
Semantic Audio Analysis

- 1 Semantic Audio Analysis
- 2 Organizing Sound Mixtures
- 3 Applications for Audio Semantics
- 4 Open Questions

Dan Ellis <dpwe@ee.columbia.edu>

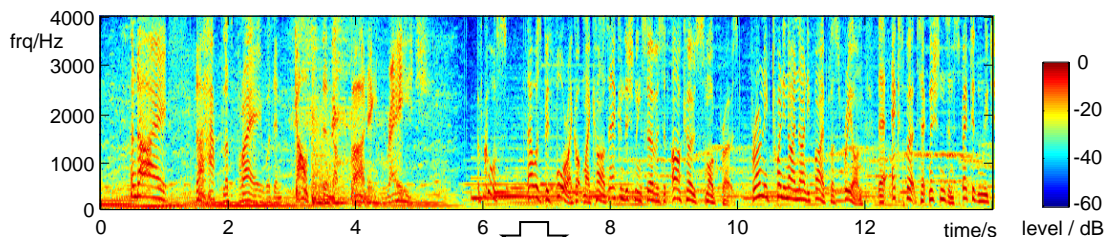
Laboratory for Recognition and Organization of Speech and Audio
(Lab**ROSA**)
Columbia University, New York
<http://labrosa.ee.columbia.edu/>



1

Semantic Audio Analysis

- **Audio Semantics**
= what is the **meaning** / message?
 - “AI complete”?
- **“Semantics” is broad!**
 - used for the-stuff-we-can’t-do-yet
- **How about speech recognition?**



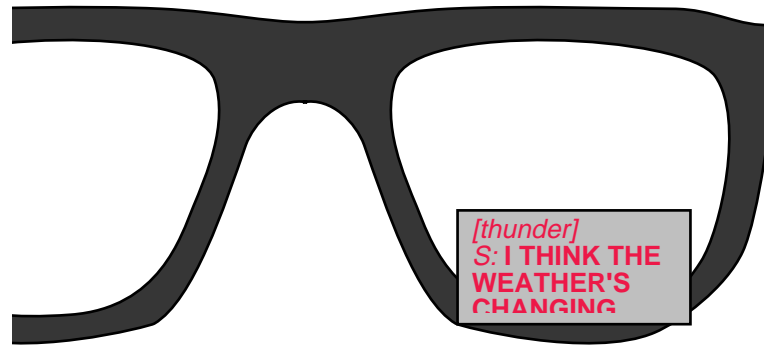
IT'S NICE TO HAVE THIS FRIDAY WAS IN DOLLARS IN THE BOMBING RAIDS WAS CLOSING THOUSAND TO FOUR GUNMEN CAUSING **CONDITIONING** SAID THE STOCK ROSE SMOKERS FROM THE **TWENTY NINE NINETY FIVE PLUS** THREE ON THE **EXPERT TECHNICIANS** EXPECTED TO REACH

- no good even if it worked!



Towards Semantic Audio Analysis

- **What do we want from SAA?**
 - describe sound in **human-recognizable terms**
 - “automatic subtitles for real life”?

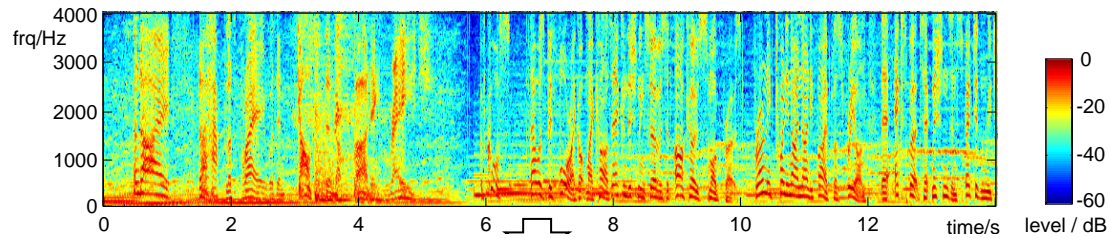


- **Key step in speech recognition is classification**
 - convert continuous signal into **discrete classes**
 - need a rich, application-dependent vocabulary
 - **hierarchy** of pattern recognition
- **Listeners perceive sound sources**
 - If SAA primitives are to be **subjective percepts...**
 - ...**source segregation** is the first problem?



SAA Applications

- **Subjective descriptions are the ultimate sound representation**



- data **compression**: “Radio ad for AARCO”
 - signal **enhancement**: “... without noise”
 - **modification**: “.. woman’s voice ..”
 - .. needs both analysis **and synthesis**?
-
- **Sound understanding useful for:**
 - indexing/retrieval
 - robots
 - prostheses?



Outline

- 1 Semantic Audio Analysis
- 2 **Organizing Sound Mixtures**
 - Human Auditory Scene Analysis
 - Organizing Mixtures by Computer
- 3 Applications for Audio Semantics
- 4 Open Questions

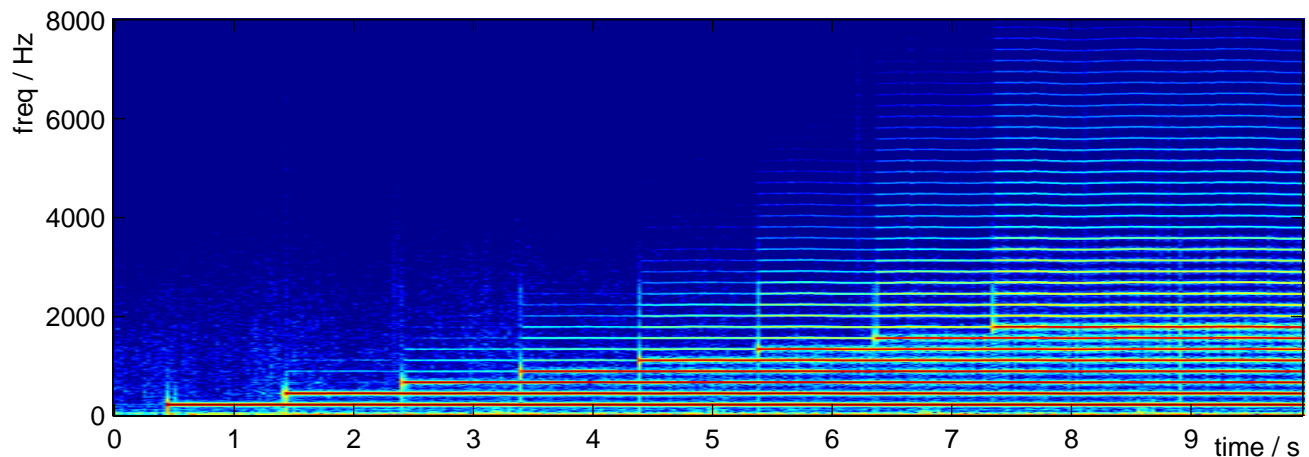


2

Auditory Scene Analysis

(Bregman 1990)

- **How do people analyze sound mixtures?**
 - break mixture into small **elements** (in time-freq)
 - elements are **grouped** in to sources using **cues**
 - sources have aggregate **attributes**
- **Elements + attributes**



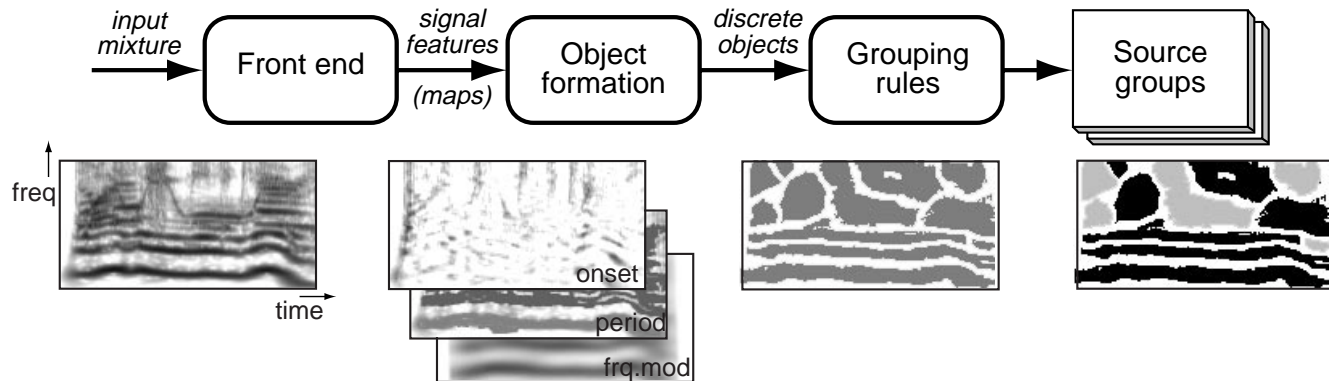
- common onset (= dependent origins)
- periodicity (= single process)
- + spatial cues etc. + familiarity, context ...



Computational Auditory Scene Analysis: The Representational Approach

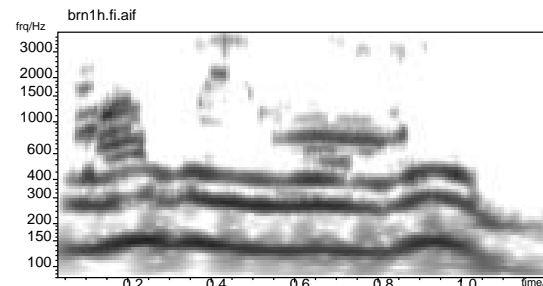
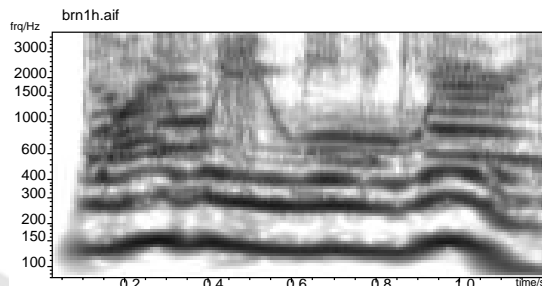
(Cooke & Brown 1993)

- Direct implementation of **psych. theory**



- 'bottom-up' processing
- uses common onset & periodicity cues

- Able to extract **voiced speech**:



Approaches to handling sound mixtures

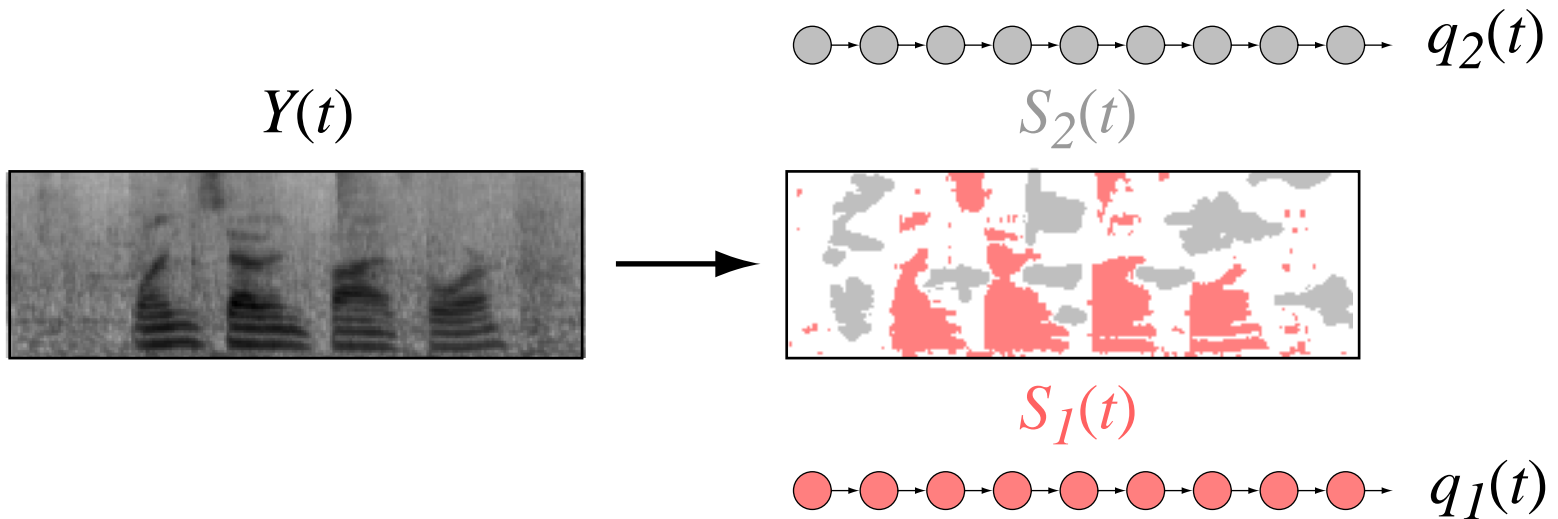
- **Separate **signals**, then recognize**
 - Computational Auditory Scene Analysis (CASA), Independent Component Analysis
 - nice, if you can make it work
- **Recognize **combined** signal**
 - ‘multicondition training’
 - combinatorics seem daunting
- **Recognize with **parallel models****
 - optimal **inference** from full joint state-space
$$p(O, x, y) \rightarrow p(x, y | O)$$
 - or: skip obscured fragments, **infer** from higher-level context
 - or do both: **missing-data recognition**



Multi-source decoding

(Barker, Cooke & Ellis 2003)

- **Missing Data recognizes from a subset;**
Can search for more than one source



- **Mutually-dependent data masks**
- **Use e.g. CASA features to propose masks**
 - locally coherent regions
- **Issues in models, representations, inference...**



Outline

- 1 Semantic Audio Analysis
- 2 Organizing Sound Mixtures
- 3 **Applications for Audio Semantics**
 - Meeting recordings
 - Audio diary analysis
 - Semantics of musical signals
- 4 Open Questions



3

The Meeting Recorder Project

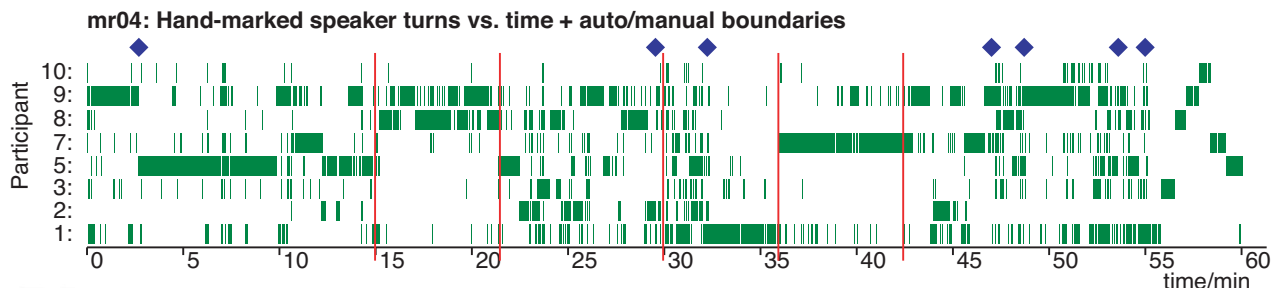
(with ICSI, UW, IDIAP, SRI, Sheffield)

- **Microphones in conventional meetings**



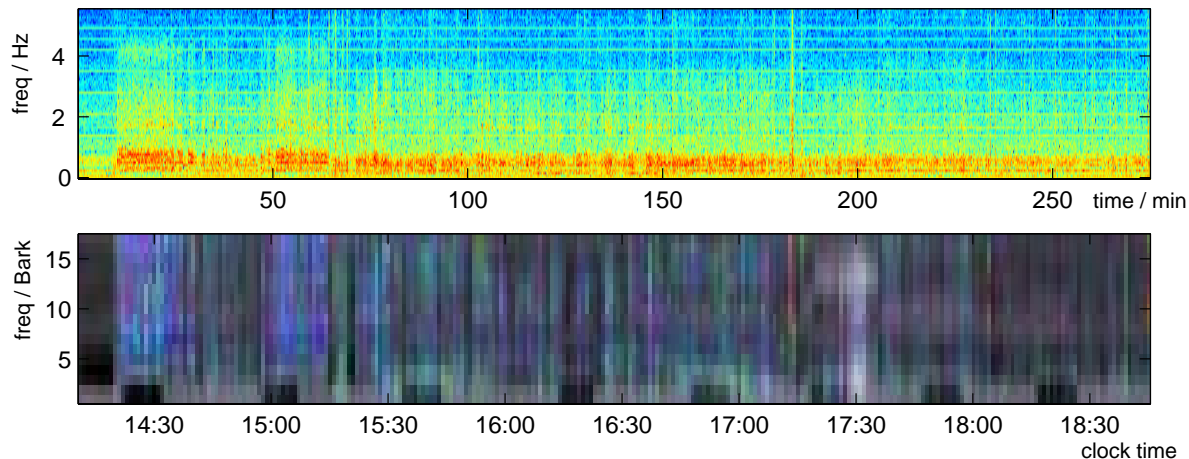
- for summarization / retrieval / behavior analysis
- informal, overlapped speech (→ ASR...)

- **Behavioral: Look for patterns of speaker turns**



Personal Audio: The Listening Machine

- **Smart PDA** records everything you hear
- **Only useful if we have index, summaries**
 - semantic descriptions (real time?)
- **Features appropriate for 1 minute segments...**



- Bark band variance, spectral entropy, ...



Music Information Retrieval

(Berenzweig & Ellis 2003)

- **Apply search concepts to music?**
 - “musical Google” – beat human annotation?
 - application: finding new music
- **Construct music space where near = “similar”**

The screenshot shows the Playola website interface. At the top, there is a search bar with the text "Search:" and a dropdown menu for "Artist". Below the search bar, there are several links: [About], [Help], [Turn Samples Off], [Turn Debug On], [Turn Popups Off], and [Logout dpwe].

Below the search bar, there is a section for "Get Playola Selections:" with a dropdown menu for "20 songs", a dropdown menu for "you recently heard", and a "Go!" button. To the right, there are links for "Browse: Artists", "Albums", "Playlists", and a "Range:" dropdown menu set to "0-C".

The main content area shows the artist "The Woodbury Muffin Outbreak" with a link to their "band web page" and a "[Play!]" button. Below this, there is a "Playlist:" dropdown menu set to "-New Playlist-" and links for "[Add to]" and "[View]".

The main content area is divided into two columns. The left column is a table with the following columns: "Song Title", "Artist", "Time", and "Rating". The right column is a "Music-Space Browser" with a "Feature" column and a "Less" to "More" scale for each feature.

Song Title	Artist	Time	Rating
The Ballad of Tabitha	The Woodbury Muffin Outbreak	4:00	
Monkey Dreams	The Woodbury Muffin Outbreak	2:57	
A Cold Dark Night (Live)	The Woodbury Muffin Outbreak	3:13	
Leo, The Ballad of	The Woodbury Muffin Outbreak	1:48	
Baby I Forgot To Tell You	The Woodbury Muffin Outbreak	4:04	

Feature	Less	More
AltNGrunge		
CollegeRock		
Country		
DanceRock		
Electronica		
MetalNPunk		
NewWave		
Rap		
RnBSoul		
SingerSongwriter		
SoftRock		
TradRock		
Female		
HiFi		



Outline

- 1 Semantic Audio Analysis
- 2 Organizing Sound Mixtures
- 3 Applications for Audio Semantics
- 4 Open Questions**



4

Open Questions

- **Semantics**
 - What are the abstract **perceptual attributes** of a sound?
How can we describe them?
- **Mixtures**
 - How do people **organize** sound mixtures into separate source percepts?
 - How can we represent generic sound **knowledge**?
- **Applications**
 - **Search/retrieval**: What terminology is most natural for users querying sound databases?
 - What is the best **balance** between machine and human listening?
 - What are the **problems** for which machine listening can be most useful?



Extra slides

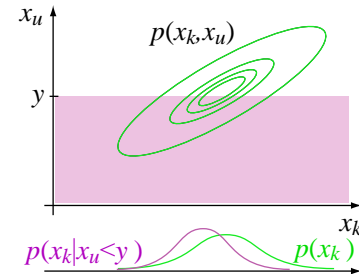


Missing Data Recognition

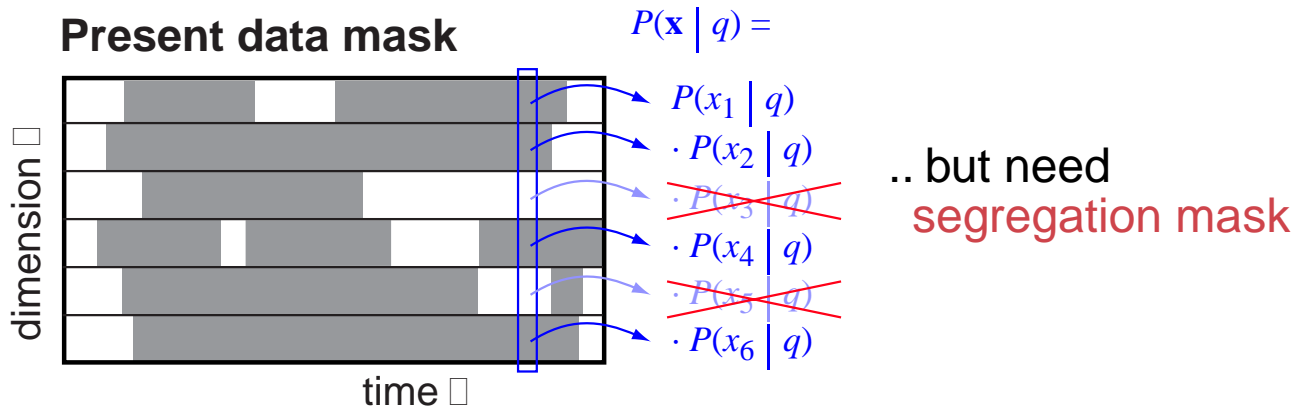
(Barker, Cooke & Ellis '03)

- Can evaluate speech models $p(\mathbf{x}|m)$ over a subset of dimensions x_k

$$p(\mathbf{x}_k | m) = \int p(\mathbf{x}_k, \mathbf{x}_u | m) d\mathbf{x}_u$$



- Hence, **missing data recognition**:



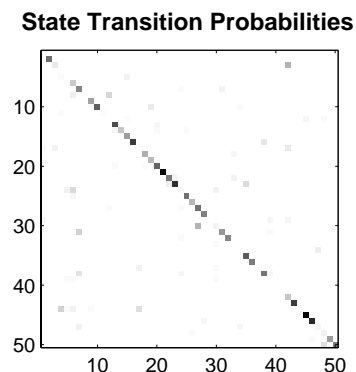
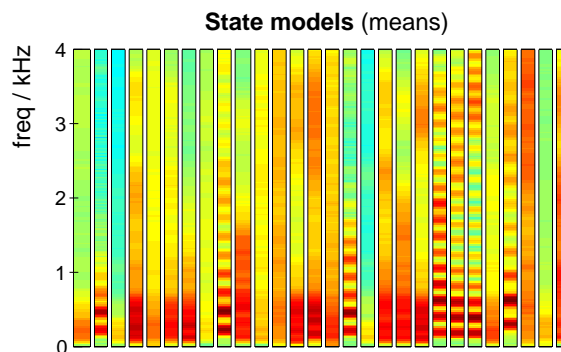
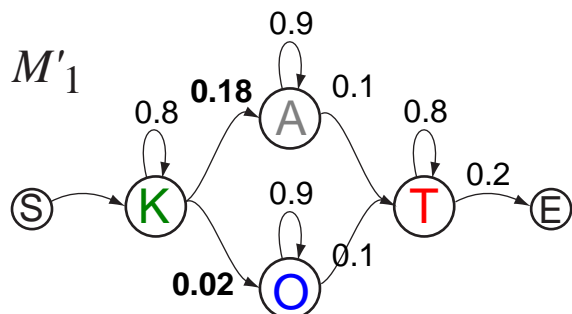
- Fit **model** and **segregation** given obs'n:

$$P(M, S | Y) = P(M) \int P(X | M) \cdot \frac{P(X | Y, S)}{P(X)} dX \cdot P(S | Y)$$

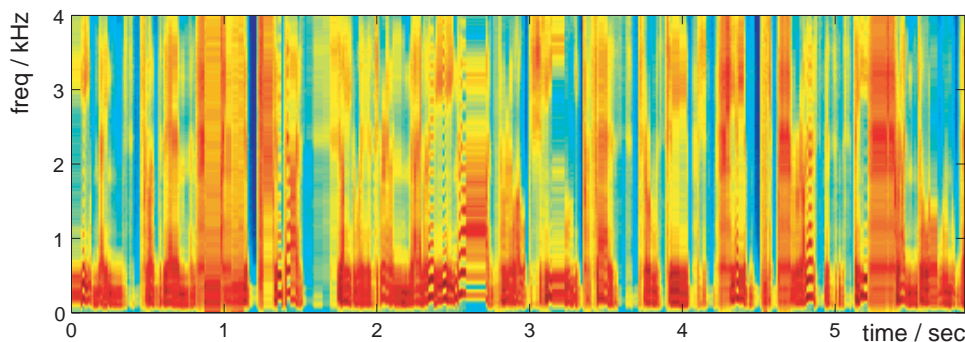


What a speech HMM contains

- **Markov model structure: states + transitions**



- **A generative model**
 - but not a good speech generator!

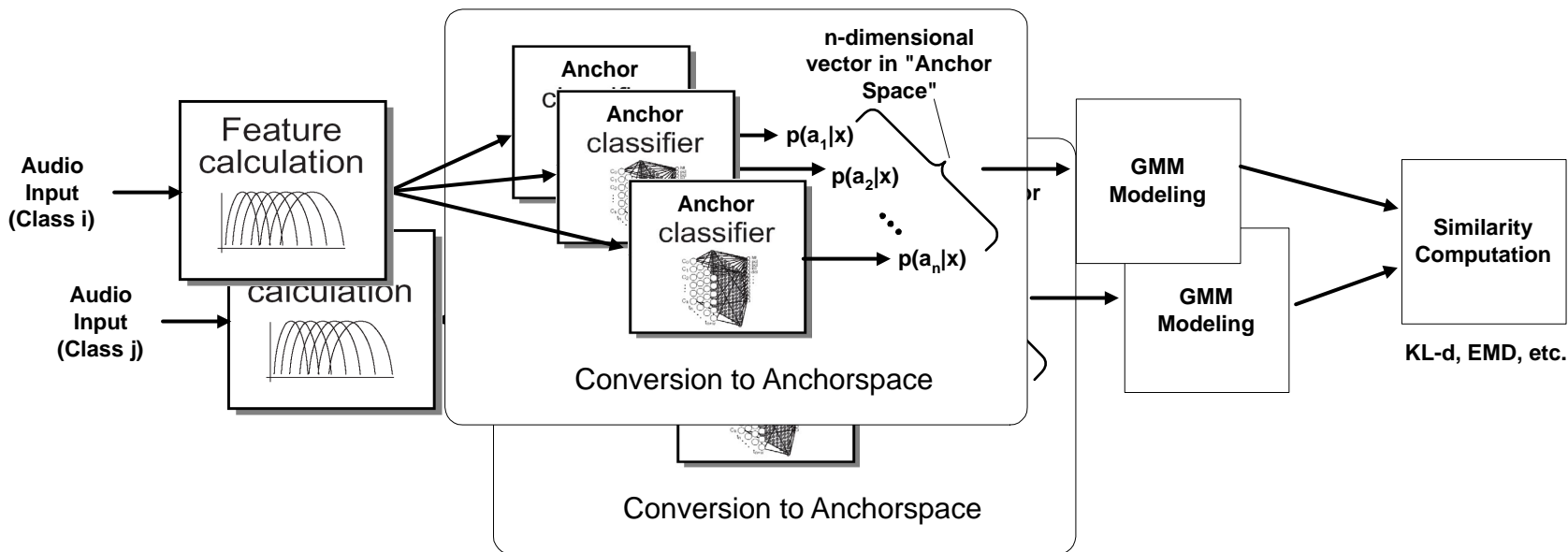


- only meant for **inference** of $p(X|M)$



Music similarity from Anchor space

- A classifier trained for one artist (or genre) will respond **partially** to a similar artist
- Each artist evokes a particular **pattern** of responses over a set of classifiers
- We can treat these **classifier outputs** as a new **feature space** in which to estimate similarity



- **“Anchor space”** reflects subjective qualities?

