” to form the
  final complex sounds that we perceive.
- Notes from music
- Phoneme-like structures combine in utterances.
- The individual component sounds “combine” to form the
  final complex sounds that we perceive.
- Notes from music
- Phoneme-like structures combine in utterances.
- The individual component sounds “combine” to form the
  final complex sounds that we perceive.
- Notes from music
- Phoneme-like structures combine in utterances.
- The individual component sounds “combine” to form the
  final complex sounds that we perceive.
- Notes from music
- Phoneme-like structures combine in utterances.
- The individual component sounds “combine” to form the
  final complex sounds that we perceive.
- Notes from music
- Phoneme-like structures combine in utterances.
- The individual component sounds “combine” to form the
  final complex sounds that we perceive.
- Notes from music
- Phoneme-like structures combine in utterances.
- The individual component sounds “combine” to form the
  final complex sounds that we perceive.
- Notes from music
- Phoneme-like structures combine in utterances.
- The individual component sounds “combine” to form the
  final complex sounds that we perceive.
- Notes from music
- Phoneme-like structures combine in utterances.
- The individual component sounds “combine” to form the
  final complex sounds that we perceive.
- Notes from music
- Phoneme-like structures combine in utterances.
- The individual component sounds “combine” to form the
  final complex sounds that we perceive.
- Notes from music
- Phoneme-like structures combine in utterances.
- The individual component sounds “combine” to form the
  final complex sounds that we perceive.
- Notes from music
- Phoneme-like structures combine in utterances.
- The individual component sounds “combine” to form the
  final complex sounds that we perceive.
- Notes from music
- Phoneme-like structures combine in utterances.
- The individual component sounds “combine” to form the
  final complex sounds that we perceive.
- Notes from music
- Phoneme-like structures combine in utterances.
- The individual component sounds “combine” to form the
  final complex sounds that we perceive.
- Notes from music
- Phoneme-like structures combine in utterances.
- The individual component sounds “combine” to form the
  final complex sounds that we perceive.
- Notes from music
- Phoneme-like structures combine in utterances.
- The individual component sounds “combine” to form the
  final complex sounds that we perceive.
Two pickers

- The probability of selecting a particular pot is different for each picker. The pots are the same, and the probability of drawing a ball marked with a particular number is the same for both pickers.

- The probability of selecting a particular pot is independently computed for each picker.

- The probability of selecting pot color $P(X|\text{color})$ for a particular color is different for each picker.

- Let $P_i(x)$ and $P_i'(x)$ be the probability of drawing a number $x$ for the $i$th picker.

- The probability of drawing a number $x$ for the $i$th picker is $P_i(x) = P_i(\text{red})P(X|\text{red}) + P_i(\text{blue})P(X|\text{blue})$.

- The probability of drawing a number $x$ for both pickers is $P_{12}(x) = P_{1}(\text{red})P_{2}(X|\text{red}) + P_{1}(\text{blue})P_{2}(X|\text{blue})$.

- But combine the counts of balls in the pots separately for each caller. Complete the probability of selecting pots separately for each caller. Analyze each of the callers separately.

- Two tables with two pickers.

- Problem: Given the set of numbers called out by both pickers, estimate $P_1(\text{color})$ and $P_2(\text{color})$ for both colors. $P(X|\text{red})$ and $P(X|\text{blue})$ for all values of $X$.

- PICKER 1

- PICKER 2

- PICKER 1 picked the balls.
In Squiggles

Given a sequence of observations $O_{k,1}, O_{k,2}, ..$ from the $k$th picker

$N_{k,X}$ is the number of observations of color $X$ drawn by the $k$th picker

Initialize $P_k(Z), P(X|Z)$ for pots $Z$ and colors $X$

Iterate:

For each color $X$, for each pot $Z$ and each observer $k$:

Update the probability of numbers for the pots:

Update the mixture weights: probability of urn selection for each picker
A generative model for one frame of a spectrogram

A magnitude spectral vector obtained from a DFT represents spectral magnitude against discrete frequencies. This may be viewed as a histogram of draws from a multinomial distribution over frequency bins.

The balls are marked with discrete frequency indices from the DFT.

The picker has a fixed set of urns, each with a different probability distribution over frequency bins. In the first draw, he selects an urn according to some probability $P_0(z)$, and then draws the spectrum for the first frame. In the second draw, he selects an urn according to some probability $P_1(z)$, and then draws the spectrum for the second frame. And so on, until he has constructed the entire spectrogram.

Two aspects of the multinomial model are combined: the prior probability with which he selects any urn, and the multinomial probability distribution over the urn. The urns represent different models of the spectral envelope.

A more complex model might use multiple urns, each with a different multinomial distribution. The picker has a fixed set of urns, each with a different probability distribution over frequency bins. In the first draw, he selects an urn according to some probability $P_0(z)$, and then draws the spectrum for the first frame. In the second draw, he selects an urn according to some probability $P_1(z)$, and then draws the spectrum for the second frame. And so on, until he has constructed the entire spectrogram.
The Picker Generates a Spectrogram

- The picker has a fixed set of Urns
  - Each urn has a different probability distribution over $f$

- He draws the spectrum for the first frame
  - In which he selects urns according to some probability $P_0(z)$

- Then draws the spectrum for the second frame
  - In which he selects urns according to some probability $P_1(z)$

- And so on, until he has constructed the entire spectrogram

The number of draws in each frame represents the RMS energy in that frame.

By "learning" the mixture multinomial model for any sound source, we "discover" the latent spectral structures that characterize that source.

The model can be learned from spectrograms of a small amount of audio from the source with the EM algorithm.

EM learning of bases

- Initialize $P(f|z)$
- For each frame
  - Must decide on the number of urns $l$ for all $f$
  - Initialize bases $A_l$f

For the given sound source

- The mixture multinomial representation of latent spectral structures (bases)

For the $l$th frame

- Each component multinomial (urn) is actually a normalized histogram $P(f|z)$
- (normalized histogram over frequencies)
- Each component multinomial (urn) is actually a normalized histogram $P(f|z)$

The amount of activity between frames is explained as additive

The mixture model can be written as

\[
\sum_l \sum_{z} P(l|z)P(z|f) = P(f)
\]

The mixture weights $P(l|z)$ can be estimated using the EM algorithm.

By estimating the mixture multinomial model, we can "discover" the latent spectral structures that characterize the source.

The model can be learned from spectrograms of a small amount of audio from the source with the EM algorithm.
How the bases compose the signal

- The overall signal is the sum of the contributions of individual urns
- Each urn contributes a different amount to each frame
- The contribution of the z-th urn to the t-th frame is given by
  \[ P(f|z)P(t|z)S(f) \]

EM Update Equations

- Iterative process:
  1. Compute a posteriori probability of the z-th urn for the source for each f
  2. Compute mixture weight of z-th urn
  3. Compute the probabilities of the frequencies for the z-th urn

\[ \sum_{z} P(z) = 1 \]

Learning Structures

- Bag of Spectrograms PLCA Model
  - Compose the entire spectrogram all at once
  - Urns include two types of balls
    - One set of balls represents frequency F
    - The second has a distribution over time T
  - Each draw:
    1. Select an urn
    2. Draw "F" from frequency pot
    3. Draw "T" from time pot
    4. Increment histogram at (T,F)

Estimating the bag of spectrograms

- EM update rules
  1. Can learn all parameters
  2. Can learn P(T|Z) and P(Z) only given P(f|Z)
  3. Can learn only P(Z)

\[ \sum_{z} P(z) = 1 \]
How meaningful are these structures?

- Are these really the "notes" of sound?
- To investigate, let's go back in time..

The Engineer and the Musician

Once upon a time a rich potentate discovered a previously unknown recording of a beautiful piece of music. Unfortunately it was badly damaged.

He greatly wanted to find out what it would sound like if it were not damaged.

So he hired an engineer and a musician to solve the problem.

Who do you think won the princess?

The Prize

The Engineer and the Musician

- The Engineer worked for many years. He spent much money and published many papers.
- The musician listened to the music carefully for a day, transcribed it, broke out his trusty keyboard and replicated the music.

What the musician can do

- Notes are distinctive
- The musician knows notes (of all instruments)
- He can identify notes and their cadence

The Engineer and the Musician

- The Engineer works on the signal to reconstruct the music.
  - He uses these skills to reconstruct the music.
  - He took many years to learn these skills.
  - He can identify notes and their cadence.
  - He knows how music is composed.

The musician works on his familiarity with music.

- The musician knows notes (of all instruments)
- Notes are distinctive
- He can recognize individual components

The Engineer and the Musician

- The musician listened to the signal.
  - He used his trusty keyboard and carefully transcribed it.
- Finally he had a somewhat scratchy restoration of the music.
  - The musician listened to the signal.
  - He spent much money and published many papers.
  - He spent many years working for many years.

- The engineer worked on the signal to reconstruct the music.
  - He used his trusty keyboard and carefully transcribed it.
  - Finally he had a somewhat scratchy restoration of the music.

The Prize

- The Engineer and the Musician
  - The musician listened to the signal.
  - He spent much money and published many papers.
  - He spent many years working for many years.
  - He used his trusty keyboard and carefully transcribed it.
  - Finally he had a somewhat scratchy restoration of the music.
  - The engineer worked on the signal to reconstruct the music.
  - He used his trusty keyboard and carefully transcribed it.
  - Finally he had a somewhat scratchy restoration of the music.

The Engineer and the Musician

- What do you think won the princess?
- The engineer worked on the signal to reconstruct the music.
  - He used his trusty keyboard and carefully transcribed it.
  - Finally he had a somewhat scratchy restoration of the music.
- The musician listened to the music carefully for a day, transcribed it, broke out his trusty keyboard and replicated the music.

The Engineer and the Musician

- The Engineer worked for many years. He spent much money and published many papers.
- The musician listened to the music carefully for a day, transcribed it, broke out his trusty keyboard and replicated the music.

- The musician knows notes (of all instruments)
- Notes are distinctive
- He can recognize individual components

The Engineer and the Musician

- The Engineer works on the signal to reconstruct the music.
  - He uses these skills to reconstruct the music.
  - He took many years to learn these skills.
  - He can identify notes and their cadence.
  - He knows how music is composed.

The musician works on his familiarity with music.

- The musician knows notes (of all instruments)
- Notes are distinctive
- He can recognize individual components

The Engineer and the Musician

- The engineer worked on the signal to reconstruct the music.
  - He used his trusty keyboard and carefully transcribed it.
  - Finally he had a somewhat scratchy restoration of the music.
- The musician listened to the music carefully for a day, transcribed it, broke out his trusty keyboard and replicated the music.

The Engineer and the Musician

- What do you think won the princess?
- The engineer worked on the signal to reconstruct the music.
  - He used his trusty keyboard and carefully transcribed it.
  - Finally he had a somewhat scratchy restoration of the music.
- The musician listened to the music carefully for a day, transcribed it, broke out his trusty keyboard and replicated the music.

The Engineer and the Musician

- The Engineer works on the signal to reconstruct the music.
  - He uses these skills to reconstruct the music.
  - He took many years to learn these skills.
  - He can identify notes and their cadence.
  - He knows how music is composed.

The musician works on his familiarity with music.

- The musician knows notes (of all instruments)
- Notes are distinctive
- He can recognize individual components

The Engineer and the Musician

- The engineer worked on the signal to reconstruct the music.
  - He used his trusty keyboard and carefully transcribed it.
  - Finally he had a somewhat scratchy restoration of the music.
- The musician listened to the music carefully for a day, transcribed it, broke out his trusty keyboard and replicated the music.
We must estimate the draws of the unseen frequencies. However, we are only able to observe the number of draws of some frequencies and not the others.

The musician must restore them.

Bandwidth Expansion

The picker has drawn the histograms for every frame in the signal.

However, we are only able to observe the number of draws of some frequencies and not the others.

We must estimate the draws of the unseen frequencies.

Can we estimate the rest of the frequencies?

Telephone quality speech.

Problem: a given speech signal only has frequencies in the 300Hz-3KHz range.

Music over a telephone

The musician must restore it.
Step 1: Learning
- From a collection of full-bandwidth training data that are similar to the bandwidth-reduced data, learn spectral bases.
- Each magnitude spectral vector is a mixture of a common set of bases.
- Use the EM to learn bases from them.

Step 2: Estimation
- Using only the observed frequencies in the bandwidth-reduced data, estimate mixture weights for the bases learned in step 1.
- Find out which notes were active at what time.

Step 3 and Step 4: Recompose
- Compose the complete probability distribution for each frame using the mixture weights estimated in step 2.

The negative multinomial
- The unobserved frequencies
  - The complete probability distribution $P(z)$ to predict
  - The observed set of observed frequencies
  - The total number of observed counts $N$

$$P(z) = \frac{1}{Z} \prod_{i=1}^{S} \left( \sum_{j=1}^{M} \left( \frac{N_{i,j}}{N} \right) \right)$$

Predicting from $P(z)$: Simplified Example
- A single urn with only red and blue balls.
  - Given that out of an unknown number of draws, exactly $m$ were red, how many were blue?
  - A simple solution:
    - Total number of draws $N = m / P(\text{red})$
    - The number of tails drawn $= N \times P(\text{blue})$
  - Actual multinomial solution is much slightly more complex.

Step 2: Estimation
- Composite mixture weights of $z_{t}$ for each frame $f$
  - Speaker for each $f$
  - Compute a posterior probability of the $z_{t}$ urn for the utterance process: “Transcribe”

Step 1: Learning
- Basically learning the “notes”
  - Use the EM to learn bases from them.
  - Each magnitude spectral vector is a mixture of a common set of bases.
  - Learn spectral bases that are similar to the bandwidth-reduced data.

Bandwidth Expansion: Step 2 – Estimation
- From a collection of full-bandwidth training data.

Bandwidth Expansion: Step 1 – Learning
- Using only the observed frequencies in the bandwidth-reduced data.
Estimating unobserved frequencies

- Expected value of the number of draws from a negative multinomial:
  \[ \sum_{s} \sum_{t} \mathbb{E} \left[ f_{s, t} \right] = \sum_{s} \sum_{t} f_{s, t} \]

- Estimated spectrum in unobserved frequencies
  \[ \hat{P}(f_{n}, f_{s}, S_{t}, t, t) \]

Overall Solution

- Learn the "urns" for the signal source from broadband training data
- For each frame of the reduced bandwidth test utterance, find mixture weights for the urns
- Ignore (marginalize) the unseen frequencies
- Given the complete mixture multinomial distribution for each frame, estimate spectrum (histogram) at unseen frequencies

Signal Separation from Monaural Recordings

- Multiple sources are producing sound simultaneously
- The combined signals are recorded over a single microphone
- The goal is to selectively separate out the signal
- At least to enhance the signals from a selected source

Resolving the components

- The musician wants to follow the individual tracks in the recording.
- Effectively "separate" or "enhance" them against the background
- The musician wants to follow the individual tracks in the recording.

Prediction of Audio

- An example with random spectral holes

Estimating unobserved frequencies

- Expected value of the number of draws from a negative multinomial:
  \[ \mathbb{E}[N] = \sum_{f} \sum_{i} f_{i, f} \]

- Estimated spectrum in unobserved frequencies
  \[ \hat{P}(f_{n}, f_{s}, S_{t}, t, t) \]
Supervised separation: Example with two sources

A simple maximum a posteriori estimator

1. $s_i(t)$ – The spectrum at frequency $f_i$ of the mixed signal
2. $Z(t)$ – The spectrum at frequency $f_i$ of the unmixed signal
3. $S_i(t)$ – The spectrum of the separated signal for the $i$-th source

Supervised separation: Example with two sources

Goal: Estimate the contribution of individual sources

Each source has its own basis

All bases combine to generate the mixed signal

Can be learned from unmixed recordings of the source

Separating the sources: Cleaner Solution

Find mixture weights for all bases for each frame

Separating the contribution of bases from each source

Supervised separation: Example with unknown sources
Separating Mixed Signals: Examples

- “Raise my rent” by David Gilmour
  - Background music bases learnt from 5 seconds of music-only segments within the song
  - Lead guitar bases learnt from the rest of the song

- Norah Jones singing “Sunrise”
  - A more difficult problem

A more difficult problem:

Separate overlapping speech

- Bases for both speakers learnt from 5 second recordings of individual speakers
  - Shows improvement of about 5dB in Speaker-to-Speaker ratio for both speakers
  - Improvements are worse for same-gender mixtures of individual speakers.

Where it works

- When the spectral structures of the two sound sources are distinct
  - E.g. lead guitar and music
  - A more difficult problem

Can be improved

- Yes
  - More bases per source
  - More training data per source
  - Tweaking
- More training data per source
  - More bases per source

More on the topic

- Shift-invariant representations
  - Newest-highest, neighbor representations
  - Sparser overcomplete representations
  - And/or algorithmic improvements

Can it be improved?

- Yes
  - More training data per source
  - More bases per source
  - Adjusting FFT sizes and windows in the signal processing

Where it works

- Not as effective when the sources are similar
  - E.g. lead guitar and music
  - Don’t look much like one another

Separating overlapping speech

- From the rest of the song
  - Lead guitar
  - Bass
  - “Raise my rent” by Paul Simon

Examples

- From 5 seconds of music-only
  - Background music, “Sunrise”
  - A more difficult problem
  - “Raise my rent” by Paul Simon

Patterns extend beyond a single frame

- Etc.
  - Newest-highest, neighbor representations
  - Sparser overcomplete representations
  - And/or algorithmic improvements

Can be improved

- Tweaking
  - More training data per source

Separating overlapping speech

- Not as effective when the sources are similar
  - E.g. lead guitar and music
  - Don’t look much like one another

Separating overlapping speech

- When the spectral structures of the two sound
The shift-invariant model

- Employs bag of spectrograms model
- Each "super-urn" (z) has two sub urns
- One suburn now stores a bi-variate distribution
  - Each ball in the other suburn merely has a time "T" marked on them
  - The fragment in each super-urn is further fragmented into each time-shift

Maximum likelihood estimate follows fragmentation and counting strategy

- Two-step fragmentation
  - Each instance is fragmented into the super-urns
  - Since one can arrive at a given (t,f) by selecting any T from P(T|Z) and the appropriate shift T-t from P(t,f|Z)

Given data (spectrogram S(t,f))
- Initialize P(Z), P(T|Z), P(t,f|Z)
- Iterate:
  - \[ P(T, f) = \sum_z P(z) \sum T P(T|z) P(T - t, f) \]

Another example: Dereverberation

- Assume generation by a single latent variable
- The t-f basis is the "clean" spectrogram

An Example

- Two distinct sounds occurring with different repetition rates within a signal
- Input spectrogram

An Example

- Two distinct sounds occurring with different repetition rates within a signal
- Input spectrogram
Dereverberation: an example

- "Basis" spectrum must be made sparse for effectiveness
- Dereverberation of gamma-tone spectrograms is also particularly effective for speech recognition

The shift-invariant model

- Patterns may be substructures
- Repeating patterns that may occur anywhere
- Not just in the same frequency or time location
- More apparent in image data

The two-D Shift-Invariant Model

- Both sub-pots are distributions over \((T,F)\) pairs
- One sub-pot represents the basic pattern
- The other sub-pot represents the location

\[
P(T,F|Z) = \sum_{Z} P(z) \sum_{T,F} P(T,F|z) P(T,F|z) P(T,F|z) P(T,F|z)
\]

Two-D Shift Invariance: Estimation

- Fragment and count strategy
- Fragment into superpots, but also into each \(T\) and \(F\)

\[
P(T,F|Z) = \sum_{Z} P(z) \sum_{T,F} P(T,F|z) P(T,F|z)
\]

Shift-Invariance: Comments

- \(P(T,F|Z)\) and \(P(t,f|Z)\) are symmetric
- Constraints
- Can control which of them learns patterns and which the locations
- I.e. the size of the basic patch
- Other tricks — e.g. sparsity
Shift-Invariance in Many Dimensions

The generic notion of "shift-invariance" can be extended to multivariate data

- Not just two-D data like images and spectrograms
- The generic notion of "shift-invariance" can be extended to multivariate data

Example: 2-D shift invariance

Example: 3-D shift invariance

Pitch Tracking

The constant Q transform

Spectrographic analysis with a bank of constant Q filters

- The bandwidth of filters increases with center frequency
- The spacing between center frequencies increases with frequency
- Logarithmic spacing

Constant Q representation of Speech

Energy at the output of a bank of filters with logarithmically spaced center frequencies

- Like a spectrogram with non-linear frequency axis
- Changes in pitch become vertical translations of spectrogram

- Different vertical locations correspond to different notes of an instrument
- Changes in pitch become vertical translations of spectrogram
- The a spectrogram with non-linear frequency axis

Pitch Tracking

- Changing pitch becomes a vertical shift in the location of a basis
- The constant-Q spectrogram is modeled as a single pattern

\[ P(f) = \text{Kernel} \times \text{Output} \]

Different vertical locations correspond to different notes of an instrument

The original figure has multiple handwritten renderings of three characters in different colors

- The algorithm learns the three characters and identifies their locations in the figure
- The original figure has multiple handwritten renderings of three characters
- The algorithm learns the three characters and identifies their locations in the figure
- The algorithm learns the three characters and identifies their locations in the figure
In Conclusion

- Surprising use of EM for audio analysis
- Various extensions
  - Sparse estimation
  - Example-based methods
- Related deeply to non-negative matrix factorization
- TBD

The “impulse” distribution shows pitch of both separately

Example: A voice and an instrument overlaid

Pitch tracking of multiple sources

Having more than one bass (z) allows simultaneous

Left: A vocalized “song”

Right: Chord sequence

“impulse” distribution captures the “melody”