Machine Recognition of Sounds in Mixtures

Outline

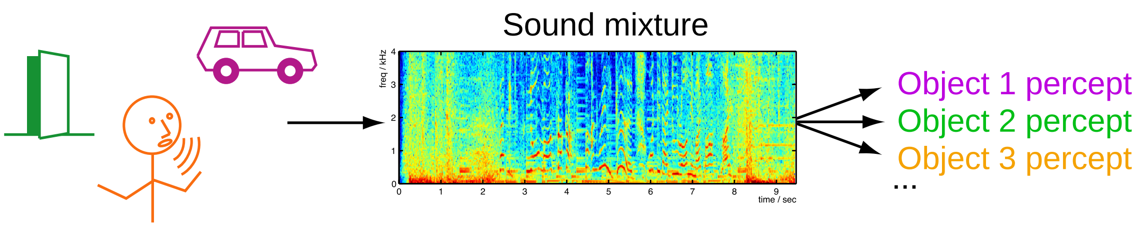
1. **Computational Auditory Scene Analysis**
2. **Speech Recognition as Source Formation**
3. **Sound Fragment Decoding**
4. **Results & Conclusions**

Dan Ellis <dpwe@ee.columbia.edu>
LabROSA, Columbia University, New York

Jon Barker <j.barker@dcs.shef.ac.uk>
SPandH, Sheffield University, UK

❶ Computational Auditory Scene Analysis (CASA)

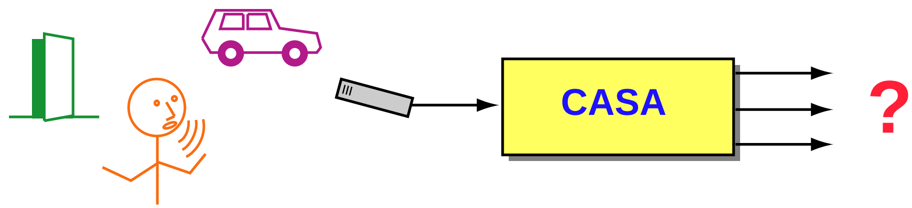
- **Human sound organization:**
  Auditory Scene Analysis



  - composite sound signal → separate percepts
  - based on ecological constraints
  - acoustic cues → perceptual grouping

- **Computational ASA:**
  Doing the same thing by computer
  ...?

What is the goal of CASA?



- **Separate signals?**
  - output is unmixed waveforms
  - underconstrained, very hard ...
  - too hard? not required?
- **Source classification?**
  - output is set of event-names
  - listeners do more than this...
- **Something in-between?**
  Identify independent sources + characteristics
  - standard task, results?

Segregation vs. Inference

- **Source separation requires attribute separation**
  - sources are characterized by attributes (pitch, loudness, timbre + finer details)
  - need to identify & gather different attributes for different sources ...
- **Need representation that segregates attributes**
  - spectral decomposition
  - periodicity decomposition
- **Sometimes values can't be separated**
  - e.g. unvoiced speech
  - maybe infer factors from probabilistic model?

$$p(O, x, y) \rightarrow p(x, y|O)$$

  - or: just skip those values, infer from higher-level context

Outline

1. Computational Auditory Scene Analysis

2. **Speech Recognition as Source Formation**
   - Standard speech recognition
   - Handling mixtures

3. Sound Fragment Decoding

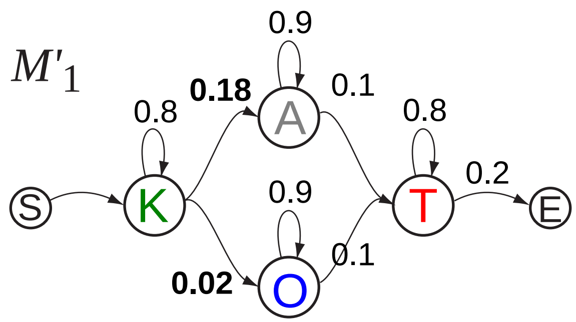
4. Results & Conclusions

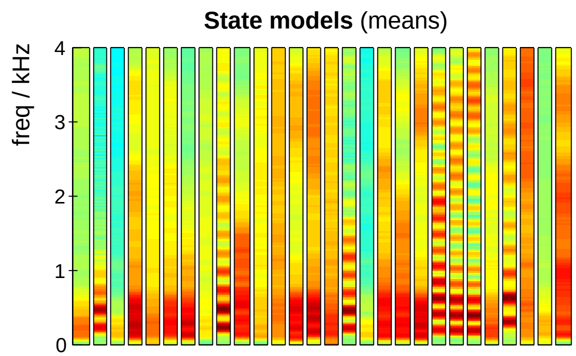
2 Speech Recognition as Source Formation

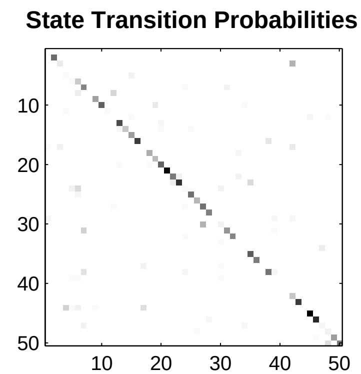
- **Automatic Speech Recognition (ASR): the most advanced sound analysis**
- **ASR extracts abstract information from sound**
  - (i.e. words)
  - even in mixtures (noisy backgrounds) .. a bit
- **ASR is not signal extraction: only certain signal information is recovered**
  - .. just the bits we care about
- **Not CASA preprocessing for ASR: Instead, approach ASR as an example of CASA**
  - words = description of source properties
  - uses strong prior constraints: signal models
  - but: must handle mixtures!

How ASR Represents Speech

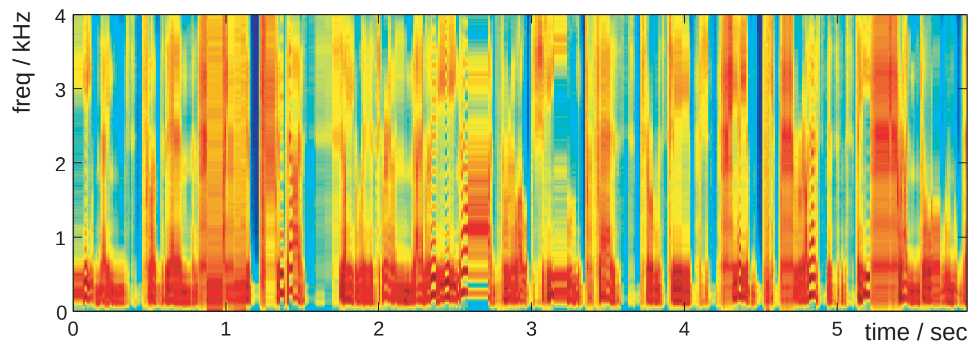
- **Markov model structure: states + transitions**







- **Generative model**
  - but not a good speech generator!



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

Sequence Recognition

- **Statistical Pattern Recognition:**

$$M^* = \underset{M}{\operatorname{argmax}}\, P(M|X) = \underset{M}{\operatorname{argmax}} \frac{P(X|M)\cdot P(M)}{P(X)}$$

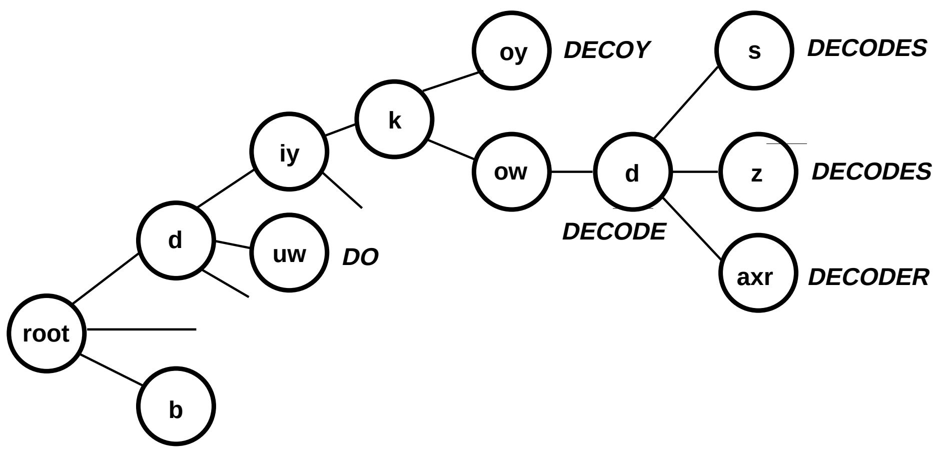
models

observations

- **Markov assumption decomposes into frames:**

$$P(X|M) = \prod_n p(x_n|m_n)\,p(m_n|m_{n-1})$$

- **Solve by searching over all possible state sequences $\{m_n\}$.. but with efficient pruning:**



Approaches to sound mixture recognition

- **Separate signals, then recognize**
  - e.g. (traditional) CASA, ICA
  - nice, if you can do it
- **Recognize combined signal**
  - ‘multicondition training’
  - combinatorics..
- **Recognize with parallel models**
  - full joint-state space?
  - divide signal into fragments, then use missing-data recognition

Outline

1. Computational Auditory Scene Analysis
2. Speech Recognition as Source Formation
3. **Sound Fragment Decoding**
   - Missing Data Recognition
   - Considering alternate segmentations
4. Results & Conclusions

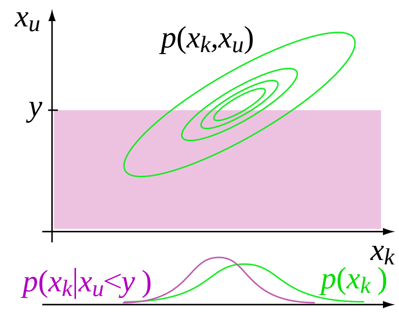
3 Sound Fragment Decoding

- **Signal separation is too hard!**
  Instead:
  - segregate features into partially-observed sources
  - then classify
- **Made possible by missing data recognition**
  - integrate over uncertainty in observations for true posterior distribution
- **Goal:**
  Relate clean speech models $P(X|M)$
  to speech-plus-noise mixture observations
  - .. and make it tractable

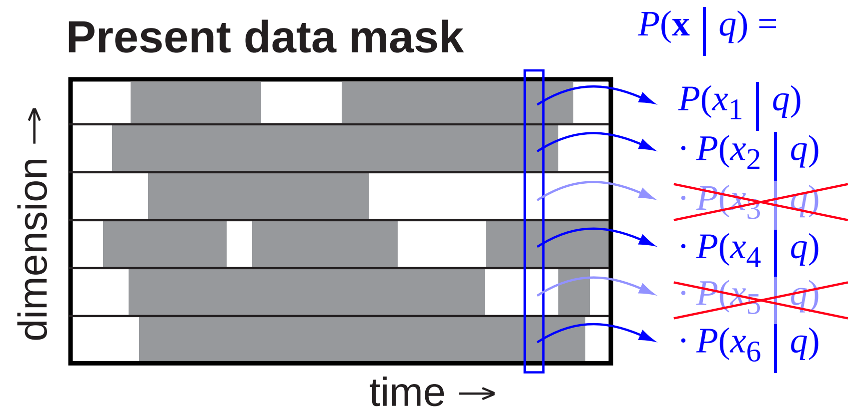
Missing Data Recognition

- **Speech models $p(\mathbf{x}|m)$ are multidimensional...**
  - i.e. means, variances for every freq. channel
  - need values for all dimensions to get $p(\bullet)$
- **But: can evaluate over a subset of dimensions x_k**

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



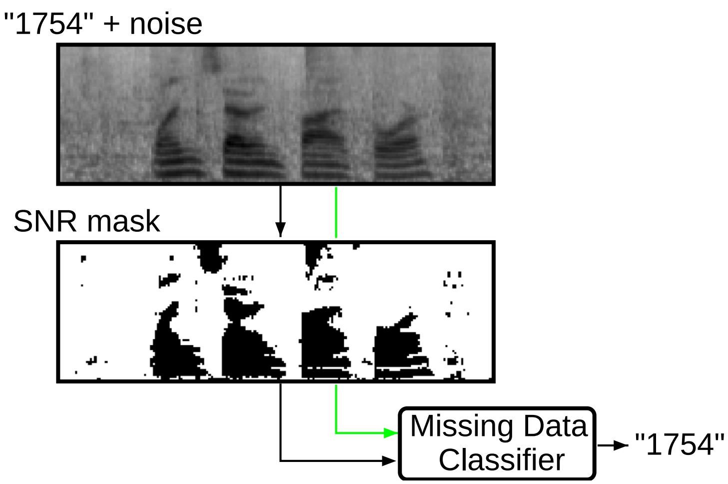
- **Hence, missing data recognition:**

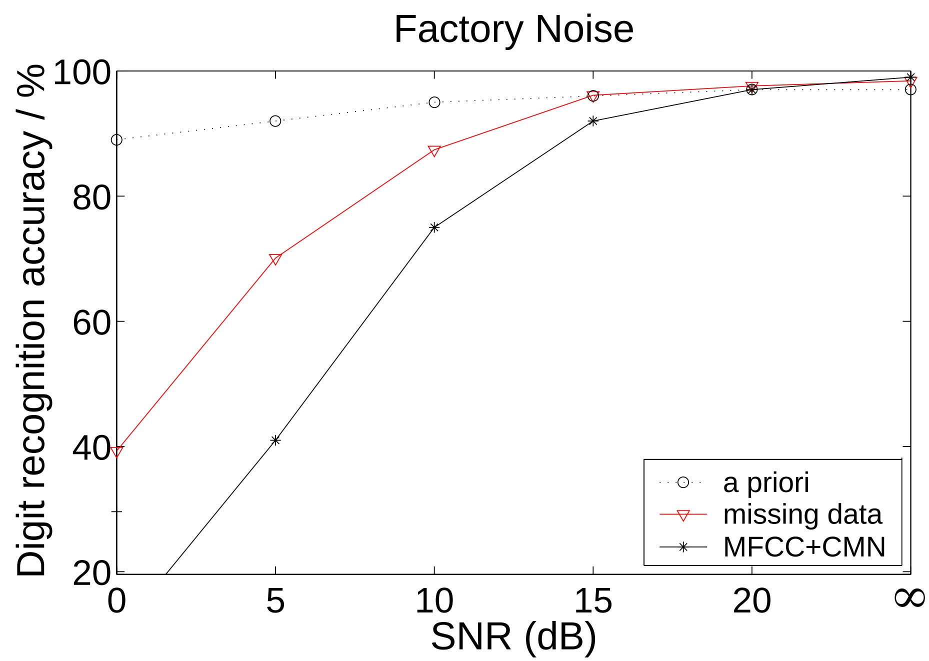


  - hard part is finding the mask (segregation)

Missing Data Results

- **Estimate static background noise level $N(f)$**
- **Cells with energy close to background are considered "missing"**





  - must use spectral features!

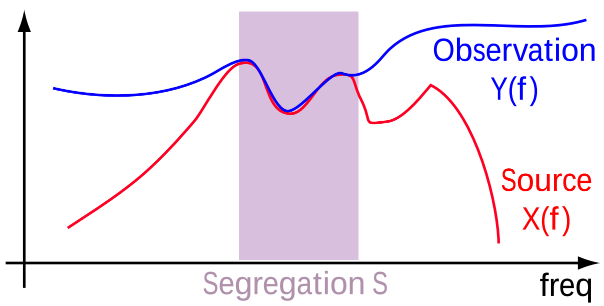
- **But: nonstationary noise → spurious mask bits**
  - can we try removing parts of mask?

Comparing different segregations

- **Standard classification chooses between models M to match source features X**

$$M^* = \underset{M}{\operatorname{argmax}}\ P(M|X) = \underset{M}{\operatorname{argmax}}\ P(X|M) \cdot \frac{P(M)}{\cancel{P(X)}}$$

- **Mixtures: observed features Y, segregation S, all related by $P(X|Y,S)$**



- **Joint classification of model and segregation:**

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

  - $P(X)$ no longer constant

Calculating fragment matches

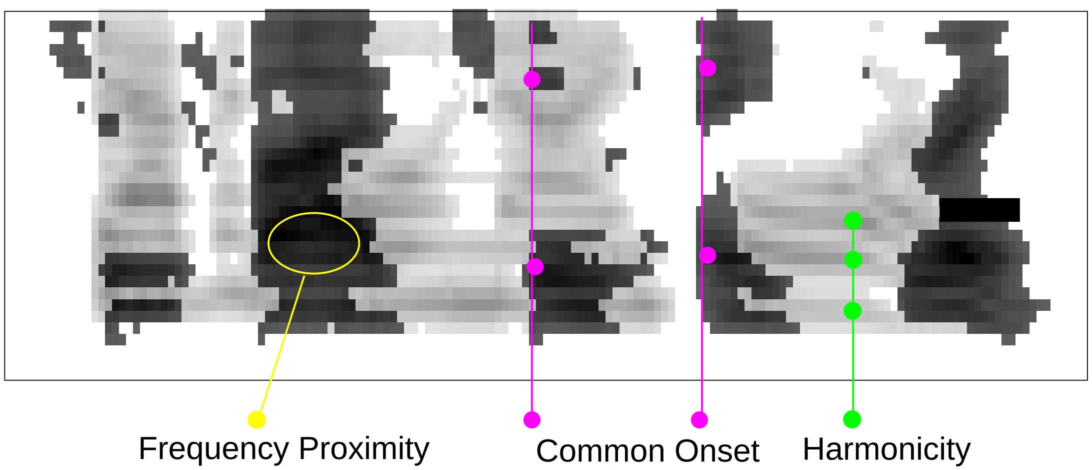
$$P(M,S|Y) = P(M)\int P(X|M) \cdot \frac{P(X|Y,S)}{P(X)} dX \cdot P(S|Y)$$

- $P(X|M)$ **- the clean-signal feature model**
- $P(X|Y,S)/P(X)$ **- is X 'visible' given segregation?**
- **Integration collapses some bands...**
- $P(S|Y)$ **- segregation inferred from observation**
  - just assume uniform, find S for most likely M
  - or: use extra information in Y to distinguish S's...
- **Result:**
  - probabilistically-correct relation between clean-source models $P(X|M)$ and inferred, recognized source + segregation $P(M,S|Y)$

Using CASA features

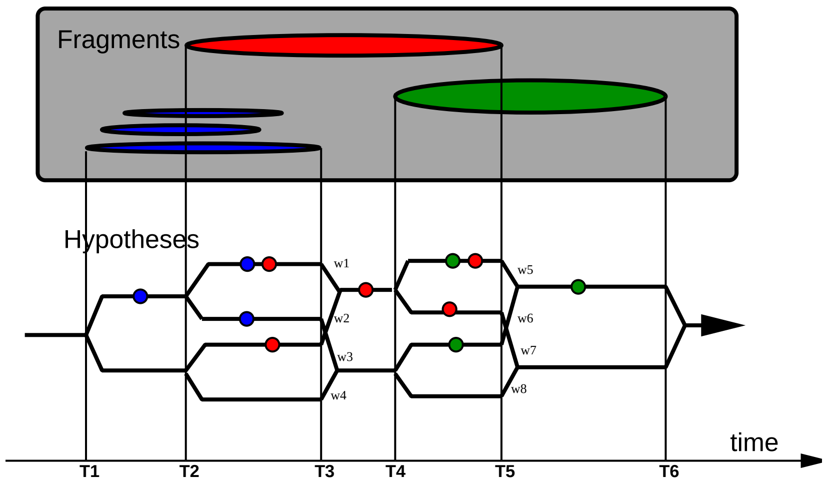
- *P*(*S*|*Y*) **links acoustic information to segregation**
  - is this segregation worth considering?
  - how likely is it?

- **Opportunity for CASA-style information to contribute**
  - periodicity/harmonicity:
    these different frequency bands belong together
  - onset/continuity:
    this time-frequency region must be whole



Fragment decoding

- **Limiting S to whole fragments makes hypothesis search tractable:**

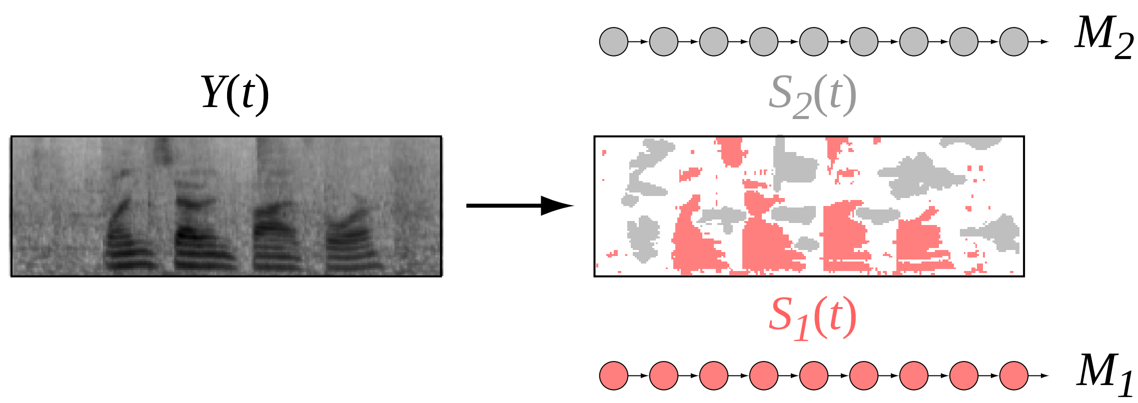


  - choice of fragments reflects $P(S|Y) \cdot P(X|M)$
    i.e. best combination of segregation
    and match to speech models

- **Merging hypotheses limits space demands**
  - .. but erases specific history

Multi-Source Decoding

- **Match multiple models at once?**



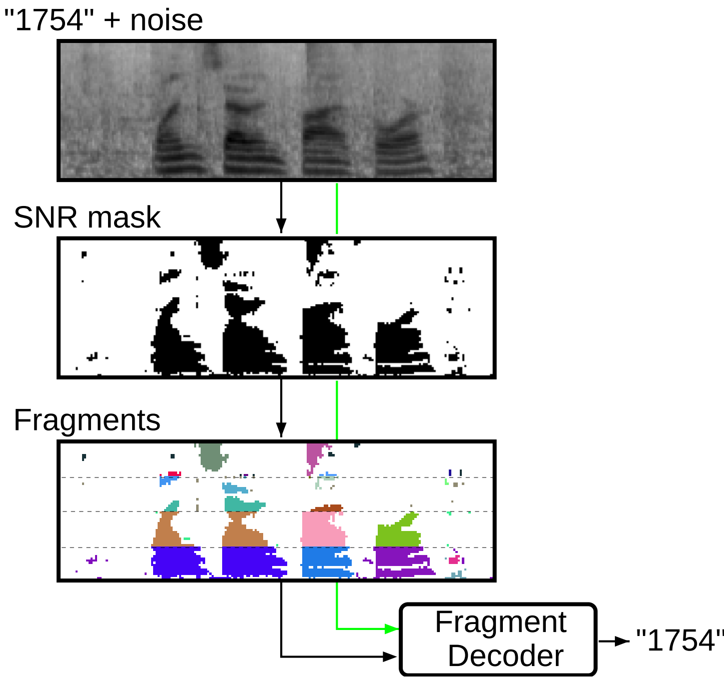
- disjoint subsets of cells for each source
- each model match $P(M_x|S_x,Y)$ is independent
- masks are mutually dependent: $P(S_1,S_2|Y)$

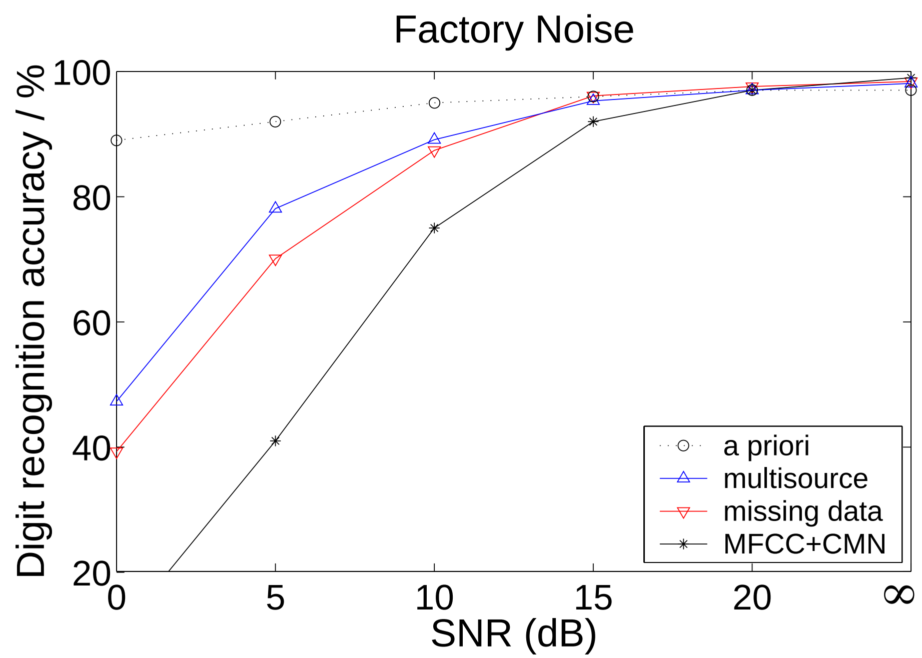
Outline

1. Computational Auditory Scene Analysis
2. Speech Recognition as Source Formation
3. Sound Fragment Decoding
4. **Results & Conclusions**
   - Speech recognition
   - Alarm detection

❹ Speech fragment decoder results

- **Simple $P(S|Y)$ model forces contiguous regions to stay together**
  - big efficiency gain when searching S space

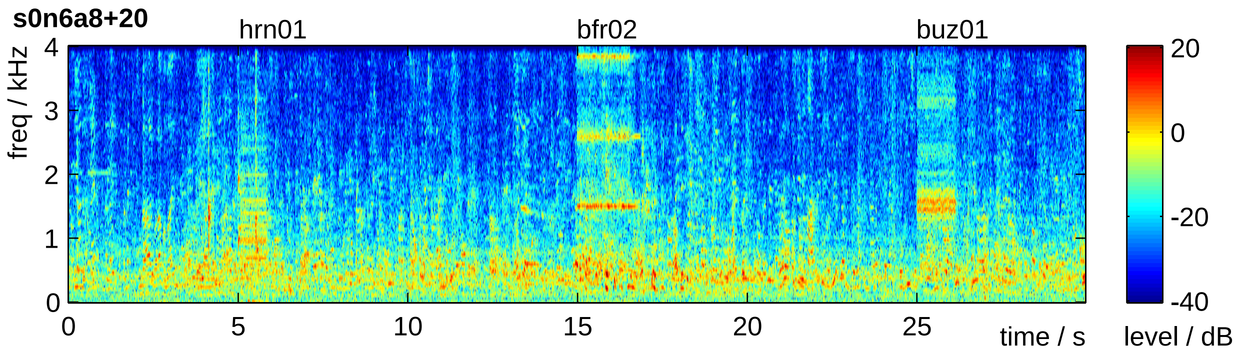




- **Clean-models-based recognition rivals trained-in-noise recognition**

Alarm sound detection

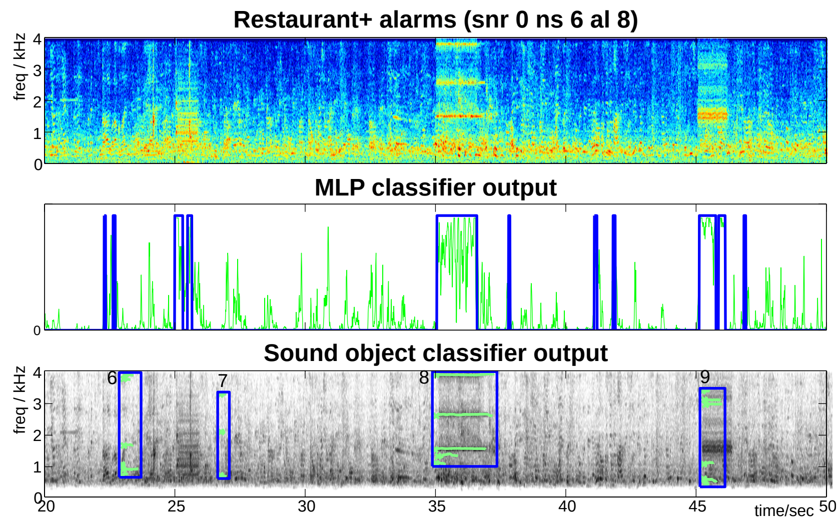
- **Alarm sounds have particular structure**
  - people 'know them when they hear them'
  - clear even at low SNRs



- **Why investigate alarm sounds?**
  - they're supposed to be easy
  - potential applications...

- **Contrast two systems:**
  - standard, global features, $P(X|M)$
  - sinusoidal model, fragments, $P(M,S|Y)$

Alarms: Results



- **Both systems commit many insertions at 0dB SNR, but in different circumstances:**

Noise	*Neural net system*			*Sinusoid model system*		
	Del	*Ins*	*Tot*	*Del*	*Ins*	*Tot*
1 (amb)	7 / 25	2	36%	14 / 25	1	60%
2 (bab)	5 / 25	63	272%	15 / 25	2	68%
3 (spe)	2 / 25	68	280%	12 / 25	9	84%
4 (mus)	8 / 25	37	180%	9 / 25	135	576%
Overall	**22 / 100**	170	**192%**	**50 / 100**	147	**197%**

Summary & Conclusions

- **Scene Analysis**
  - necessary for useful hearing
- **Recognition**
  - a model domain for scene analysis
- **Fragment decoding**
  - recognition with partial observations
  - combines segmentation & model fitting
- **Future work**
  - models of sources other than speech
  - simultaneous ‘perception’ of multiple sources