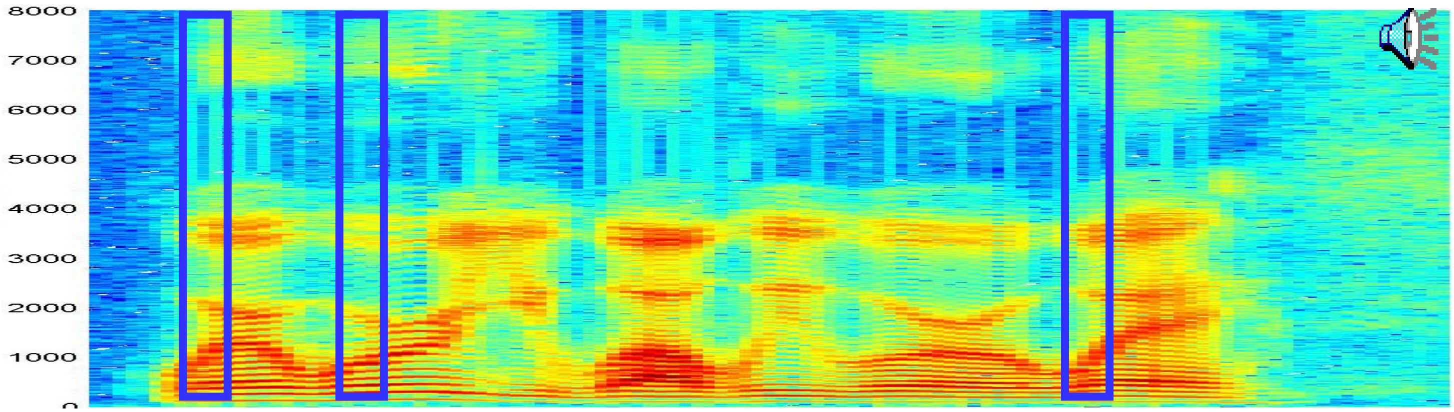
Latent Variable Models and Signal Separation

Class 12. 11 Oct 2011

Summary So Far

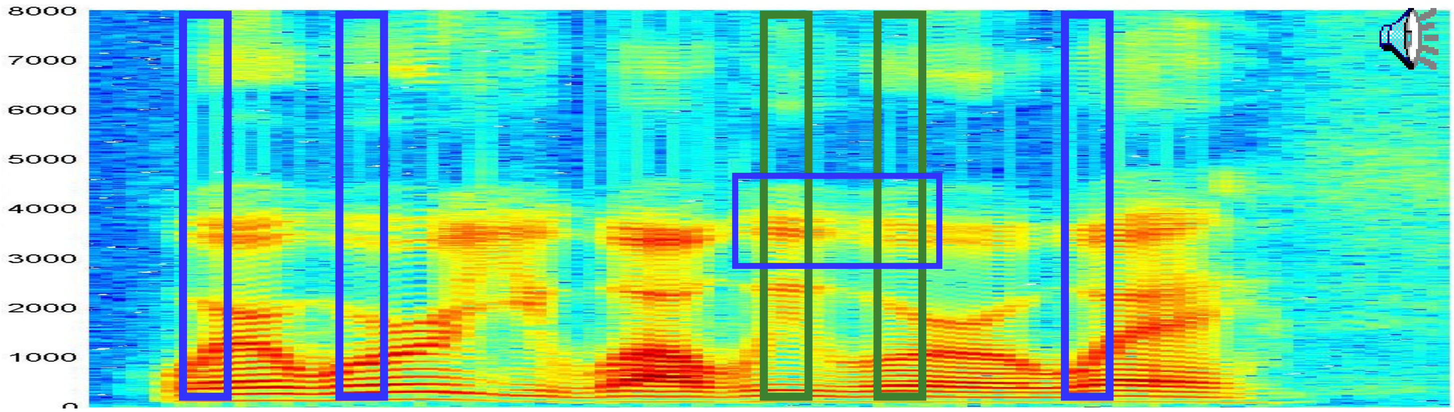
- PLCA:
  - The basic mixture-multinomial model for audio (and other data)
- Sparse Decomposition:
  - The notion of sparsity and how it can be imposed on learning
- Sparse Overcomplete Decomposition:
  - The notion of *overcomplete* basis set
- Example-based representations
  - Using the training data itself as our representation

Next up: Shift/Transform Invariance



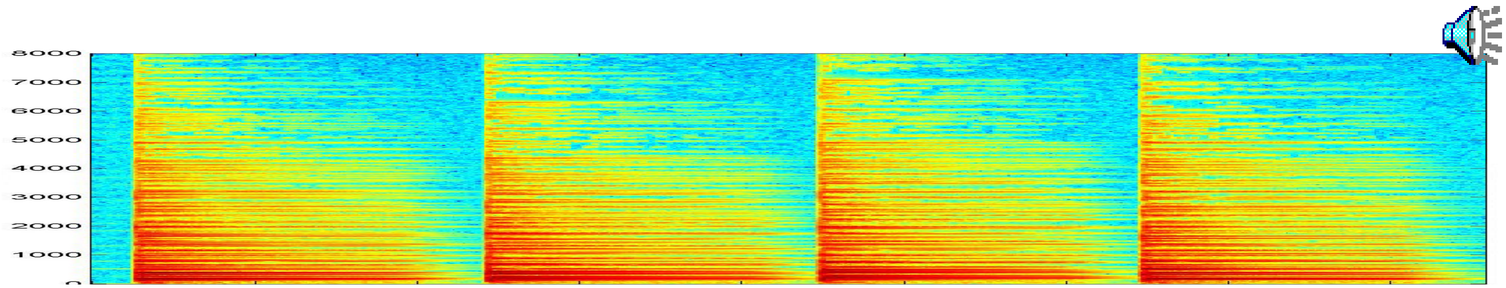
- Sometimes the “typical” structures that compose a sound are wider than one spectral frame
  - E.g. in the above example we note multiple examples of a pattern that spans several frames

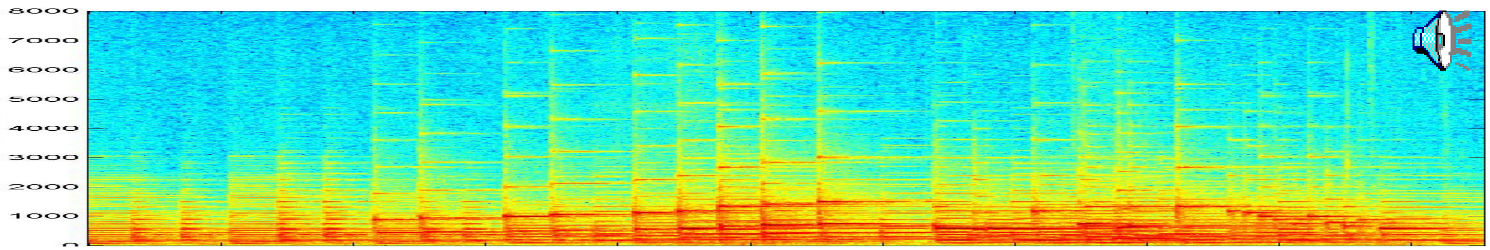
Next up: Shift/Transform Invariance



- Sometimes the “typical” structures that compose a sound are wider than one spectral frame
  - E.g. in the above example we note multiple examples of a pattern that spans several frames
- Multiframe patterns may also be local in frequency
  - E.g. the two green patches are similar only in the region enclosed by the blue box

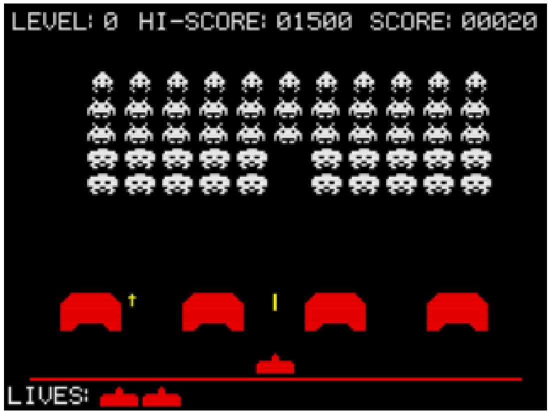
Patches are more representative than frames

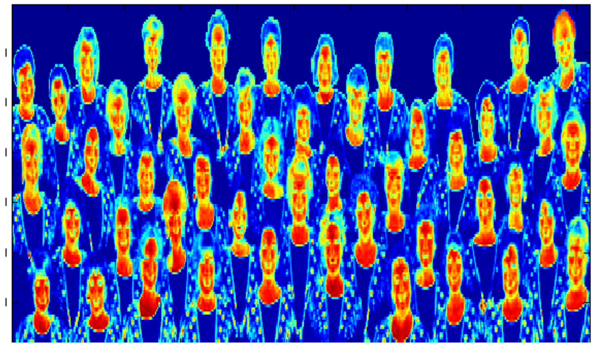




- Four bars from a music example
- The spectral patterns are actually patches
  - Not all frequencies fall off in time at the same rate
- The basic unit is a spectral patch, not a spectrum

Images: Patches often form the image



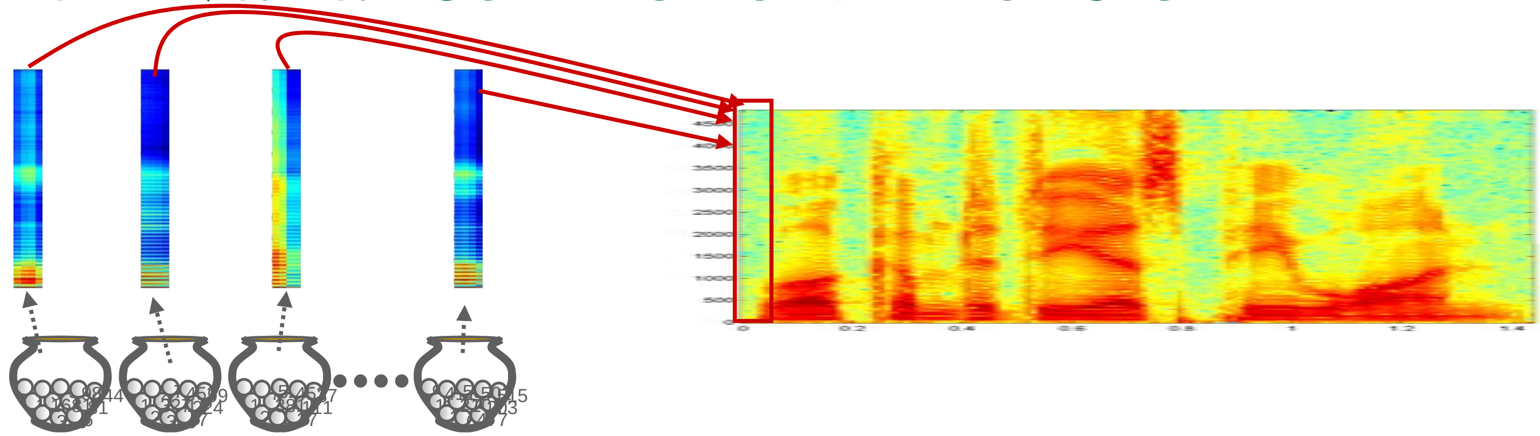


- A typical image component may be viewed as a patch
  - The alien invaders
  - Face like patches
  - A car like patch
    - overlaid on itself many times..

Shift-invariant modelling

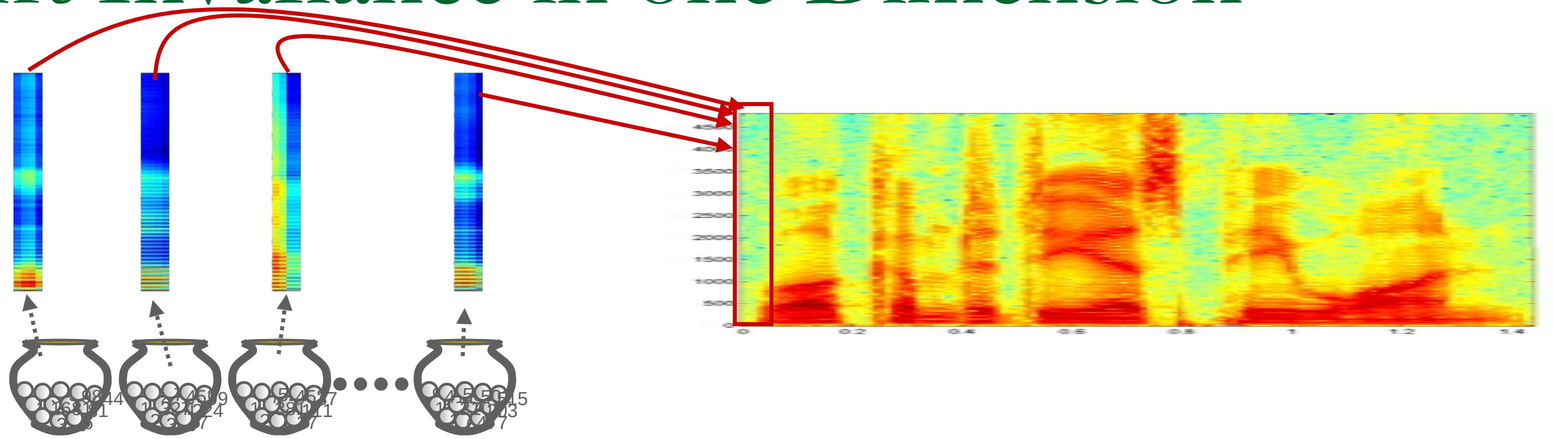
- A shift-invariant model permits individual bases to be *patches*
- Each patch composes the entire image.
- The data is a sum of the compositions from individual patches

Shift Invariance in one Dimension



- Our bases are now "patches"
  - Typical *spectro-temporal* structures
- The urns now represent patches
  - Each draw results in a (t,f) pair, rather than only f
  - *Also associated with each urn: A shift probability distribution P(T|z)*
- The overall drawing process is slightly more complex
- Repeat the following process:
  - Select an urn Z with a probability P(Z)
  - Draw a value T from P(t|Z)
  - Draw (t,f) pair from the urn
  - Add to the histogram at (t+T, f)

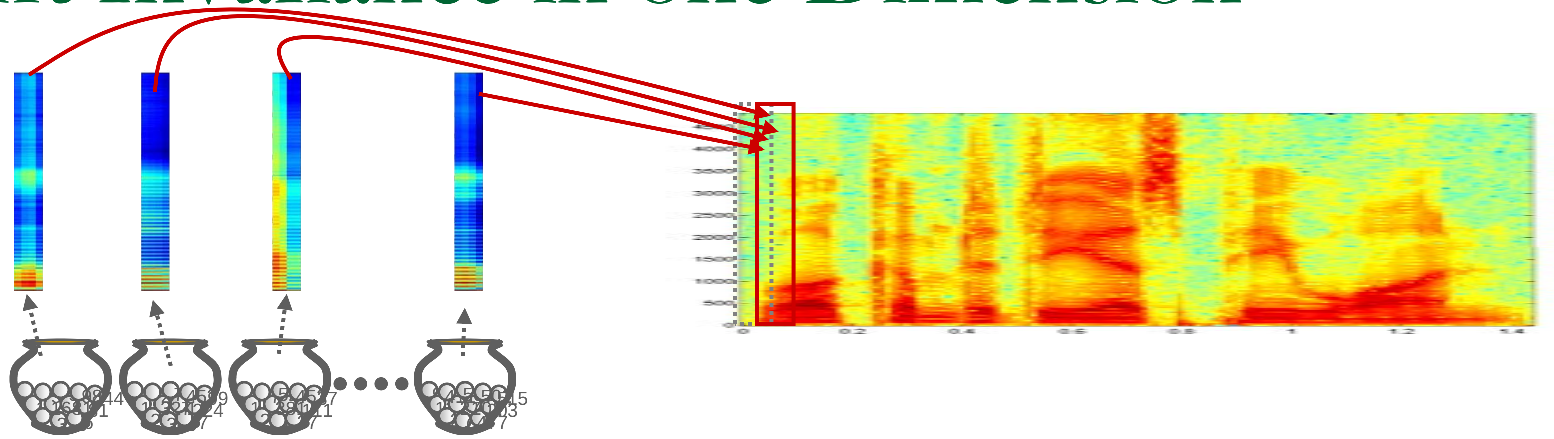
Shift Invariance in one Dimension



- The process is *shift-invariant* because the probability of drawing a shift P(T|Z) does not affect the probability of selecting urn Z

- Every location in the spectrogram has contributions from every urn patch

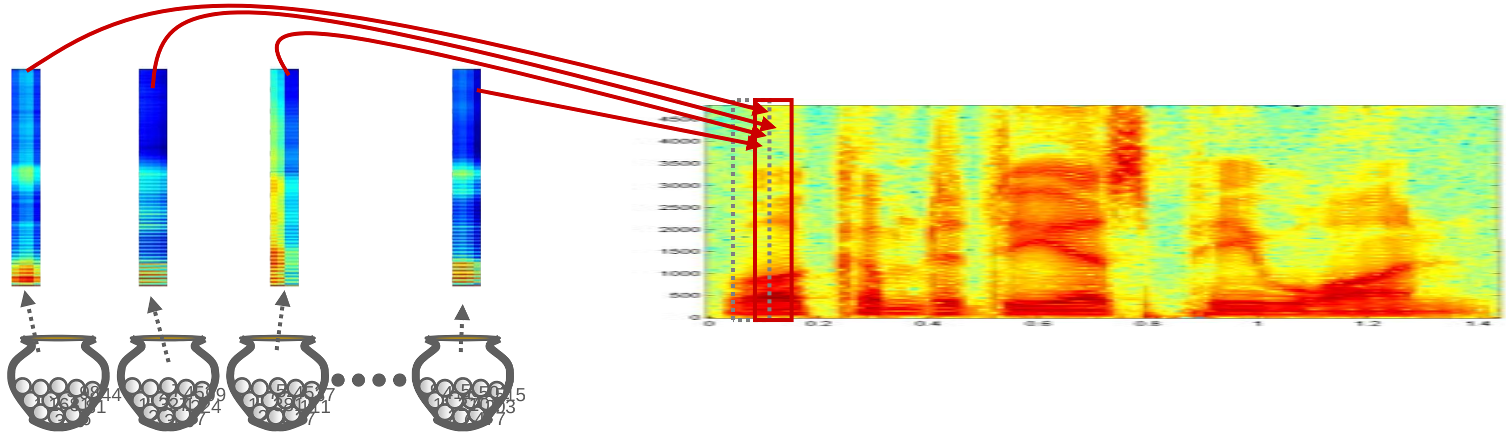
Shift Invariance in one Dimension



- The process is *shift-invariant* because the probability of drawing a shift P(T|Z) does not affect the probability of selecting urn Z

- Every location in the spectrogram has contributions from every urn patch

Shift Invariance in one Dimension



- The process is *shift-invariant* because the probability of drawing a shift P(T|Z) does not affect the probability of selecting urn Z

- Every location in the spectrogram has contributions from every urn patch

Probability of drawing a particular (t,f) combination

$$\boxed{P(t,f) = \sum_z P(z) \sum_\tau P(\tau \mid z) P(t-\tau, f \mid z)}$$

- The parameters of the model:
  - P(t,f|z) – the urns
  - P(T|z) – the *urn-specific* shift distribution
  - P(z) – probability of selecting an urn
- The ways in which (t,f) can be drawn:
  - Select any urn z
  - Draw T from the urn-specific shift distribution
  - Draw (t-T,f) from the urn
- The actual probability sums this over all shifts and urns

Learning the Model

- The parameters of the model are learned analogously to the manner in which mixture multinomials are learned

- Given observation of (t,f), it we knew which urn it came from and the shift, we could compute all probabilities by counting!
  - If shift is T and urn is Z
    - Count(Z) = Count(Z) + 1
    - For shift probability: Count(T|Z) = Count(T|Z)+1
    - For urn: Count(t-T,f | Z) = Count(t-T,f|Z) + 1
      - Since the value drawn from the urn was t-T,f

  - After all observations are counted:
    - Normalize Count(Z) to get P(Z)
    - Normalize Count(T|Z) to get P(T|Z)
    - Normalize Count(t,f|Z) to get P(t,f|Z)

- Problem: When learning the urns and shift distributions from a histogram, the urn (Z) and shift (T) for any draw of (t,f) is not known
  - These are unseen variables

Learning the Model

- Urn Z and shift T are unknown
  - So (t,f) contributes partial counts to *every* value of T and Z
  - Contributions are proportional to the *a posteriori* probability of Z and T,Z

$$P(t,f,Z) = P(Z)\sum_{T} P(T \mid Z)P(t-T,f \mid Z) \qquad P(T,t,f \mid Z) = P(T \mid Z)P(t-T,f \mid Z)$$

$$P(Z \mid t,f) = \frac{P(t,f,Z)}{\sum_{Z'} P(t,f,Z')} \qquad P(T \mid Z,t,f) = \frac{P(T,t-T,f \mid Z)}{\sum_{T'} P(T',t-T',f \mid Z)}$$

- Each observation of (t,f)
  - P(z|t,f) to the count of the total number of draws from the urn
    - Count(Z) = Count(Z) + P(z | t,f)
  - P(z|t,f)P(T | z,t,f) to the count of the shift T for the shift distribution
    - Count(T | Z) = Count(T | Z) + P(z|t,f)P(T | Z, t, f)
  - P(z|t,f)P(T | z,t,f) to the count of (t-T, f) for the urn
    - Count(t-T,f | Z) = Count(t-T,f | Z) + P(z|t,f)P(T | z,t,f)

Shift invariant model: Update Rules

- Given data (spectrogram) S(t,f)
- Initialize P(Z), P(T|Z), P(t,f | Z)
- Iterate

$$P(t,f,Z) = P(Z)\sum_T P(T|Z)P(t-T,f|Z)$$

$$P(T,t,f|Z) = P(T|Z)P(t-T,f|Z)$$

$$P(Z|t,f) = \frac{P(t,f,Z)}{\sum_{Z'} P(t,f,Z')}$$

$$P(T|Z,t,f) = \frac{P(T,t-T,f|Z)}{\sum_{T'} P(T',t-T',f|Z)}$$

$$P(Z) = \frac{\sum_t \sum_f P(Z|t,f)S(t,f)}{\sum_{Z'} \sum_t \sum_f P(Z'|t,f)S(t,f)}$$

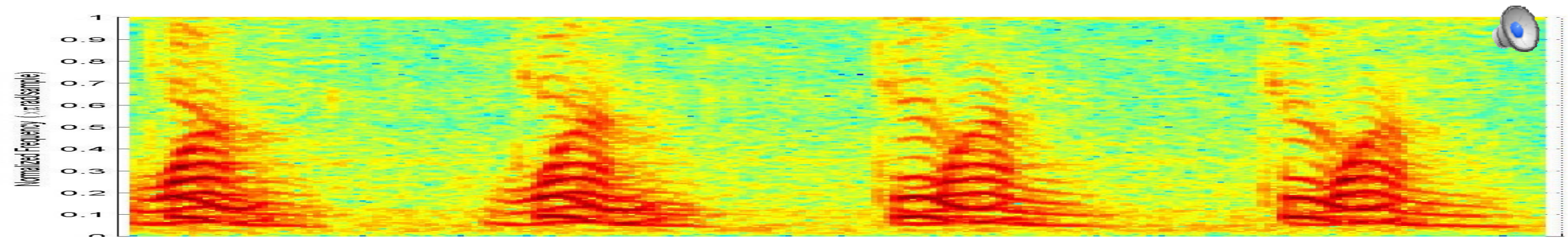
$$P(T|Z) = \frac{\sum_t \sum_f P(Z|t,f)P(T|Z,t,f)S(t,f)}{\sum_{T'} \sum_t \sum_f P(Z|t,f)P(T'|Z,t,f)S(t,f)}$$

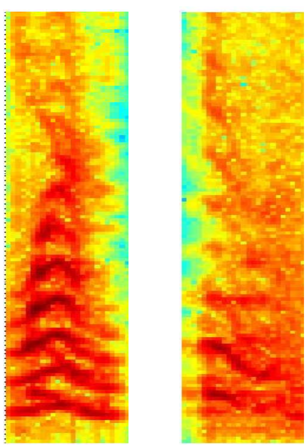
$$P(t,f|Z) = \frac{\sum_T P(Z|T,f)P(T-t|Z,T,f)S(T,f)}{\sum_{t'} \sum_T P(Z|T,f)P(T-t'|Z,T,f)S(T,f)}$$

Shift-invariance in one time: example

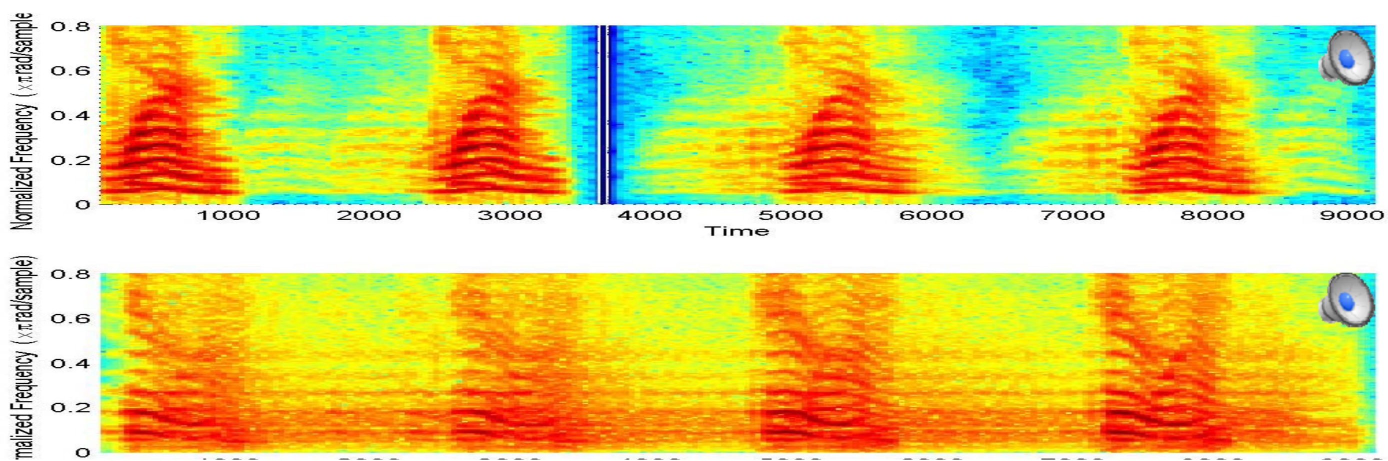
- An Example: Two distinct sounds occuring with different repetition rates within a signal
  - Modelled as being composed from two time-frequency bases
  - NOTE: Width of patches must be specified

INPUT SPECTROGRAM



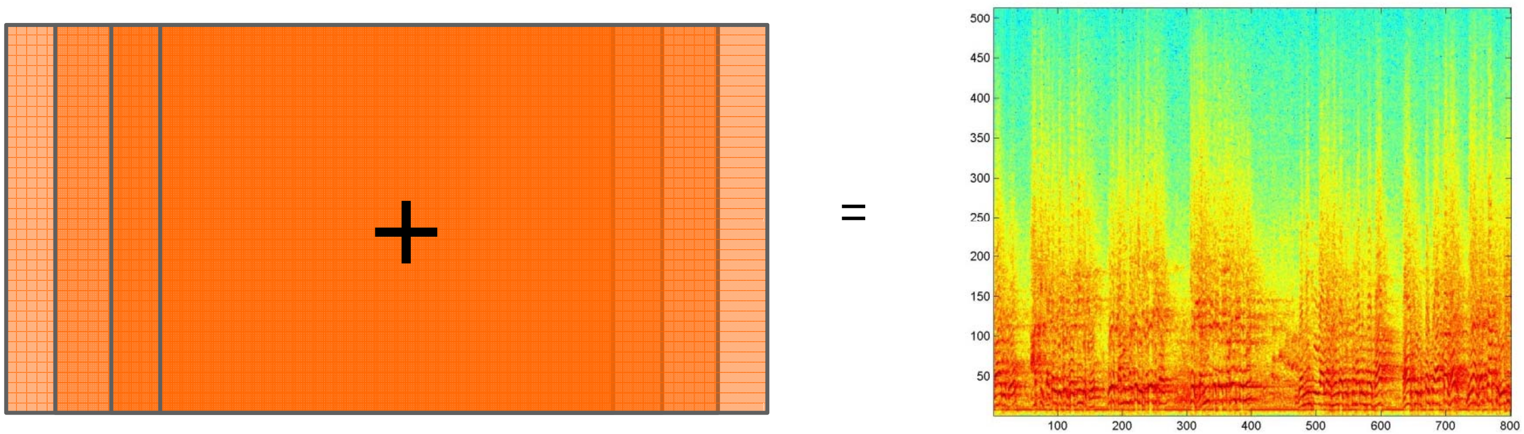


Discovered time-frequency “patch” bases (urns)



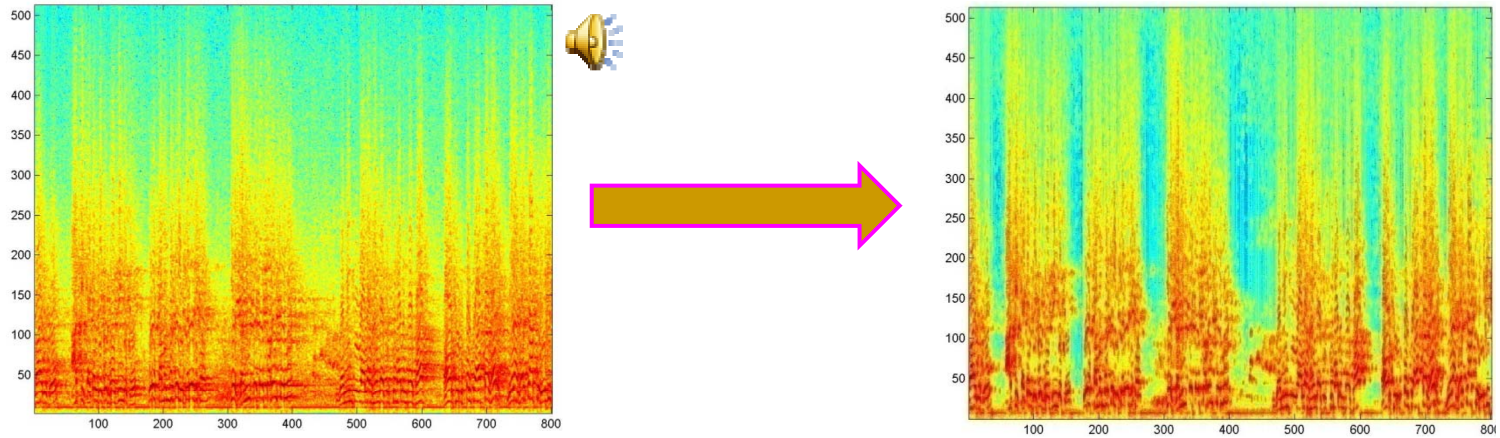
Contribution of individual bases to the recording

Shift Invariance in Time: Dereverberation



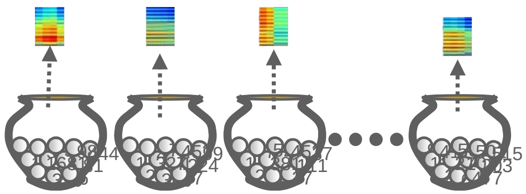
- Reverberation – a simple model
  - The Spectrogram of the reverberated signal is a sum of the spectrogram of the clean signal and several shifted and scaled versions of itself
  - A convolution of the spectrogram and a room response

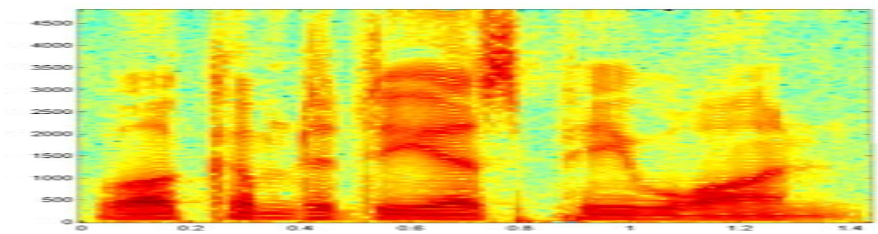
Dereverberation



- Given the spectrogram of the reverberated signal:
  - Learn a shift-invariant model with a single patch basis
    - Sparsity must be enforced on the basis
  - The "basis" represents the clean speech!

Shift Invariance in Two Dimensions





- We now have urn-specific shifts along both T and F
- The Drawing Process
  - Select an urn Z with a probability P(Z)
  - Draw SHIFT values (T,F) from $P_s(T,F|Z)$
  - Draw (t,f) pair from the urn
  - Add to the histogram at (t+T, f+F)

- This is a two-dimensional shift-invariant model
  - We have shifts in both time and frequency
    - Or, more generically, along both axes

Learning the Model

- Learning is analogous to the 1-D case

- Given observation of (t,f), it we knew which urn it came from and the shift, we could compute all probabilities by counting!
  - If shift is T,F and urn is Z
    - Count(Z) = Count(Z) + 1
    - For shift probability: ShiftCount(T,F|Z) = ShiftCount(T,F|Z)+1
    - For urn: Count(t-T,f-F | Z) = Count(t-T,f-F|Z) + 1
      - Since the value drawn from the urn was t-T,f-F

  - After all observations are counted:
    - Normalize Count(Z) to get P(Z)
    - Normalize ShiftCount(T,F|Z) to get P_s(T,F|Z)
    - Normalize Count(t,f|Z) to get P(t,f|Z)

- Problem: Shift and Urn are unknown

Learning the Model

- Urn Z and shift T,F are unknown
  - So (t,f) contributes partial counts to *every* value of T,F and Z
  - Contributions are proportional to the *a posteriori* probability of Z and T,F|Z

$$P(t,f,Z) = P(Z)\sum_{T,F} P(T,F \mid Z)P(t-T, f-F \mid Z) \qquad P(T,F,t,f \mid Z) = P(T,F \mid Z)P(t-T, f-F \mid Z)$$

$$P(Z \mid t,f) = \frac{P(t,f,Z)}{\sum_{Z'} P(t,f,Z')} \qquad P(T,F \mid Z,t,f) = \frac{P(T,F,t-T,f-F \mid Z)}{\sum_{T',F'} P(T',F',t-T',f-F' \mid Z)}$$

- Each observation of (t,f)
  - P(z|t,f) to the count of the total number of draws from the urn
    - Count(Z) = Count(Z) + P(z | t,f)
  - P(z|t,f)P(T,F | z,t,f) to the count of the shift T,F for the shift distribution
    - ShiftCount(T,F | Z) = ShiftCount(T,F | Z) + P(z|t,f)P(T | Z, t, f)
  - P(T | z,t,f) to the count of (t-T, f-F) for the urn
    - Count(t-T,f-F | Z) = Count(t-T,f-F | Z) + P(z|t,f)P(t-T,f-F | z,t,f)

Shift invariant model: Update Rules

- Given data (spectrogram) S(t,f)
- Initialize P(Z), P_s(T,F|Z), P(t,f | Z)
- Iterate

$$P(t,f,Z) = P(Z)\sum_{T,F} P(T,F \mid Z)P(t-T, f-F \mid Z) \qquad P(T,F,t,f \mid Z) = P(T,F \mid Z)P(t-T, f-F \mid Z)$$

$$P(Z \mid t,f) = \frac{P(t,f,Z)}{\sum_{Z'} P(t,f,Z')} \qquad P(T,F \mid Z,t,f) = \frac{P(T,F,t-T,f-F \mid Z)}{\sum_{T',F'} P(T',F',t-T',f-F' \mid Z)}$$

$$P(Z) = \frac{\sum_t \sum_f P(Z \mid t,f)S(t,f)}{\sum_{Z'}\sum_t \sum_f P(Z' \mid t,f)S(t,f)} \qquad P(T,F \mid Z) = \frac{\sum_t \sum_f P(Z \mid t,f)P(T,F \mid Z,t,f)S(t,f)}{\sum_{T'}\sum_{F'}\sum_t \sum_f P(Z \mid t,f)P(T',F' \mid Z,t,f)S(t,f)}$$

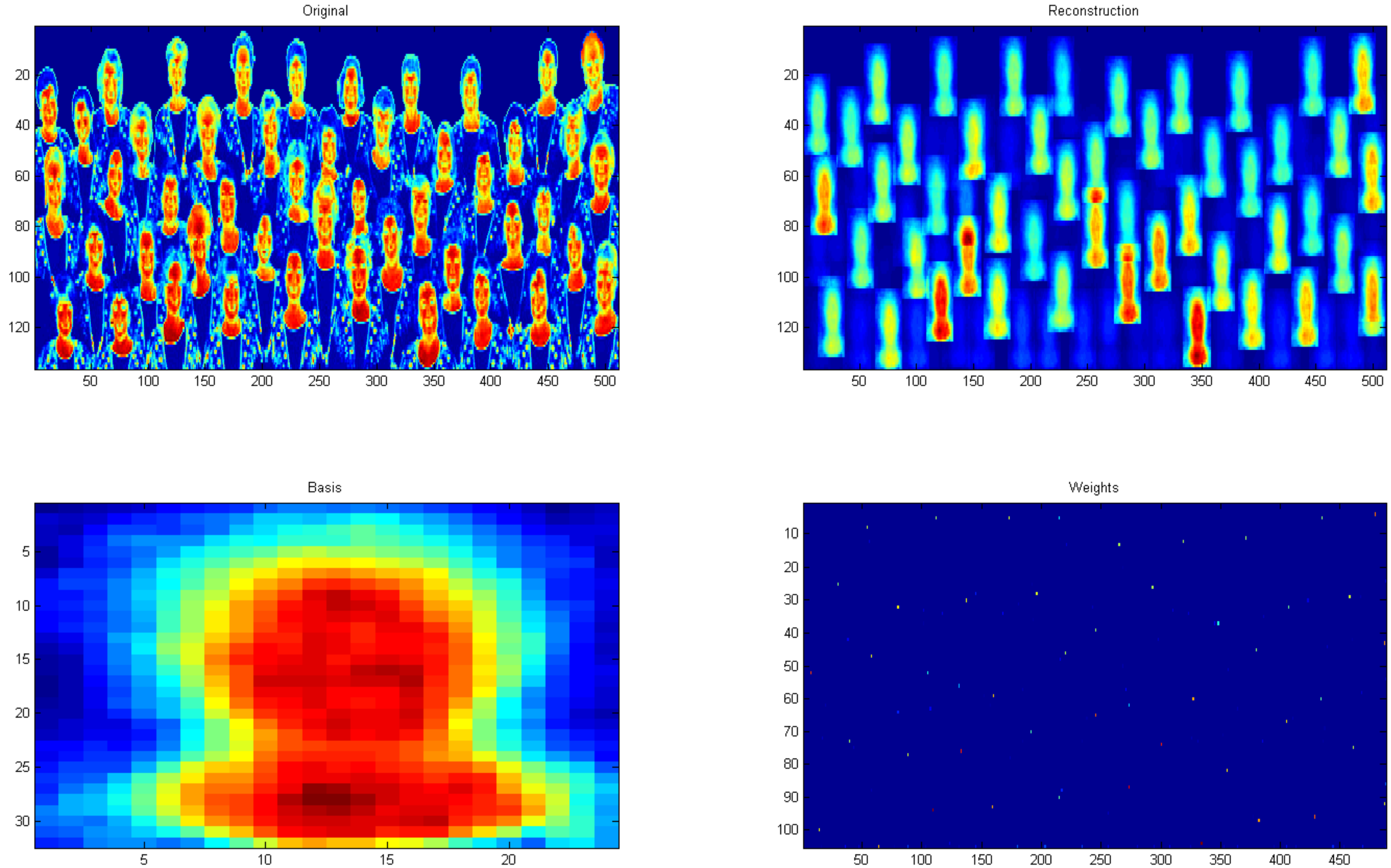
$$P(t,f \mid Z) = \frac{\sum_{T,F} P(Z \mid T,F)P(T-t,F-f \mid Z,T,F)S(T,F)}{\sum_{t',f'}\sum_{T,F} P(Z \mid T,F)P(T-t',F-f' \mid Z,T,F)S(T,F)}$$

2D Shift Invariance: The problem of indeterminacy

- $P(t,f|Z)$ and $P_s(T,F|Z)$ are analogous
  - Difficult to specify which will be the "urn" and which the "shift"

- Additional constraints required to ensure that one of them is clearly the shift and the other the urn
- Typical solution: Enforce sparsity on $P_s(T,F|Z)$
  - The patch represented by the urn occurs only in a few locations in the data

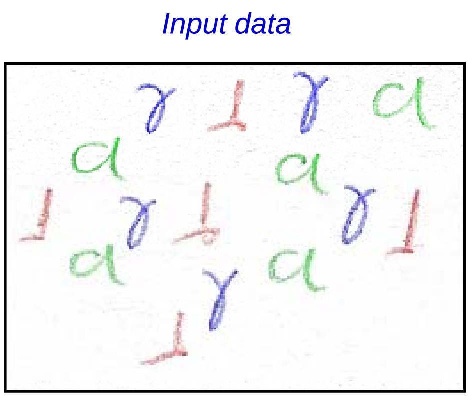
Example: 2-D shift invariance

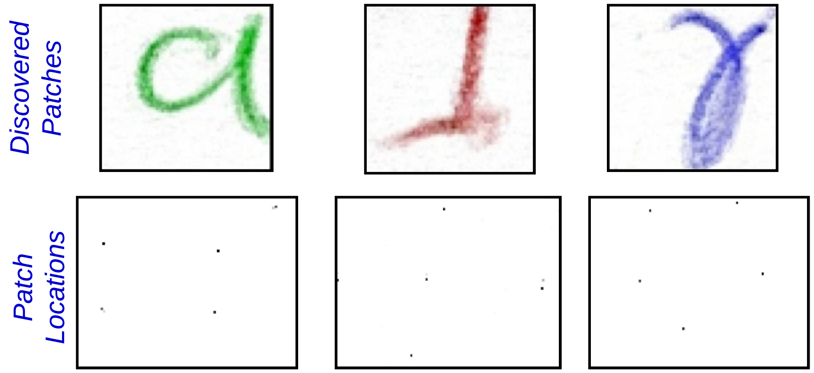


- Only one "patch" used to model the image (i.e. a single urn)
  - The learnt urn is an "average" face, the learned shifts show the locations of faces

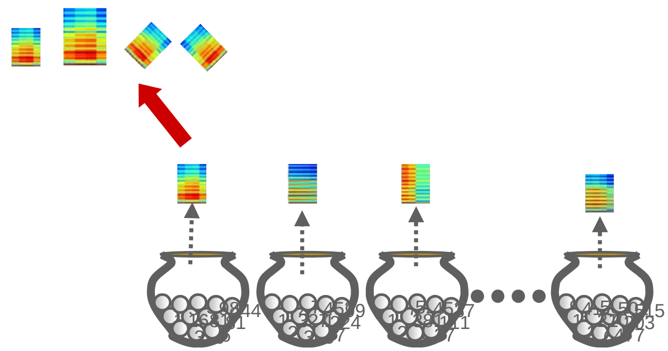
Example: 2-D shift invarince

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





Beyond shift-invariance: transform invariance



- The draws from the urns may not only be shifted, but also transformed
- The arithmetic remains very similar to the shift-invariant model
  - We must now impose one of an enumerated set of transforms to (t,f), after shifting them by (T,F)
  - In the estimation, the precise transform applied is an unseen variable

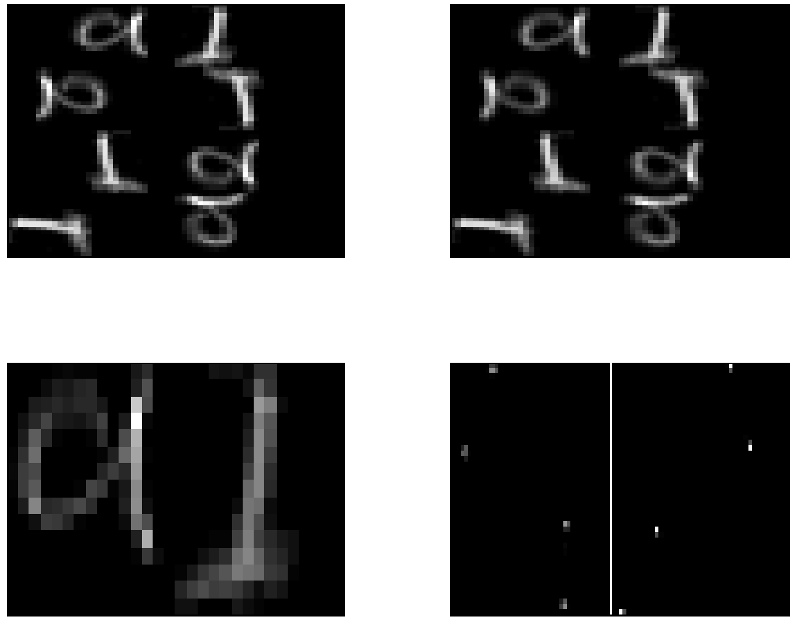
Transform invariance: Generation

- The set of transforms is enumerable
  - E.g. scaling by 0.9, scaling by 1.1, rotation right by 90degrees, rotation left by 90 degrees, rotation by 180 degrees, reflection
  - Transformations can be chosen by draws from a distribution over transforms
    - E.g. P(rotation by 90 degrees) = 0.2..
    - Distributions are URN SPECIFIC
- The drawing process:
  - Select an urn Z (patch)
  - Select a shift (T,F) from P_s(T, F| Z)
  - Select a transform from P(txfm | Z)
  - Select a (t,f) pair from P(t,f | Z)
  - *Transform* (t,f) to txfm(t,f)
  - Increment the histogram at txfm(t,f) + (T,F)

Transform invariance

- The learning algorithm must now estimate
  - P(Z) – probability of selecting urn/patch in any draw
  - P(t,f|Z) – the urns / patches
  - P(txfm | Z) – the urn specific distribution over transforms
  - P_s(T,F|Z) – the urn-specific shift distribution
- Essentially determines what the basic shapes are, where they occur in the data and how they are transformed
- The mathematics for learning are similar to the maths for shift invariance
  - With the addition that each instance of a draw must be fractured into urns, shifts AND transforms
- Details of learning are left as an exercise
  - Alternately, refer to Madhusudana Shashanka's PhD thesis at BU

Example: Transform Invariance



- Top left: Original figure
- Bottom left – the two bases discovered
- Bottom right –
  - Left panel, positions of “a”
  - Right panel, positions of “l”
- Top right: estimated distribution underlying original figure

Transform invariance: model limitations and extensions

- The current model only allows *one* transform to be applied at any draw
  - E.g. a basis may be rotated or scaled, but not scaled *and* rotated
- An obvious extension is to permit combinations of transformations
  - Model must be extended to draw the combination from some distribution
- Data dimensionality: All examples so far assume only *two* dimensions (e.g. in spectrogram or image)
- The models are trivially extended to higher-dimensional data

Transform Invariance: Uses and Limitations

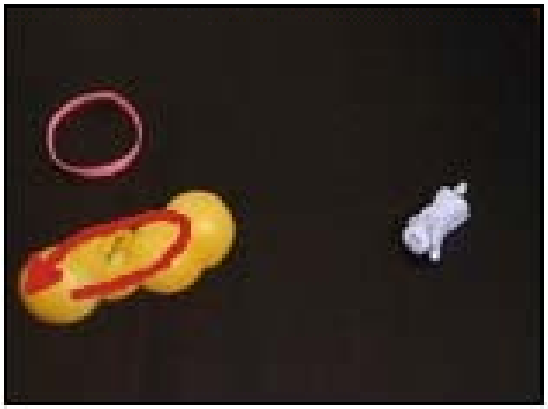
- Not very useful to analyze audio
- May be used to analyze images and video

- Main restriction: Computational complexity
  - Requires unreasonable amounts of memory and CPU
  - Efficient implementation an open issue

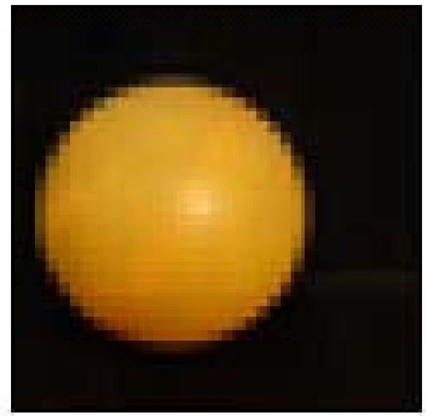
Example: Higher dimensional data

- Video example

Description of Input



Kernel 1



Kernel 2

Kernel 3