The University Of Sheffield.

Probabilistic frameworks for scene analysis

Jon Barker

Speech and Hearing Research Group
Department of Computer Science
University of Sheffield, UK

Probabilistic frameworks for scene analysis

- A **scene** is an acoustic mixture made up of multiple sound sources.
- **Analysis** is the process of recovering high-level descriptions of the sound sources from the mixture.
  - A large dog is barking in the distance.
  - A car is approaching rapidly from the right.
  - A person nearby is asking me my name.
- This talk will focus on a specific scene analysis task: *transcribing speech signals in complex acoustic scenes…*
- … but I will try to draw back to the more general problem at the end.

Contrasting research goals

1. **Engineering effective scene analysis systems**
   - e.g. building a speech recogniser which makes smallest number of recognition errors.
   - May ignore human constraints, e.g. microphone arrays.
   - Perhaps ASA-inspired but not necessarily CASA.
   - Motivation: developing effective machine-listening applications.

2. **Modelling human performance**
   - e.g. building a system which mimics human speech recognition ability. – microscopic models of speech intelligibility.
   - Adopts human constraints and tries to maintain pyschological plausibility. Might describe itself as CASA.
   - Evaluated by comparison to human performance.
   - Motivation: understanding human hearing, models can be used to design acoustic environments, human-robot communication, better signal processing for hearing prostheses.

Some questions

- How can we interpret scenes containing an unknown and time-varying number of sound sources?
- How can we ensure computational cost stays bounded regardless of the complexity of the scene?
- How can we interpret scenes that may contain unfamiliar sound sources?
- How can we learn models of individual sound sources when presented with acoustic mixtures?

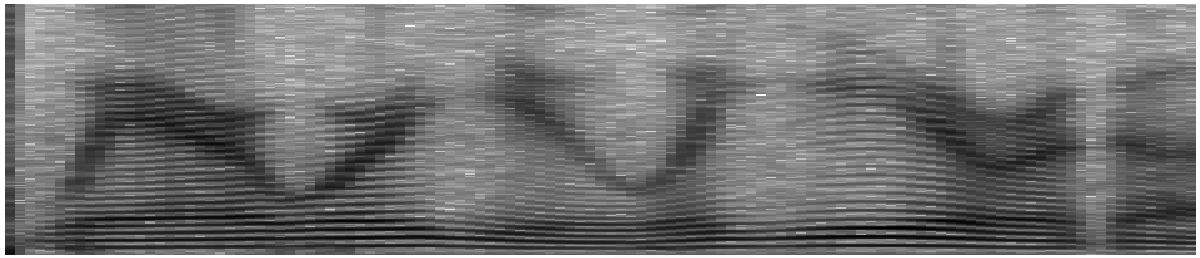
Overview

- Some motivating observations.
- Probabilistic models of scene analysis.
- Application to robust speech recognition.
- Concluding questions.

Listen …

Speech vs SWS

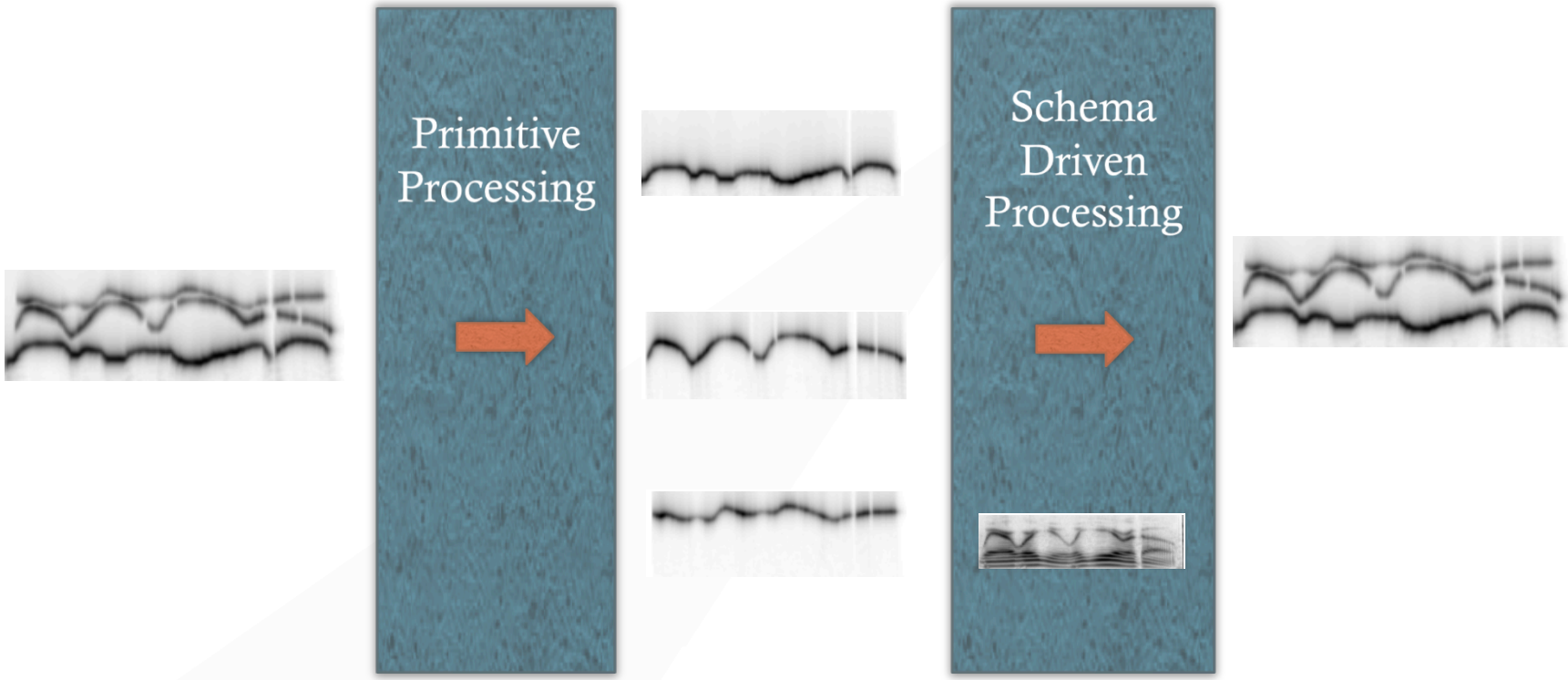
8 kHz



"Where were you a year ago ?"

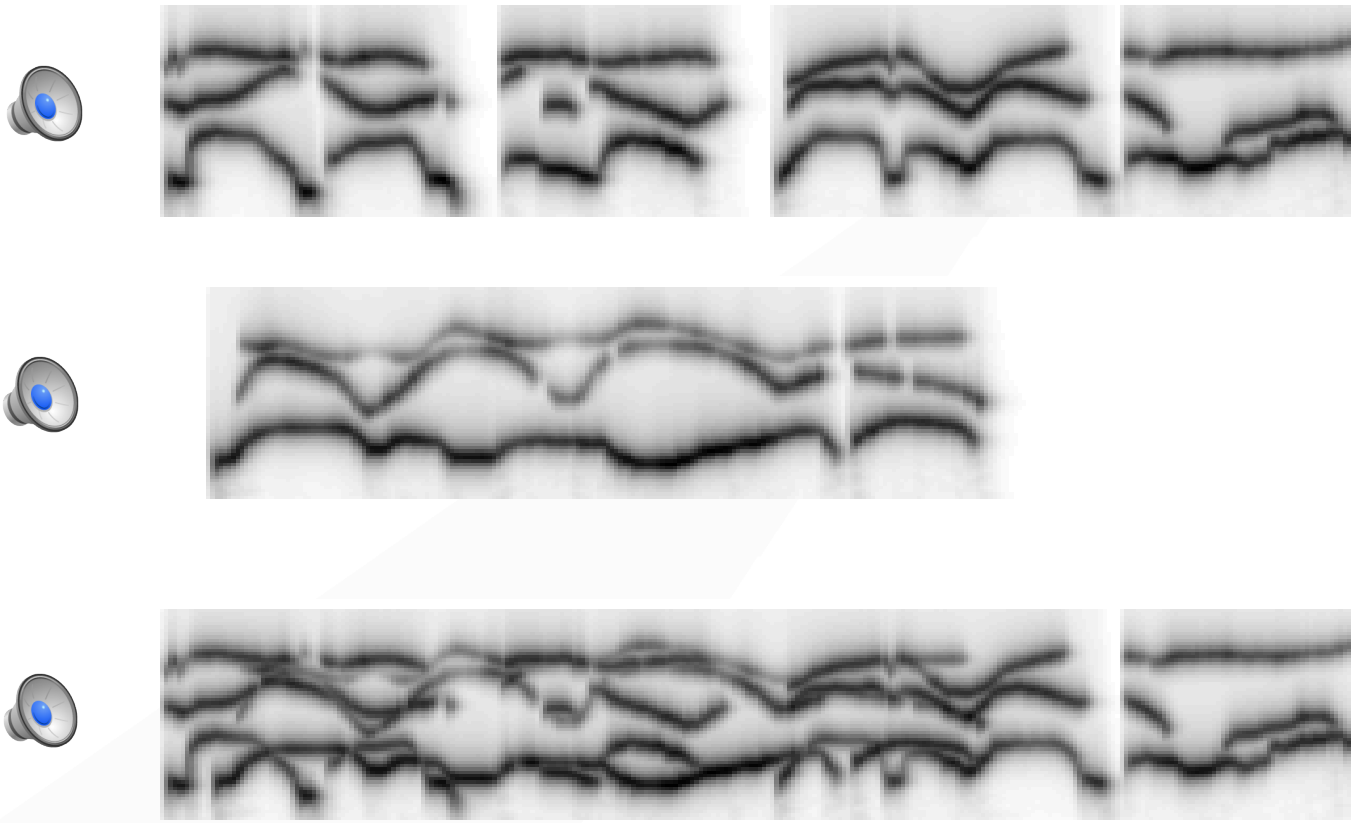
Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. (1981). Science, 212, 947–950.

An Auditory Scene Analysis view of SWS perception

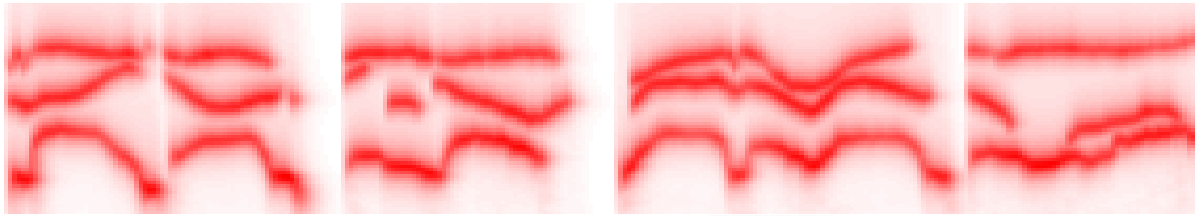


Listen

A SWS scene

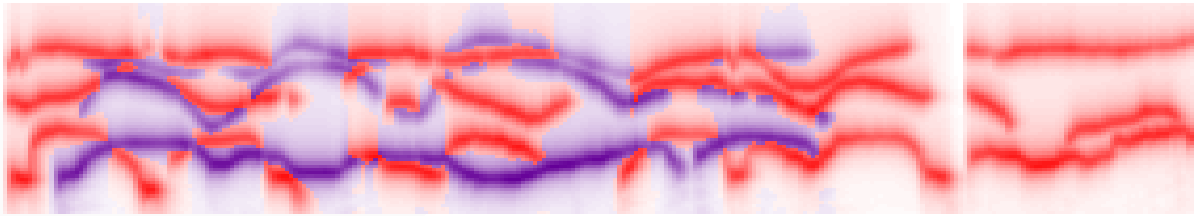


A SWS scene



My dog bingo ran around the wall

Where were you a year ago ?

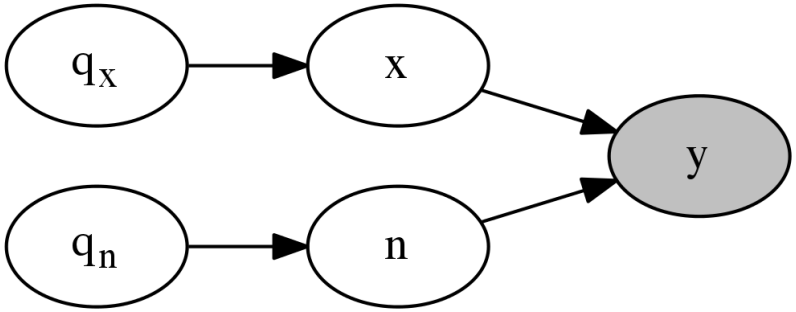


Barker, J., & Cooke, M. (1999). Speech Communication, 27, 159–174.

Visual analogy

Model Combination

q - model states
x - target source spectrum
n - noise source spectrum
y - observed noisy spectrum



$$q_x, q_n = \underset{q_x, q_n}{\operatorname{argmax}} \; p(q_x, q_n | y)$$

$$q_x = \underset{q_x}{\operatorname{argmax}} \sum_{q_n} p(q_x, q_n | y)$$

Model Combination

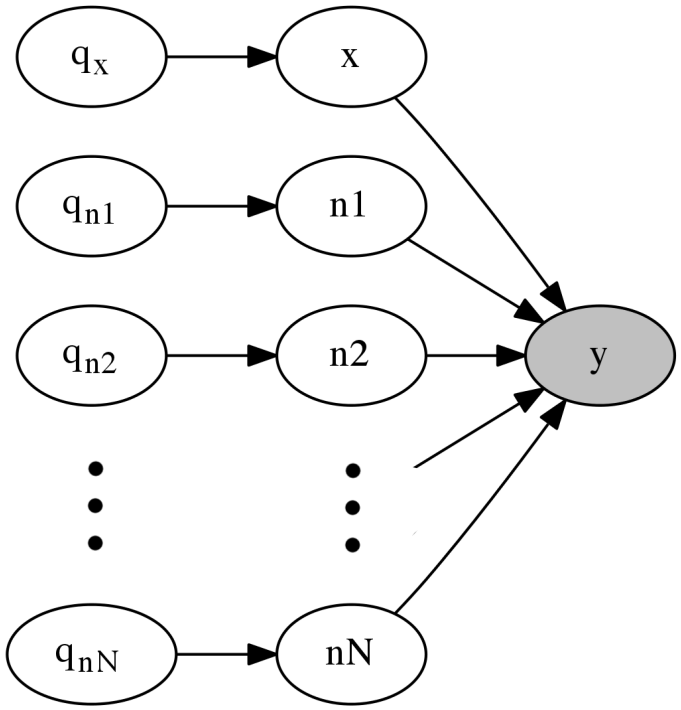
Model combination can be easily generalised to consider multiple sound sources.

But,

- Combinatorial explosion of the state space as more sources are added.

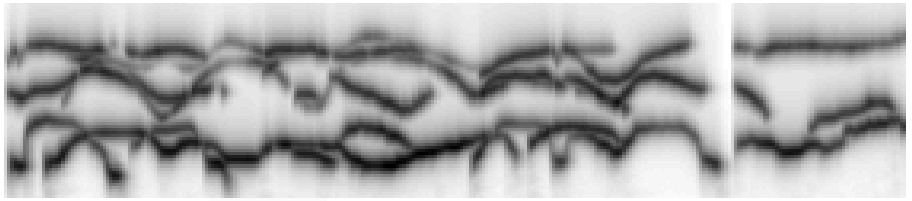
$$q_x = \underset{q_x}{\text{argmax}} \sum_{q_{n1}} \sum_{q_{n2}} \dots \sum_{q_{nN}} p(q_x, q_{n1}, q_{n2}, \dots, q_{nN} | y)$$

- And how should model complexity be determined?

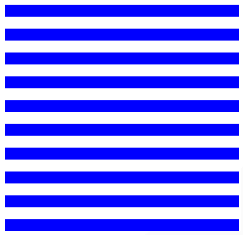


Foreground/Background Segmentation

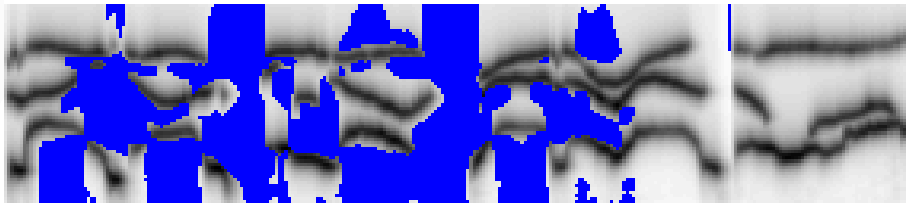
Mixed signal



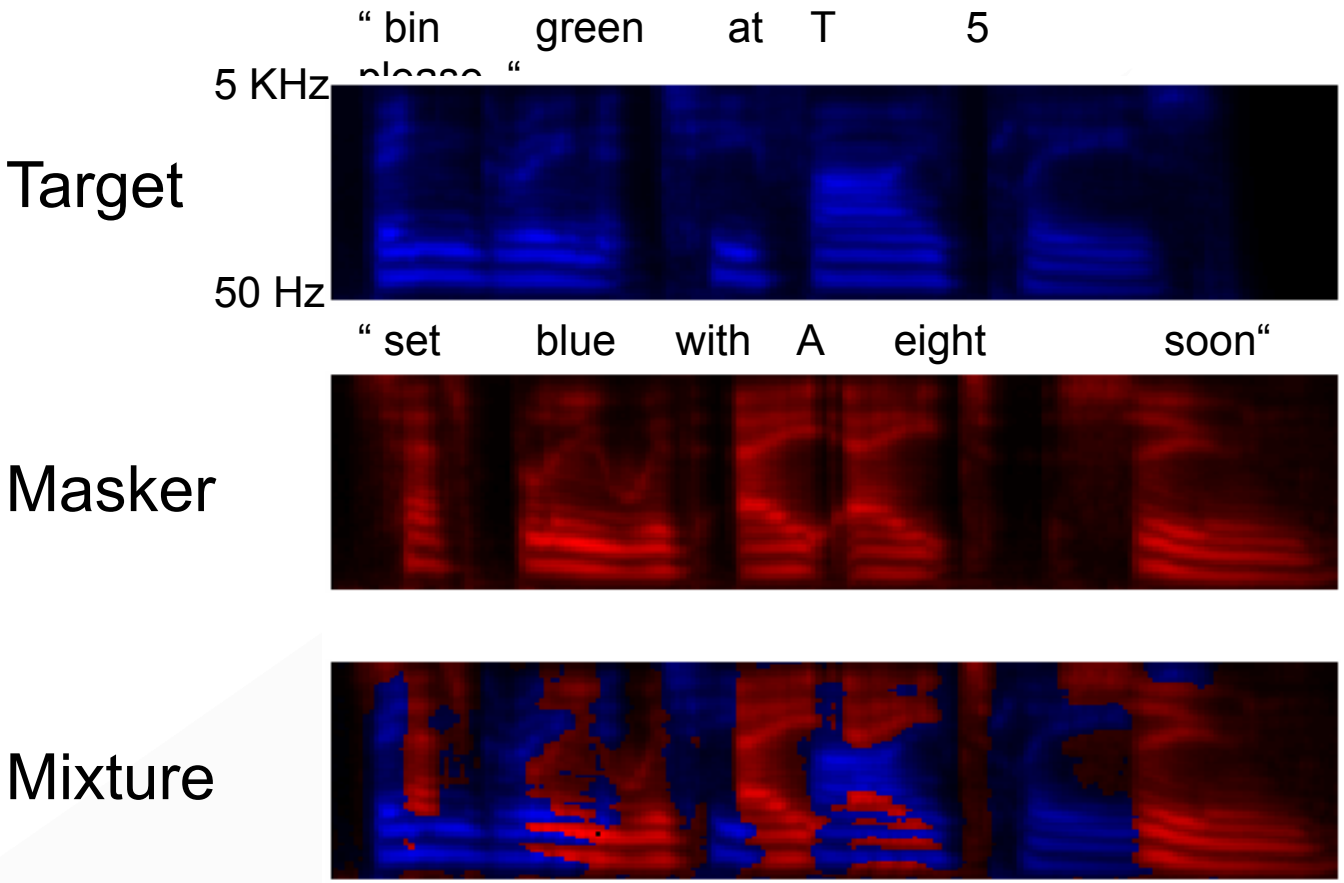
Segmentation mask



Masked mixture

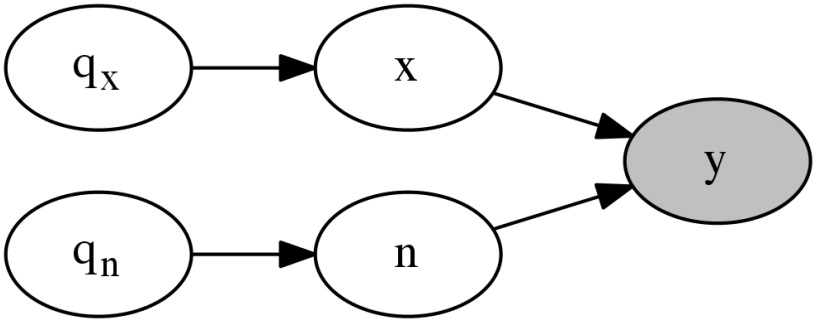


Speech mixture at 0 dB

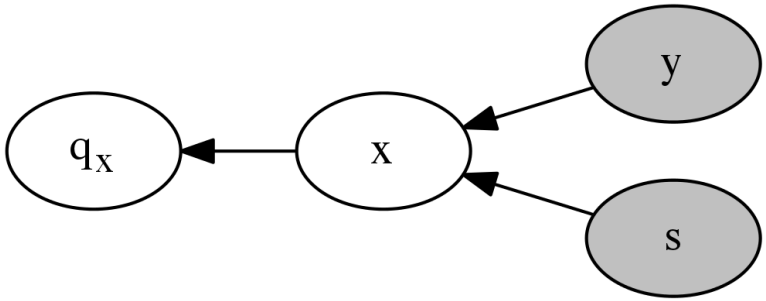


Missing data ASR

q - model states
x - target source spectrum
n - noise source spectrum
y - observed noisy spectrum



q - model states
x - target source spectrum
y - observed noisy spectrum
s - segmentation



- The background is not modeled explicitly.
- Need model for p(x|y, s) for which p(x|y, s) = p(x|y, s, n)

Missing data ASR

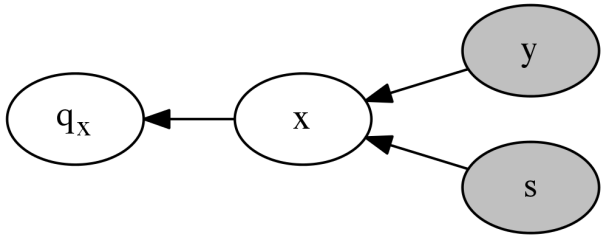
$$q_x = \underset{q_x}{\operatorname{argmax}}\ p(q_x|y,s)$$

$$p(q_x|y,s) = \int_x p(q_x,x|y,s)\mathrm{d}x$$

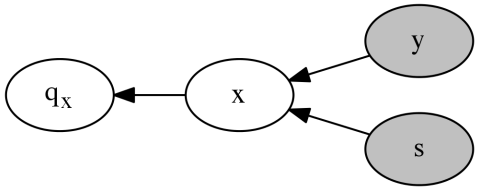
$$= \int_x p(q_x|x,y,s)p(x|y,s)\mathrm{d}x$$

$$= \int_x p(q_x|x)p(x|y,s)\mathrm{d}x$$

$$= \int_x p(x|q_x)\frac{p(x|y,s)}{p(x)}\mathrm{d}x\ \ p(q_x)$$

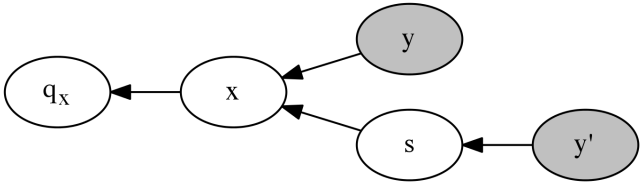


Meta-missing data



$$q_x = \underset{q_x}{\text{argmax}}\ p(q|y,s)$$

$$= \underset{q_x}{\text{argmax}} \int_x p(x|q_x)\frac{p(x|y,s)}{p(x)}\mathrm{d}x\ \ p(q_x)$$

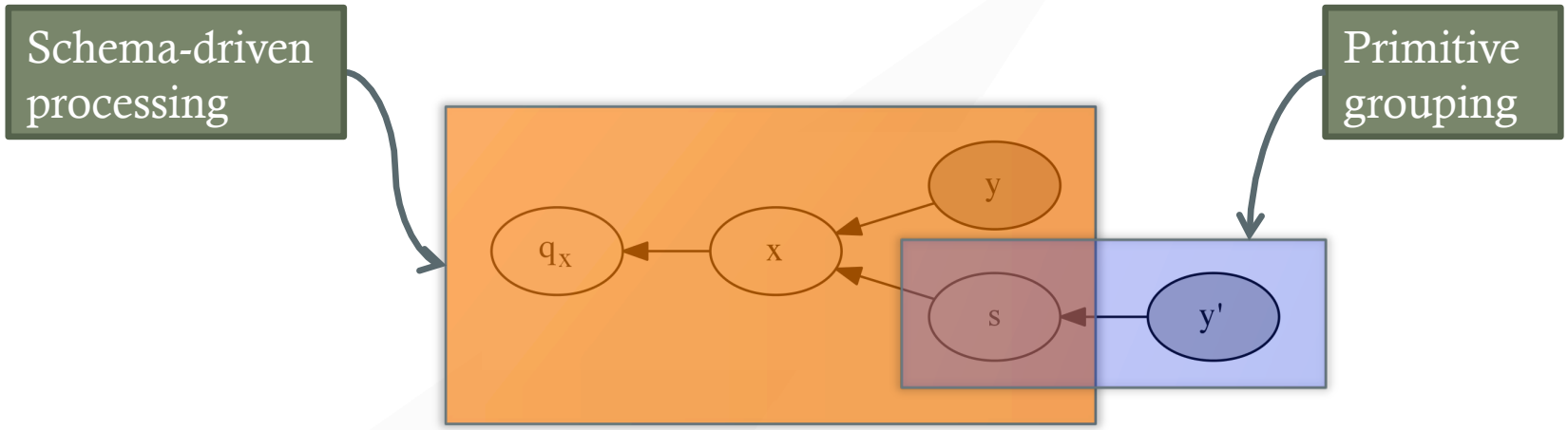


$$q_x = \underset{q_x}{\text{argmax}}\ p(q|y,y')$$

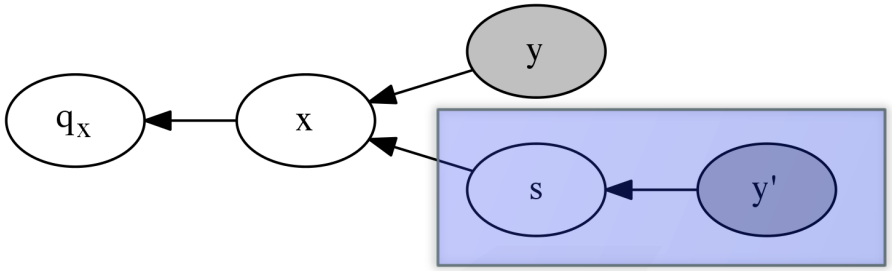
$$= \underset{q_x}{\text{argmax}} \sum_s \left(\int_x p(x|q_x)\frac{p(x|y,s)}{p(x)}\mathrm{d}x\ \ p(s|y') \right) p(q_x)$$

$$q_x, s = \underset{q_x,\, s}{\text{argmax}} \int_x p(x|q_x)\frac{p(x|y,s)}{p(x)}\mathrm{d}x\ \ p(s|y')p(q_x)$$

Relation to ASA



Fragment decoding



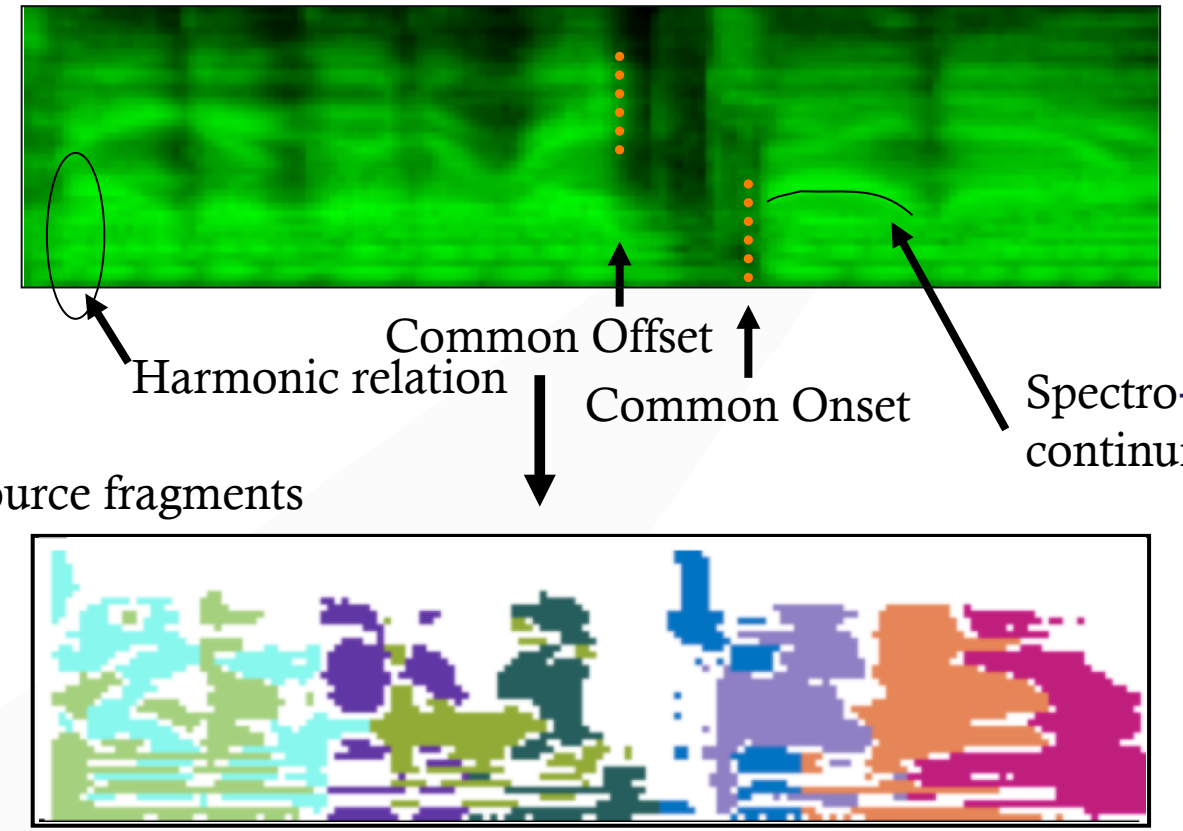
Simultaneously search over all target model states and all possible segmentations

$$q_x, s = \underset{q_x,\, s}{\operatorname{argmax}} \int_x p(x|q_x) \frac{p(x|y,s)}{p(x)} \mathrm{d}x \;\; p(s|y') p(q_x)$$

p(s|y') assigned constant value for all segmentations that are consistent with primitive grouping rules, else p(s|y') = 0

Sound source fragments

Identify fragments – i.e. local source segmentation -- using `primitive cues'.



Fragment generation

Filterbank output

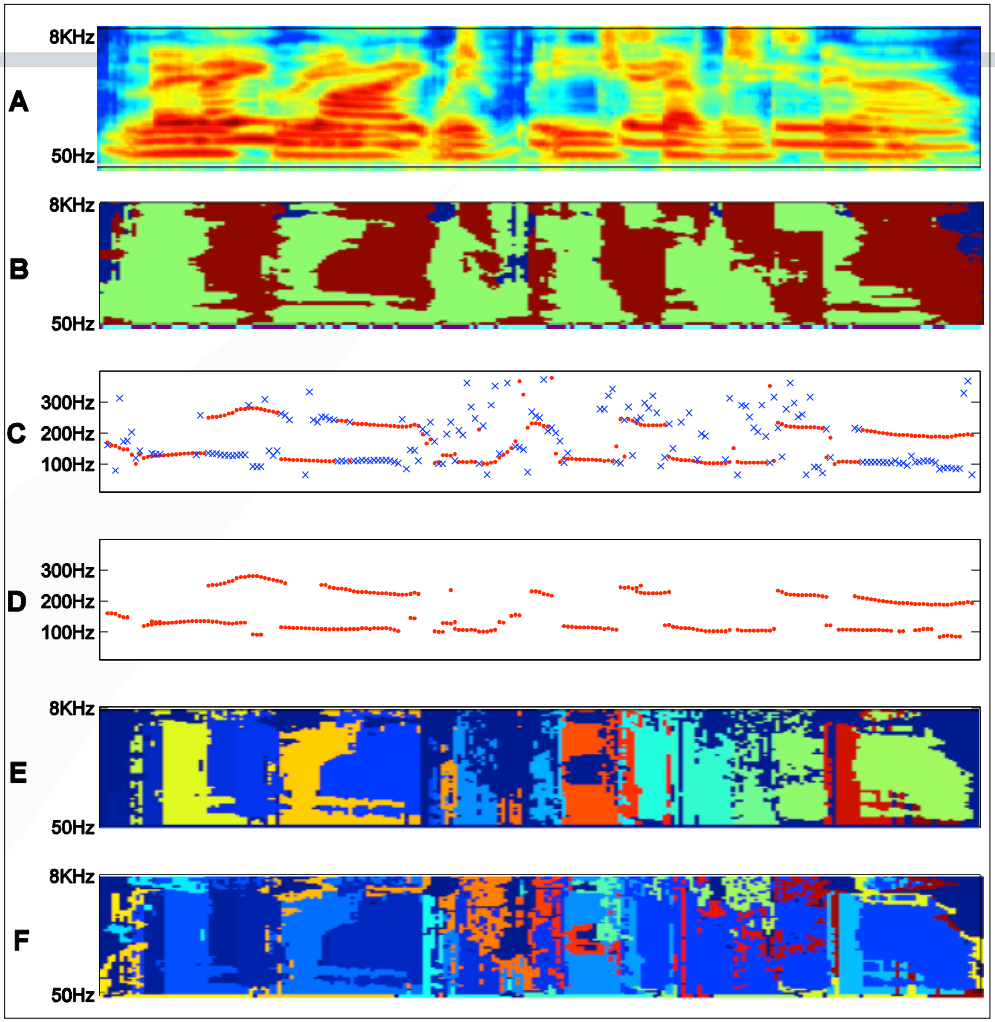
`Ideal' segmentation

Pitch candidates

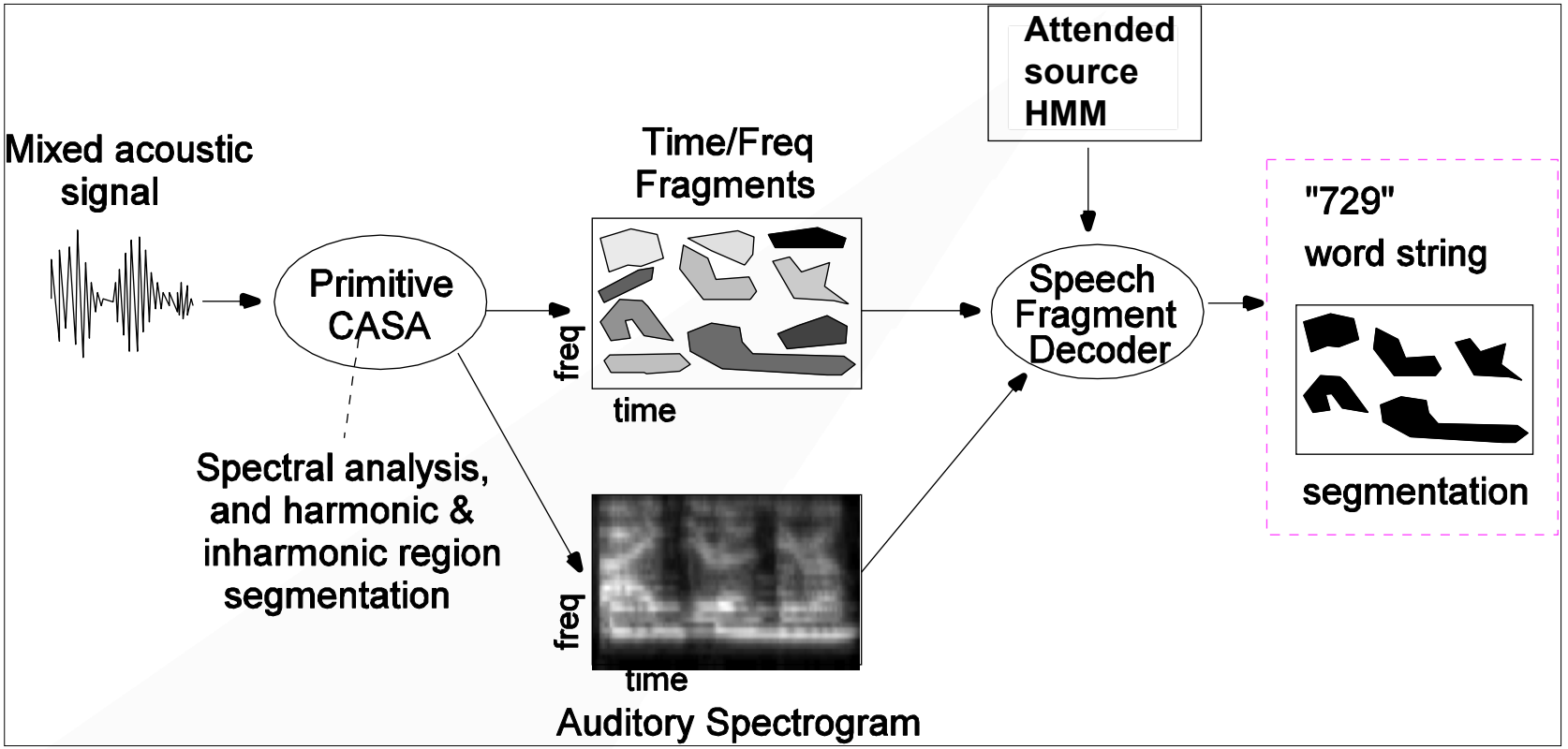
Pitch tracking

Harmonic fragments…

…+ inharmonic fragments



Speech Fragment Decoding



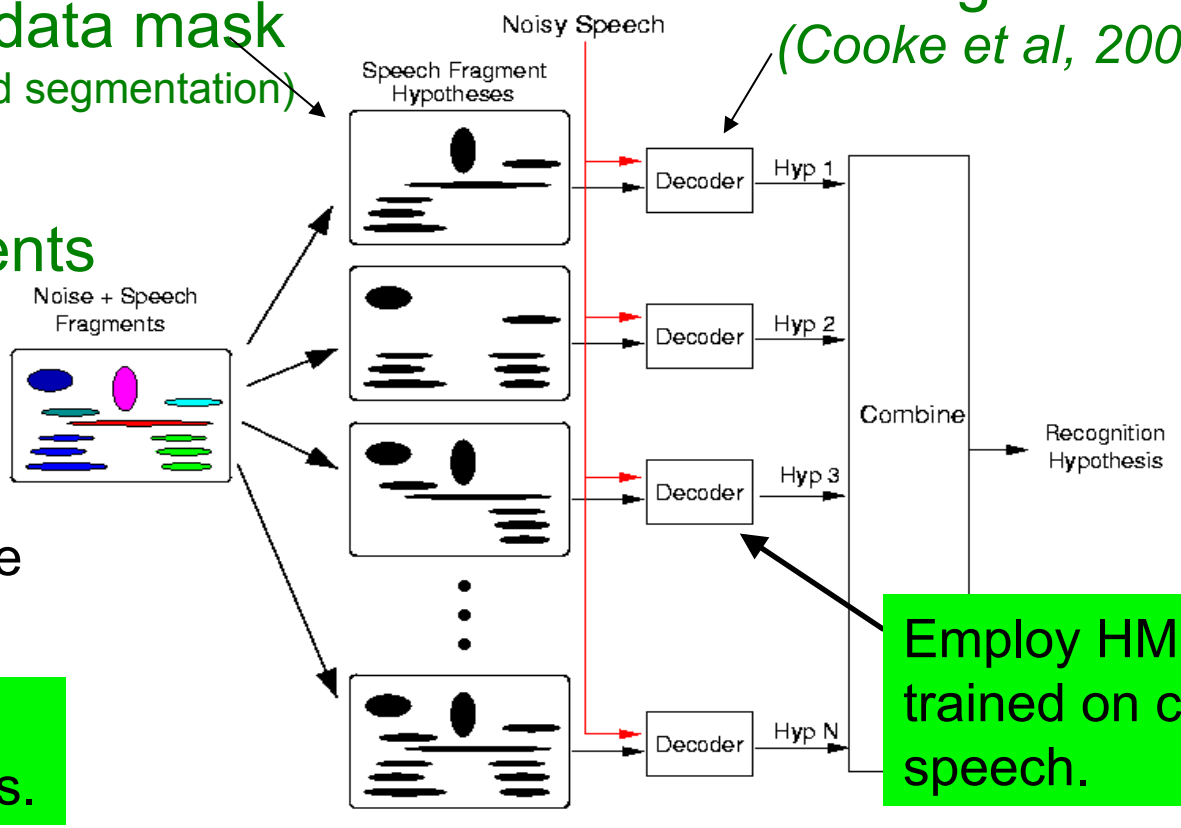
The fragment decoder

A simple (but terribly inefficient) implementation:

Missing data mask
(i.e. foreground/background segmentation)

Source fragments

Missing data ASR
(Cooke et al, 2001)

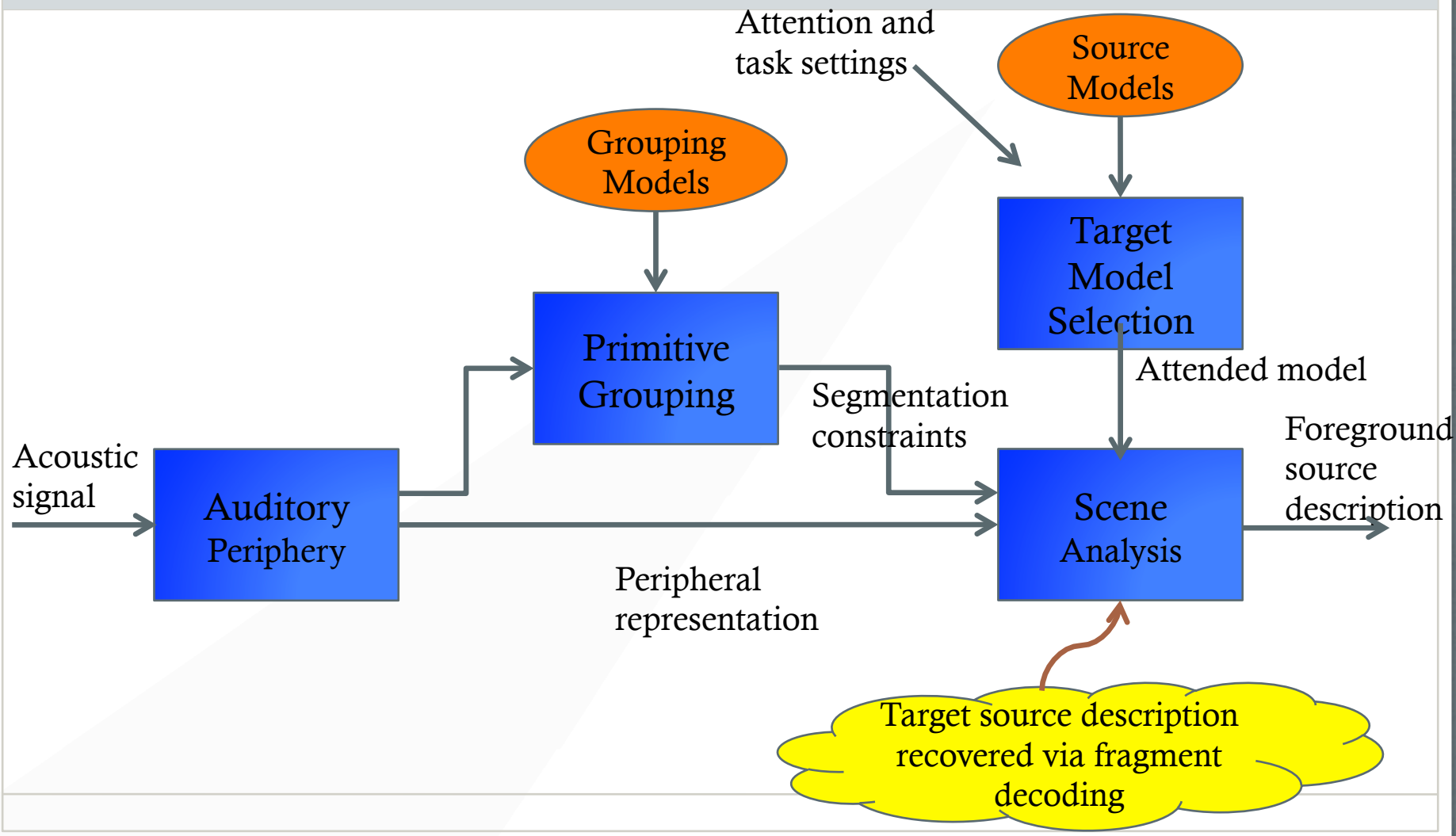


But for N fragments there are 2^N possible mask hypotheses.

Luckily, an efficient implementation exists.

Employ HMMs trained on clean speech.

A model of scene analysis



The Grid corpus

- Small vocabularly, read speech

VERB	COLOUR	PREP.	LETTER	DIGIT	ADVERB
bin	blue	at	a-z	1-9	again
lay	green	by	(no ‘w’)	and zero	now
place	red	on			please
set	white	with			soon

- 34 speakers, 1000 utterances from each speaker

Simultaneous Speech Experiments

- Target and Masker utterances mixed at a range of SNRs -9 dB to 6 dB
- Target utterances always use keyword `white', e.g.

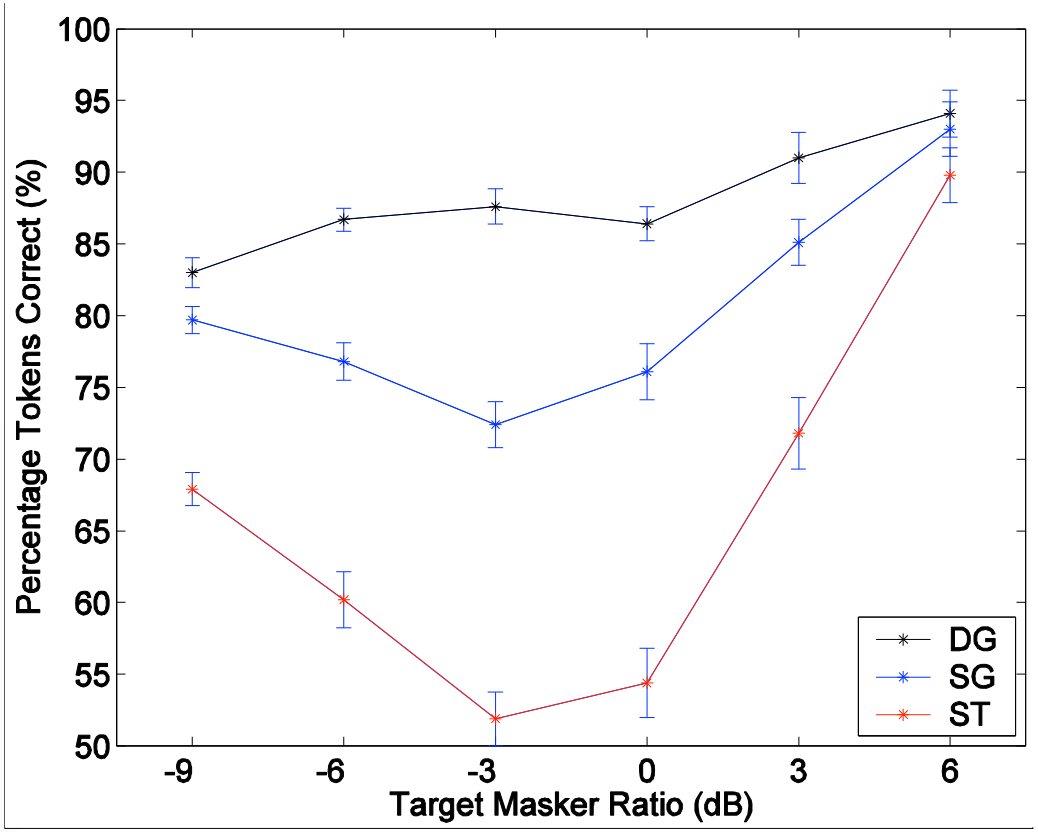
  "bin WHITE at k 2 please"
- Task is to report the Grid reference.

Human performance

Diff. gender

Same gender

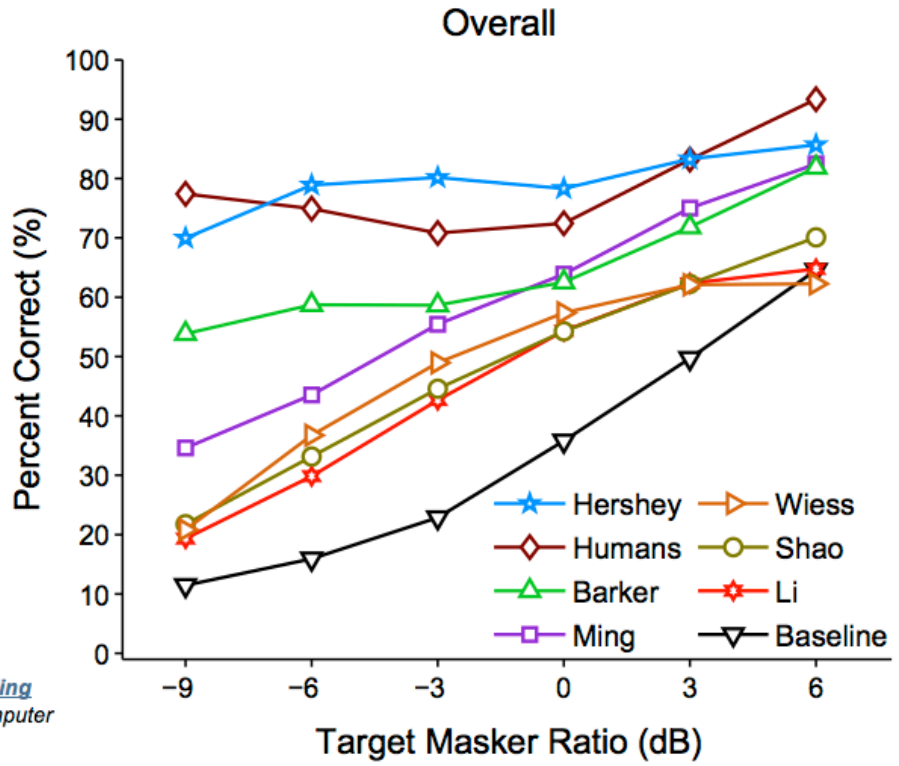
Same talker



ASR performance

PASCAL Speech Separation Challenge, Interspeech 2006

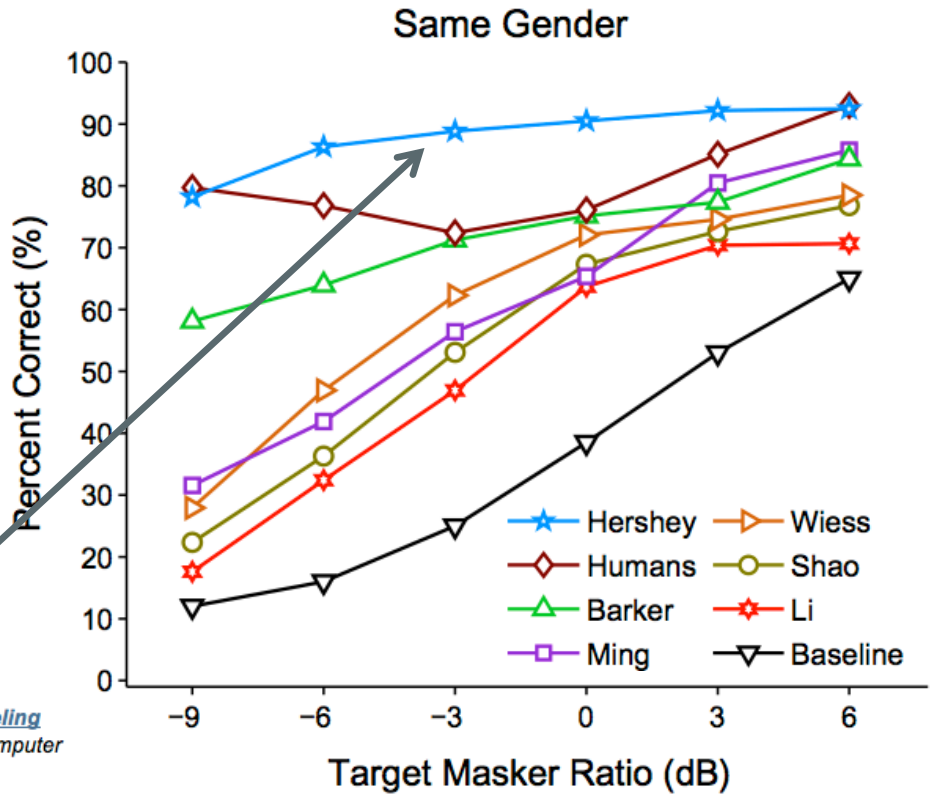
Brown – Human
Blue – Model combination
(Hershey et al, 2010)
Green – Fragment decoding



Super-human multi-talker speech recognition: A graphical modeling approach. J R Hershey, S J Rennie, P A Olsen, T T Kristjansson, *Computer Speech & Language*, Elsevier, 2010.

`super human' performance

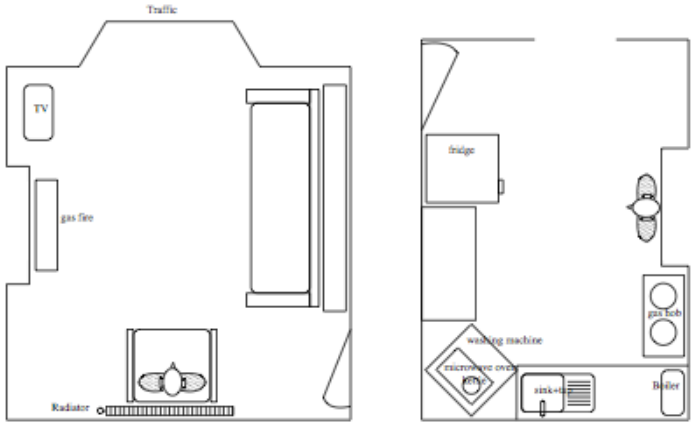
Brown – Human
Blue – Model combination
Green – Fragment decoding

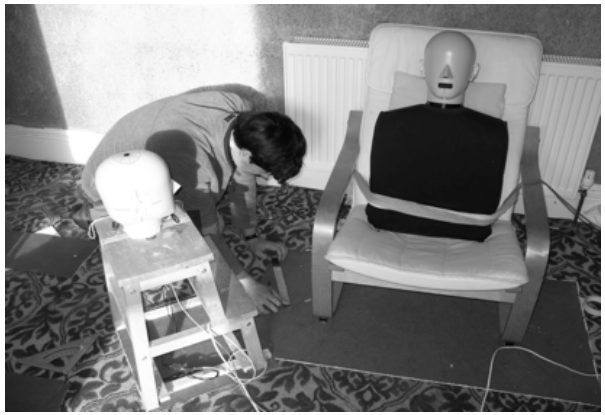


Super-human multi-talker speech recognition: A graphical modeling approach. J R Hershey, S J Rennie, P A Olsen, T T Kristjansson, *Computer Speech & Language*, Elsevier, 2010.

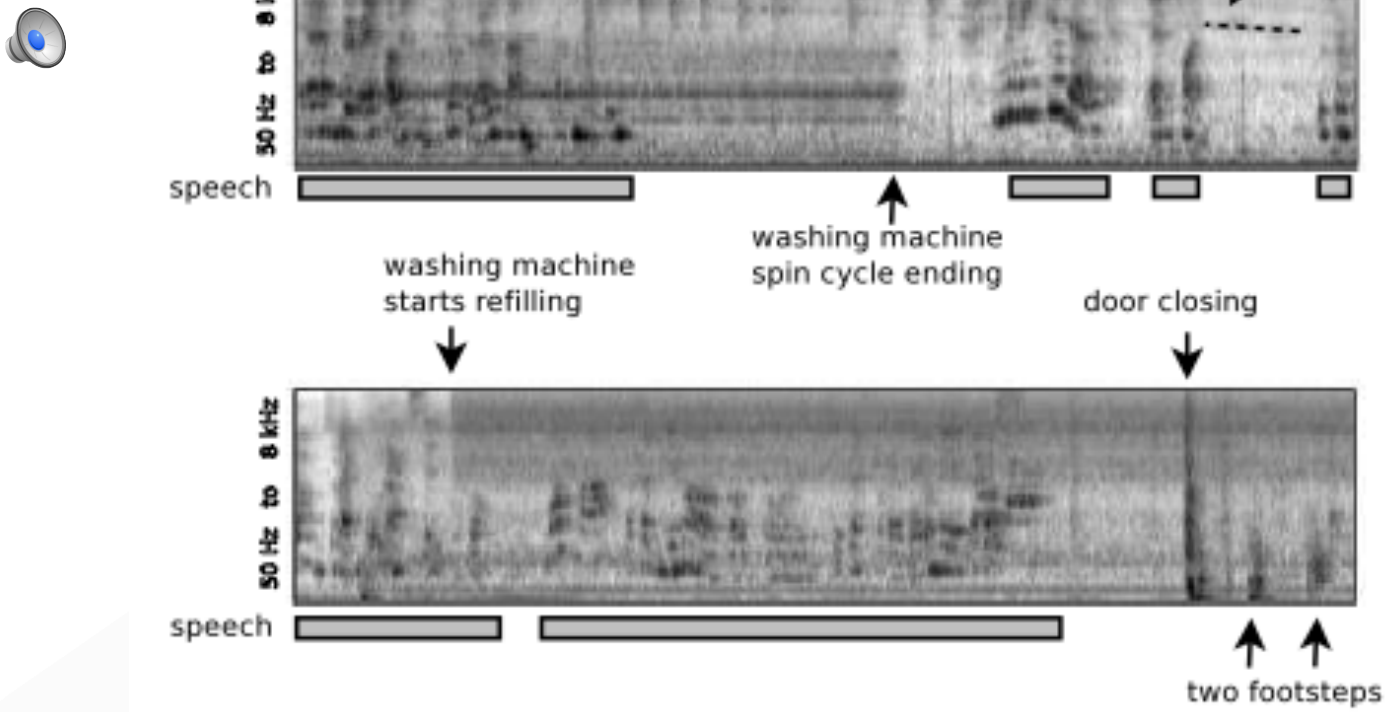
PASCAL CHiME speech separation and recognition challenge

- Speech in `multisource' noise environment.
- Binaural data recorded in noisy family home.
- Reverberant Grid utterances added at SNRs -6 to 9 dB.
- 12 international teams competing in evaluation.





The data is `acoustically cluttered'.

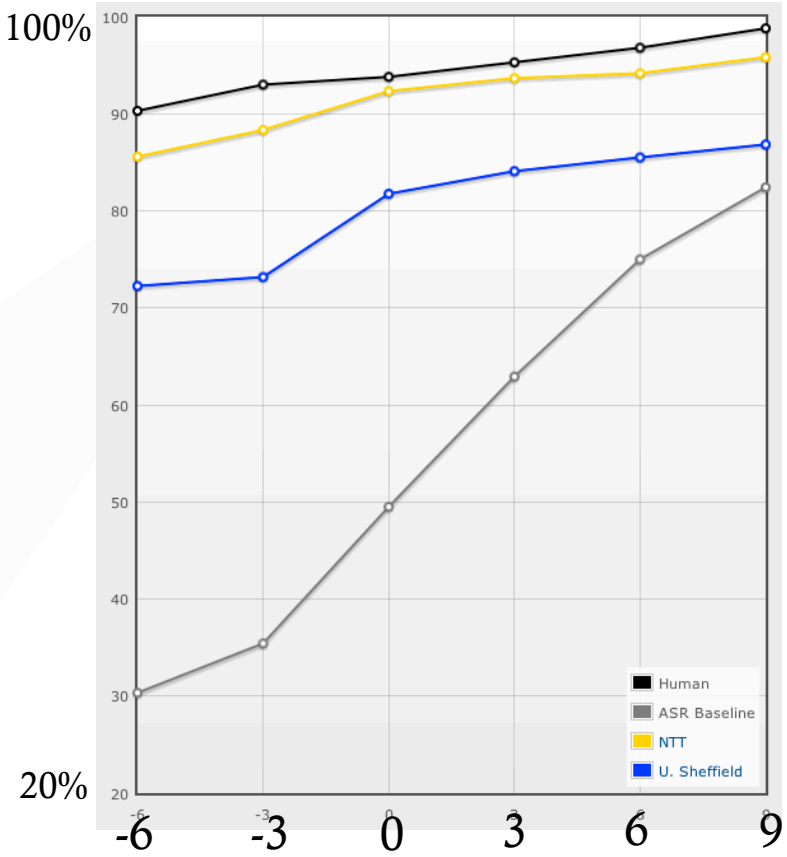


Performance

Fragment decoding system (blue) performs better than unrobust baseline (gray) …

… but well below human performance (black)

… and not as well as techniques which directly model the noise (yellow).



Summary

- A probabilistic model of ASA has been presented and demonstrated in application to robust ASR.
- The model differs from more conventional model combination approaches by only explicitly modelling the foreground source.
- Primitive grouping rules are embedded in a segmentation model which defines the space of valid foreground/background segmentations.
- The system operates by simultaneously searching over the space of valid segmentations and foreground descriptions.
- The framework has been applied to robust ASR but is sufficiently general that it has potential application to non-speech-based scene analysis tasks.

Closing questions

- How to effectively interface probabilistic grouping rules and schema-driven processes ?
- How much information about the background is used during interpretation of the foreground ?
- In situations where there are strong background models how should grouping constraints and models be combined ? (constrained model combination)

References

For more details see,

- http://staffwww.dcs.shef.ac.uk/people/J.Barker/publications.html

Acknowledgements

Thanks to colleagues,

- Ning Ma – fragment decoding, CHiME and simultaneous speech separation evaluations.
- Heidi Christensen – CHiME corpus construction

The End