Learning the parts and features of speech?

Lawrence Saul
Computer & Information Science
University of Pennsylvania

Message for experts

This talk is not for you.

(But stay anyway for the Karaoke.)

Overview

- **Learning**

  What can automatic algorithms infer from large data sets of sensory inputs?

- **Advances**

  Learning the parts of objects.
  Sparse nonnegative decompositions.

- **Applications**

  Multiple f_0 estimation.
  Robust feature detection?

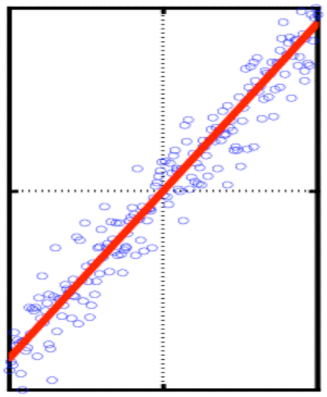
Exploratory data analysis

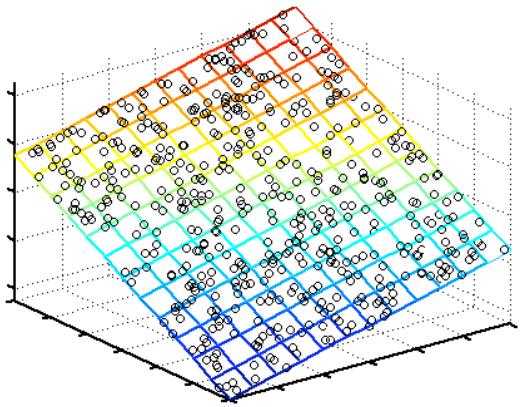
- **Dimensionality reduction**

  How to discover low dimensional structure in high dimensional data?

- **Subspace methods**

  Compute maximum variance subspace.
  Project data into subspace.





Learning parts of objects

- **Objects**

  Represent objects (such as images or speech) by high dimensional vectors.

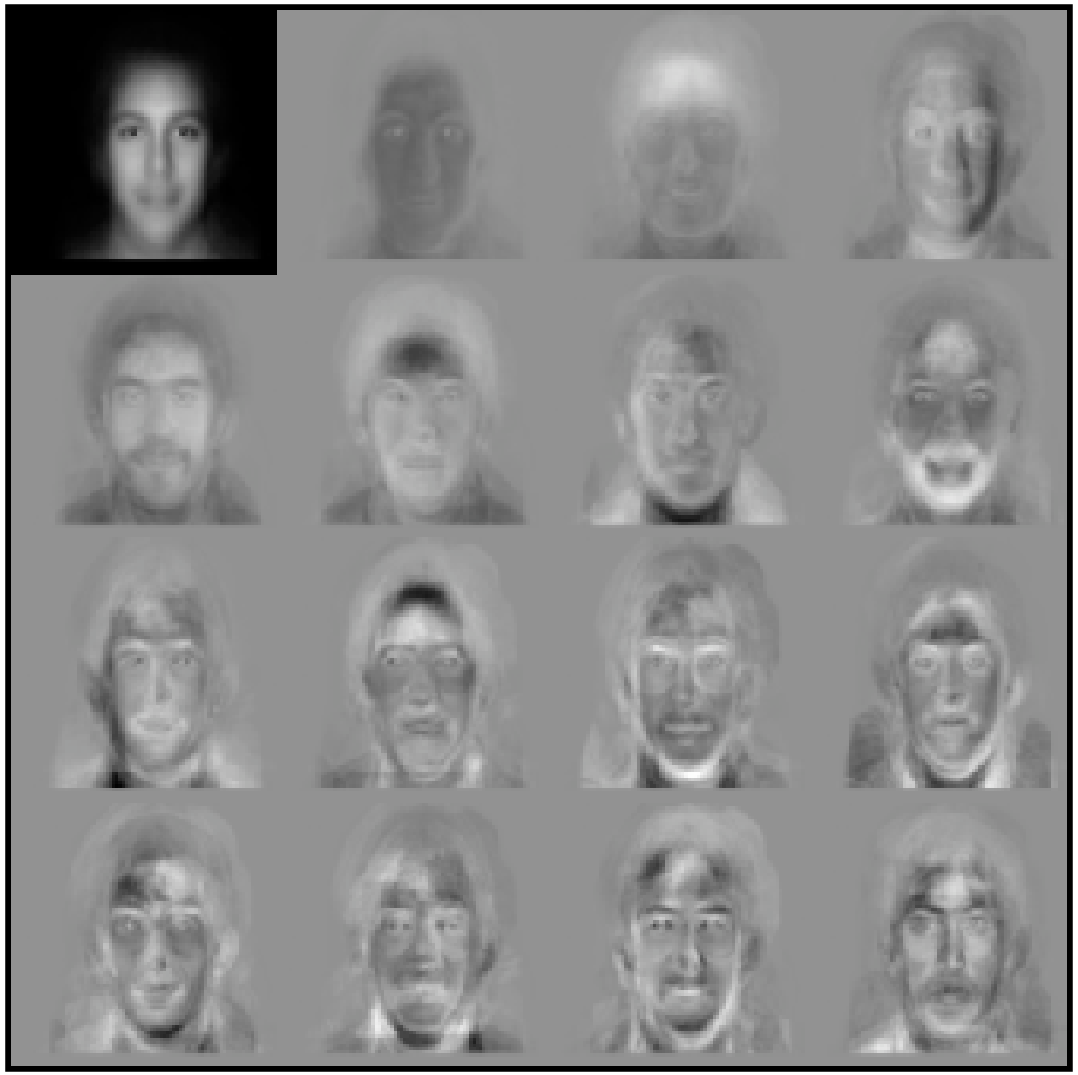
- **Parts**

  Represent parts by basis vectors of maximum variance subspace.

- **Model**

  Model objects as weighted sums of parts. Is this meaningful?

Parts of faces?



eigenfaces from 7562 images

(MIT Media Lab)

Hard to interpret as parts…

Nonnegativity

- **What if data is nonnegative?**

  Ex: pixels of images,
  power spectra of speech.

- **Dimensionality reduction**

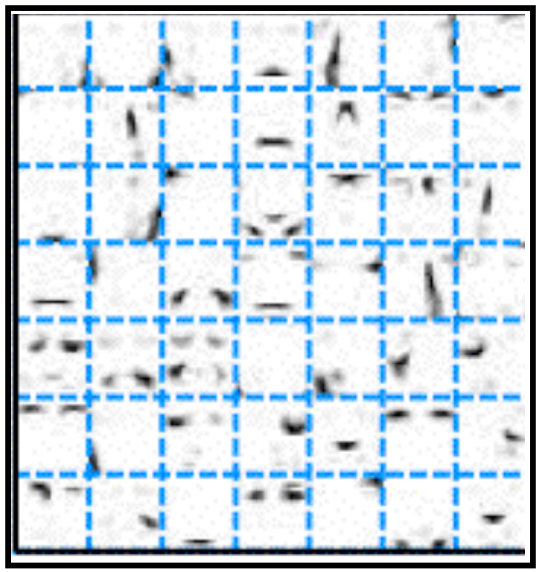
  Data lives in high dimensional orthant.
  Project into low dimensional cone.

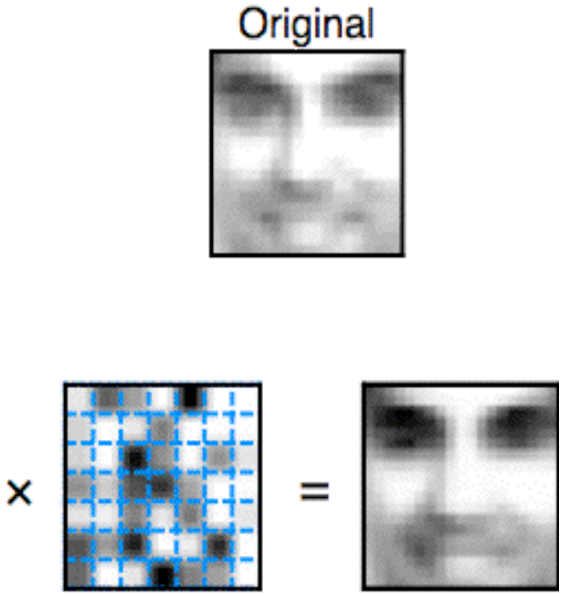
- **Constraints**

  Nonnegative basis vectors.
  Only constructive combinations.

Nonnegative parts of faces

Lee & Seung (1999)



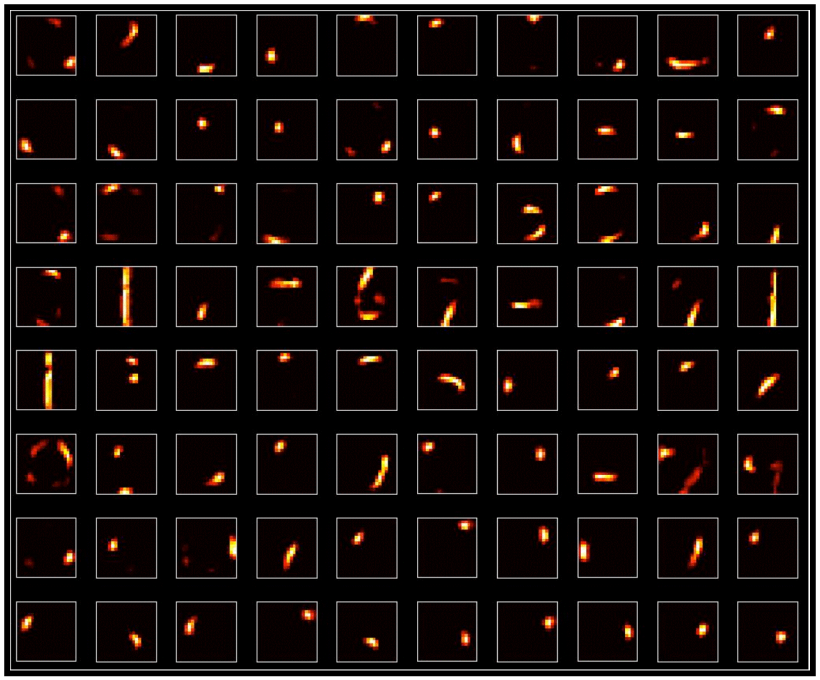


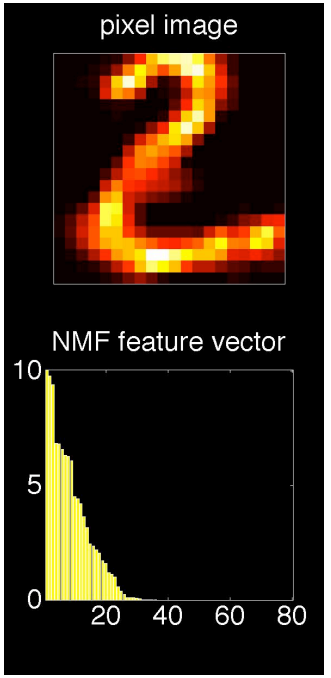
Parts resemble eyes, mouth, mustache, ...
Basis yields sparse, distributed encodings.

("Mr. Potato Head" Model)

Nonnegative parts of digits

Saul & Lee (2001)





Parts resemble cursive strokes.
Digits are modeled as sums of strokes.

Analysis by synthesis

- **To recognize a face:**

  Which eyes, mouth, nose yield best match to observed face?

- **To recognize a digit:**

  Which cursive strokes are composed to draw the digit?

- **Nonnegative deconvolution**

  Infer parts from object and basis vectors. Robust to missing parts?

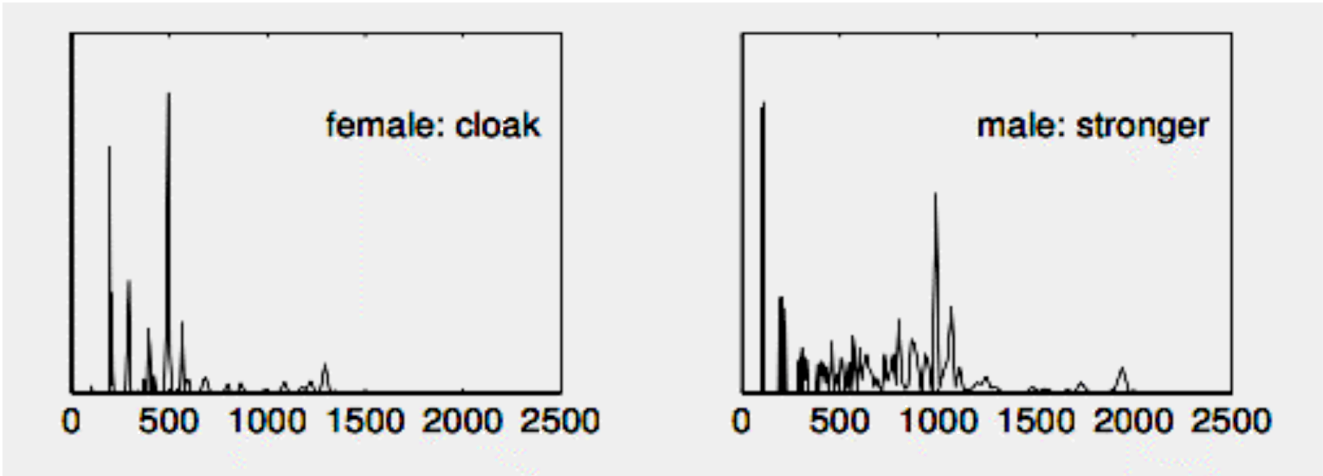
Auditory scene analysis

- **Parts as sources**

  View auditory scene as complex object.
  Constituent parts are individual sources.

- **Start simple: periodic sources**

  Basis vectors are harmonic stacks.
  Mixed signal is nonnegative superposition.



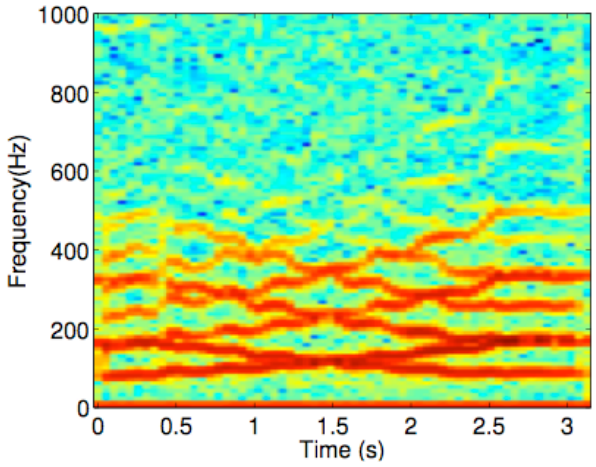
Nonnegative deconvolution for estimating multiple f_0

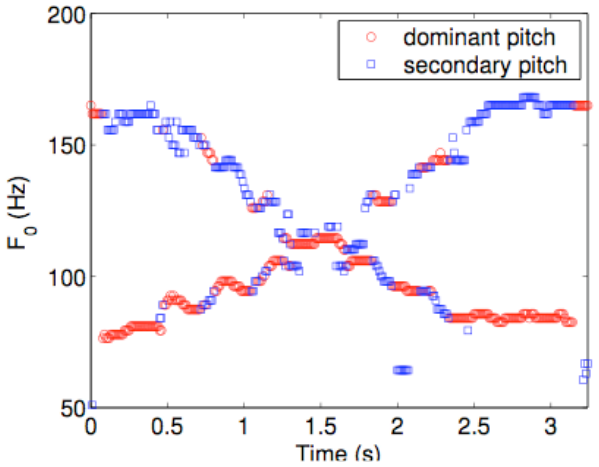
- **Polyphonic music**

  Smaragdis & Brown (2003)
  Abdallah & Plumbley (2003)

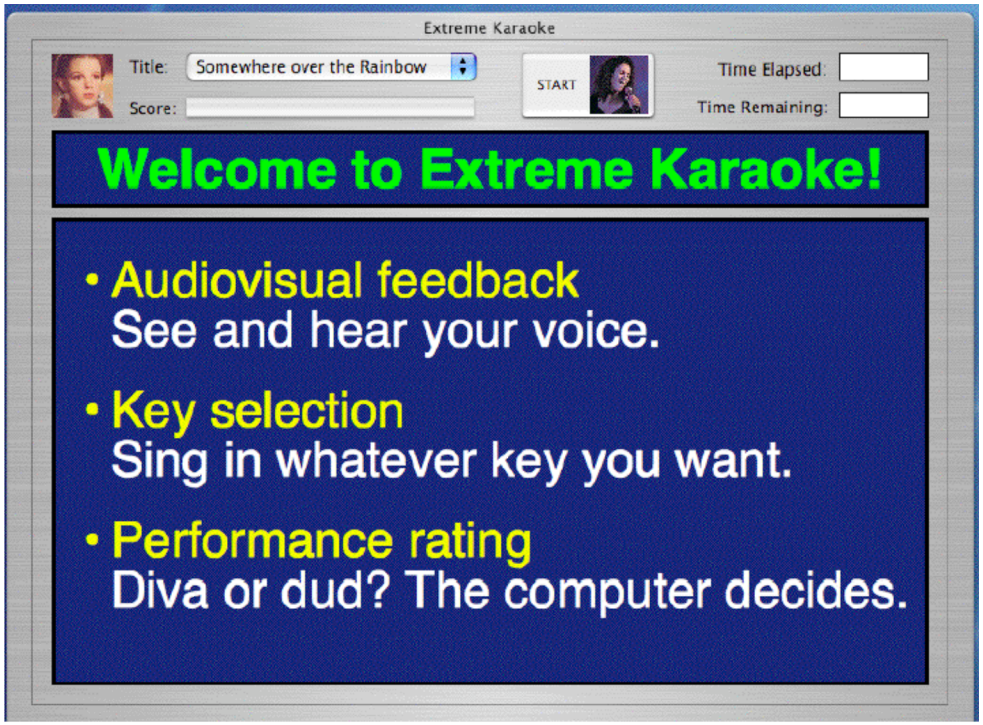
- **Overlapping voices**

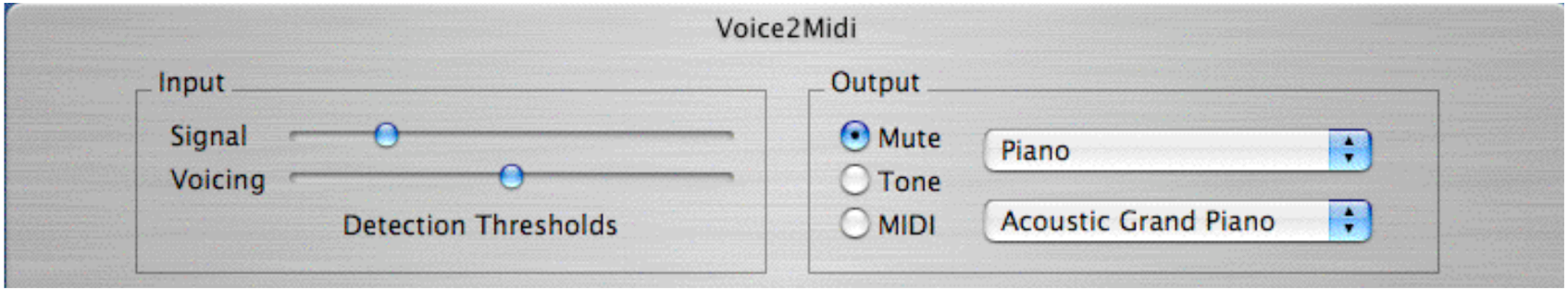
  Goto (2000), Sha & Saul (2004)





Real-time applications





Robust features for ASR

- **Parts as features**

  face = eyes + nose + mouth

  digit = sum of strokes

  phoneme = ? + ? + ?

- **Parts of speech**

  Will nonnegative decompositions yield localized patterns in time-frequency?

- **Robustness**

  Glimpses, missing data, binary masks, multiband ASR: compatible with above?

Conclusion

- **Dimensionality reduction**

  Learn parts of objects by projecting to lower dimensional spaces.

- **Nonnegativity constraints**

  Nonnegative decompositions can lead to more interpretable parts.

- **Other work**

  Nonlinear projections, manifolds, and continuous modes of variability.