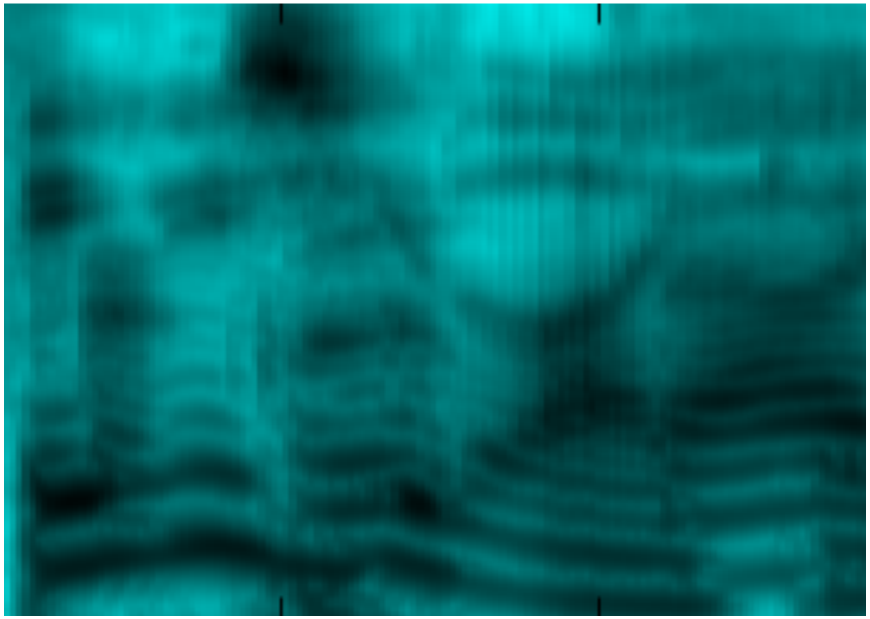
10 + 1 perspectives on speech separation and identification in listeners and machines

Martin Cooke

Speech and Hearing Research
Department of Computer Science
University of Sheffield
http://www.dcs.shef.ac.uk/~martin

I: Hard-core primitive auditory scene analysis

O Organisational cues in target speech



Principle: a sound mixture decomposed at the auditory periphery can be reassembled into its constituent sources by the application of grouping principles such as harmonicity, onset synchrony, continuity, etc.

Models: Parsons (1976), Lyons (1983), Stubbs & Summerfield (1988), Cooke (1991), Mellinger (1991), Brown (1992), Denbigh & Zhao (1992), Brown & Cooke (1994), Wang & Brown (1999), Hu & Wang (2002), …

Issues

- How to combine cues
- Grouping is not all-or-nothing
- Different thresholds for different tasks (Darwin)
- No really successful model of sequential grouping

Source property		Potential grouping cue	Illustrations	Notes
Starts and ends of events (common onset/offset)		Synchrony of transients across frequency regions	Effect of onset asynchrony on syllable identification (Darwin, 1981) and pitch perception (Darwin and Ciocca, 1992)	Offset generally weaker than onset.
Temporal modulations	slow	Correlation among envelopes in different frequency channels	Comodulation masking release (Hall *et al.*, 1984)	Common frequency modulation may lead to common amplitude modulation as energy shifts channels (Saberi and Hafter, 1995)
	fast, periodic	Channel envelopes with periodicity at f_0 (unresolved harmonics)	Segregation of two-tone complex by AM phase difference (Bregman *et al.*, 1985)	
		Harmonically-related peaks in the spectrum (resolved harmonics)	Mistuning of resolved harmonics (Moore *et al.*, 1985); effect on phonetic category (Darwin and Gardner, 1986)	
		Periodicity in fine structure (resolved and unresolved harmonics)	Perception of 'double vowels' (Scheffers, 1983)	Basis for autocorrelation models (Patterson, 1987; Meddis and Hewitt, 1991)
Spatial location		Interaural time difference due to differing source-to-pinna path lengths	Vowel identification (Hukin and Darwin, 1995). Strongest effect if direction is previously cued.	Evidence that suggests role of ITD is limited (Shackleton and Meddis, 1992) or absent (Culling and Summerfield, 1995b)
		Interaural level difference due to head shadowing	Noise-band vowel identification (Culling and Summerfield, 1995b)	
		Monaural spectral cues due to pinna interaction	Localization in the sagittal plane (Zakarauskas and Cynader, 1993)	Has not been investigated for complex, dynamic signals such as speech.
Event sequences		Across-time similarity of whole-event attributes such as pitch, timbre etc.	Sequential grouping of tones (Bregman and Campbell, 1971); sequential cueing (Darwin *et al.*, 1989, 1995)	
		Long-interval periodicity	Perception of rhythm	By-product of very-low-frequency 'spectral' analysis (e.g. Todd 1996)?
Source-specific		Conformance to learned patterns	Sine-wave speech (Remez *et al.*, 1981)	

From Cooke & Ellis (2001) Speech Communication

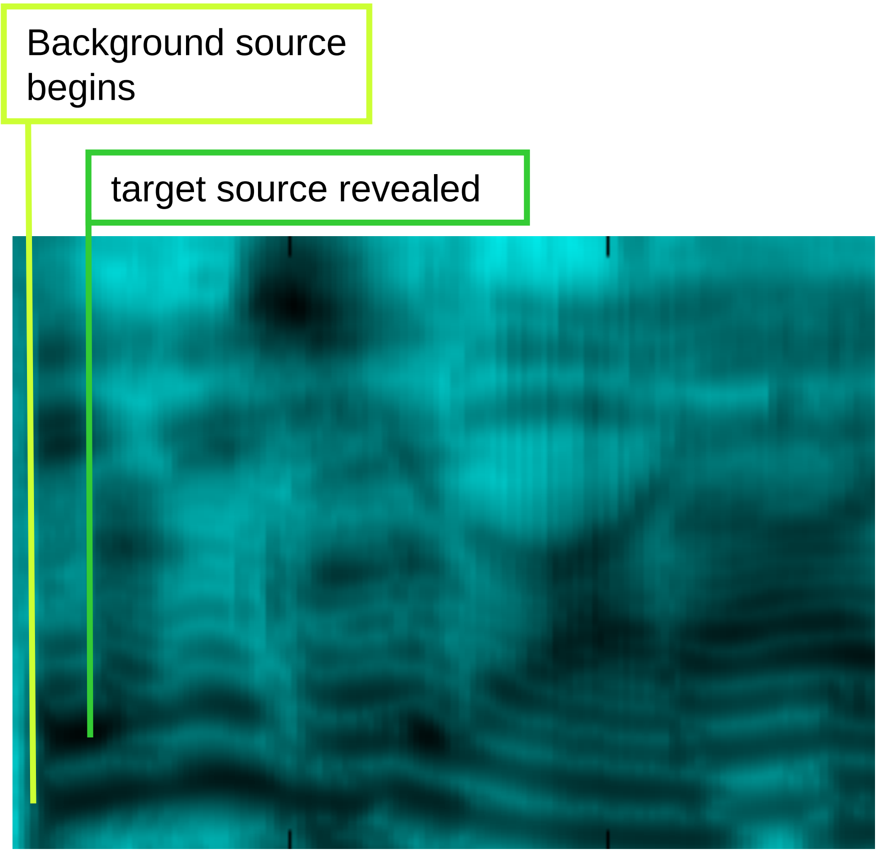
10 years of progress in primitive computational auditory scene analysis

	Original mix	Automatic separation systems		
		Cooke (1991)	Wang & Brown (1999)	Hu & Wang (2002)
Speech + telephone				
2 talkers (m/m)				
2 talkers (m/f)				

II: Full primitive auditory scene analysis

O Organisational cues in target speech

O Organisational cues in background



Principles

(i) grouping cues in the background can help unmask the target speech

(ii) unexpected energy while tracking one source can reveal the presence of another source (Bregman's old+new principle)

(iii) the residue left after extracting one or more sources can be processed to reveal further sources

Status: perceptual evidence for the power of background periodicity in helping identify the foreground

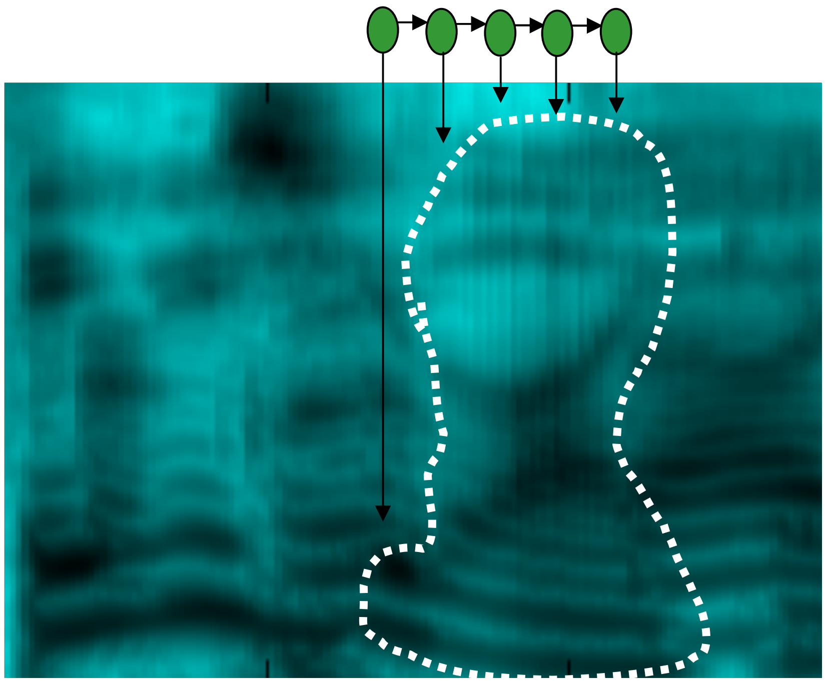
Models

(i) Cancellation models of double vowel perception (Lea, 1992, de Cheveigné, 1993++)

(ii) Residue models (eg Nakatani et al, 1998)

III: Speech is special

- O Organisational cues in target speech
- O Organisational cues in background
- O Models for target speech



Principle: speech identification processes have privileged access to the mixture signal and take what they need for classification

"*Speech is beyond the reach of Gestalt grouping principles*" (Remez et al, 1994)

Models: could actually work in practice but yet to be demonstrated computationally

Issues

- Listeners have difficult identifying speech mixtures when potential cues for organisation are degraded (cocktail party sine-wave speech)

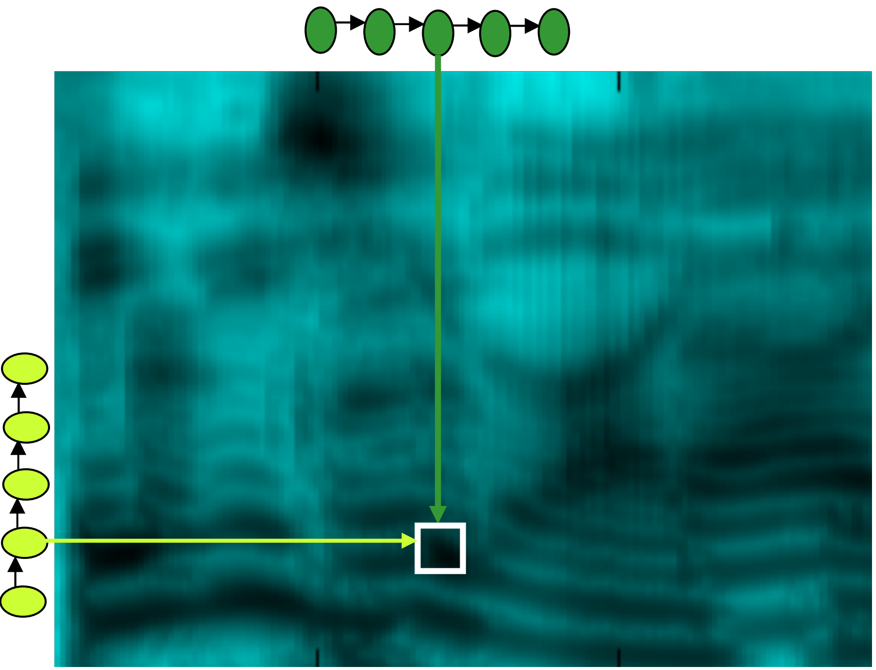
IV: Hard-core model-based explanation

O Organisational cues in target speech

O Organisational cues in background

O Models for target speech

O Models for background



Principle: all energy in the mixture can be explained by an appropriate combination of prior models for all sources present at any moment.

Models

- HMM decomposition (Varga & Moore, 1990)
- Parallel Model Decomposition (Gales & Young, 1993)
- MaxVQ (Roweis, 2001)

Issues

- Need to know how many sources are present at each time
- Need models for all possible sources
- Computationally complex for N > 2, and too complex in practice for N = 2 if the background source is non-trivial

V: Full Auditory Scene Analysis account

- O Organisational cues in target speech
- O Organisational cues in background
- O Models for target speech
- O Models for background

Principle: source separation and identification requires the action of both innate, primitive, grouping principles *and* learned schemas

Champions: Bregman; application to speech (Darwin)

Models: to some extent, the systems of Weintraub (1985) and Ellis (1996) applied bottom-up and top-down influences

Issues

- Very few CASA systems have exploited models for the speech target
- Level(s) at which primitive and schema processes could be integrated/conflicts resolved is not clear

VI: Energetic masking

- O Organisational cues in target speech
- O Organisational cues in background
- O Models for target speech
- O Models for background
- O **Energetic masking**

Principle: the intelligibility of speech in a mixture is largely determined by peripheral masking

Models: articulation index (French & Steinberg, 1947; Kryter, 1962); Speech Intelligibility Index (ANSI S3.5, 1997); Speech Transmission Index (Steeneken & Houtgast, 1980; 1999); Speech Recognition Sensitivity (Musch & Buus, 2001); Spectro-Temporal Modulation Index (Elhilali, Chi & Shamma, 2003)

Issues

- Detection of the unmasked portions
- AI, STI etc are macroscopic models of intelligibility

VII: Linguistic masking of speech by speech

- Organisational cues in target speech
- Organisational cues in background
- Models for target speech
- Models for background
- Energetic masking
- Informational masking

Principle: the intelligibility of speech in a mixture is determined not only by audibility but by the degree to which the background and foreground can be confused

'*Perceptual masking*' (Carhart et al, 1969)

Recent studies: Brungart et al (2001+); Freyman et al (2001+)

Models: None, but a prototype model of energetic and informational masking was presented by Barker & Cooke at the Hanse meeting based on competition within a speech decoder

Issues:

- Informational masking is too much of a catch-all term; factors other than foreground/background confusions may have a role over and above energetic masking eg distractors

VIII: Stationarity

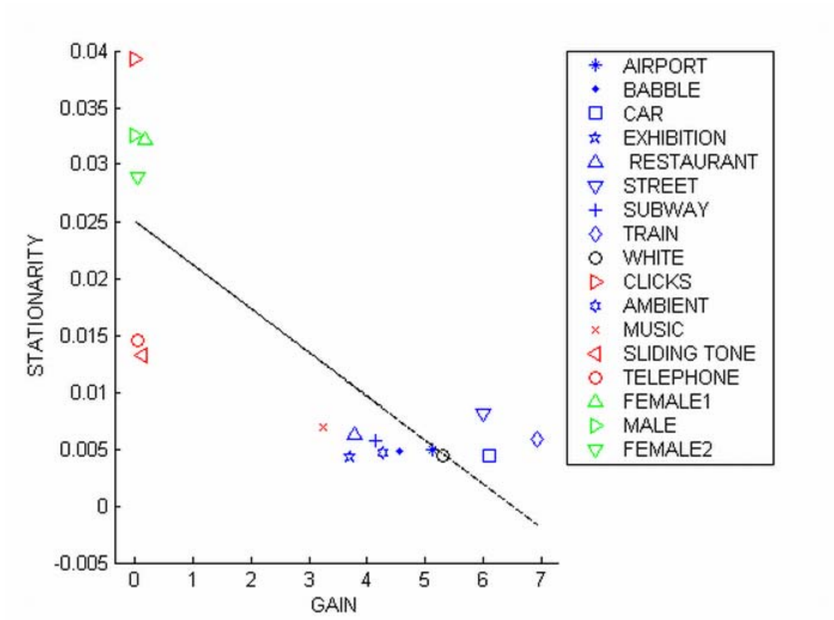
- O Organisational cues in target speech
- O Organisational cues in background
- O Models for target speech
- O Models for background
- O Energetic masking
- O Informational masking
- O **Stationarity of background**

Principle: stationary backgrounds are easily compensated

Models: lots – spectral subtraction (Boll), minimum statistics (Martin, 1993), histogram partitioning (Hirsch & Ehrlicher, 1995)

Issues

- While this is a bad approximation to everyday backgrounds, many models/algorithms embody this constraint implicitly or otherwise
- Must be used in conjunction with other processes
- Not clear to what extent listeners exploit stationarity (perhaps implicitly via enhancement of dynamics)



IX: Independence

- O Organisational cues in target speech
- O Organisational cues in background
- O Models for target speech
- O Models for background
- O Energetic masking
- O Informational masking
- O Stationarity of background
- O **Source independence**

Principle: exploit statistical independence of sources (Comon, 1994)

Models: Bell & Sejnowski (1995); Lee et al (1997); Smaragdis (2003)

Issues

- Reverberant energy correlated with direct energy
- Listeners manage with 1 or 2 sensors regardless of the number of sources
- Debate over whether "the cocktail party problem is beyond scope of ICA"

  "One of the original motivations for ICA research was the cocktail-party proble […] blind separation of audio signals is, however, much more difficult than one might expect […] due to these complications, it may be that prior information, independence and nongaussianity of the source signals are not enough" (Hyvarninen et al, 2001, *Independent Component Analysis*)

X: Sparsity and redundancy

- O Organisational cues in target speech
- O Organisational cues in background
- O Models for target speech
- O Models for background
- O Energetic masking
- O Informational masking
- O Stationarity of background
- O Source independence
- O Sparsity and redundancy

Principles

(i) spectro-temporal modulations of speech (and possibly the background too) allow relatively clear but sparse views of the target;

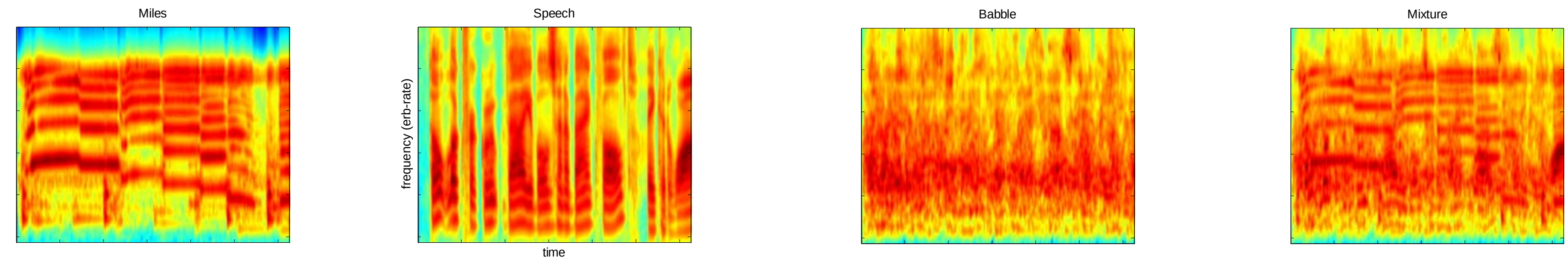
(ii) redundancy of speech makes identification possible in spite of missing information.

Models: missing data (Cooke, 1994, 2001; Raj et al, 1998, 2004; Seltzer et al, 2004); multiband ASR (Bourlard & Dupont, 1996); non-negative matrix decomposition (Smaragdis, 2003)

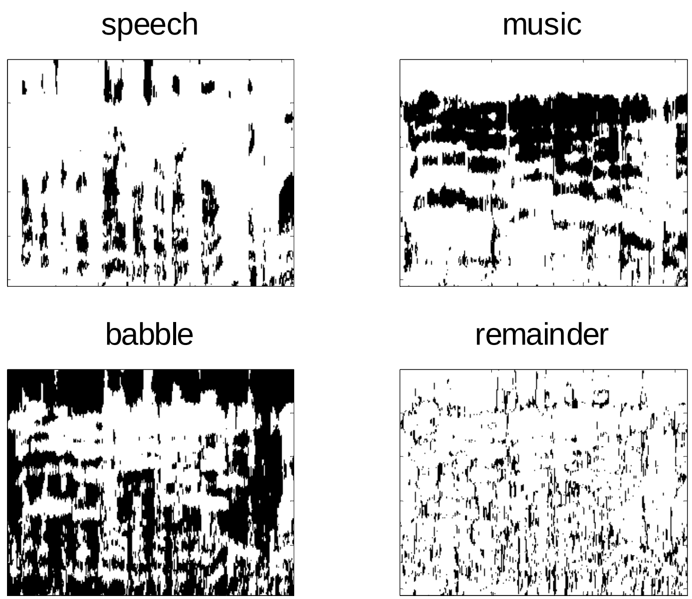
Issues

- detection and integration of sparse information in speech

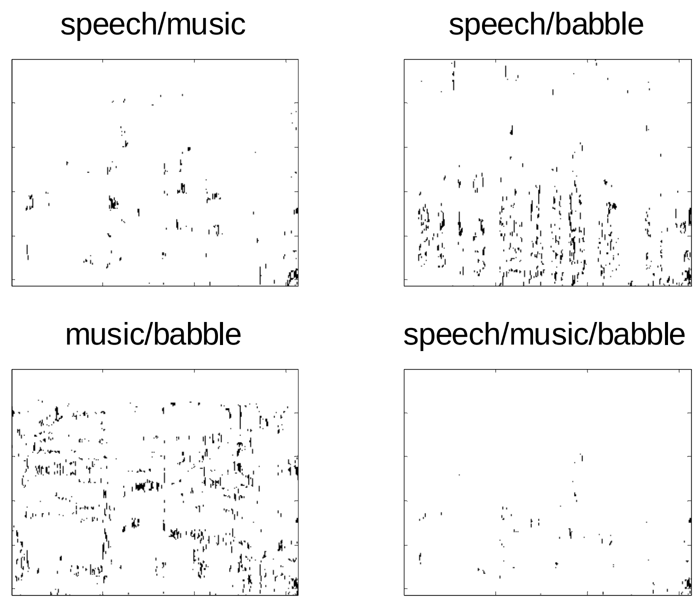
Sparse information in mixtures



Energy within 3 dB of value in mix



Energy within 3 dB of other source



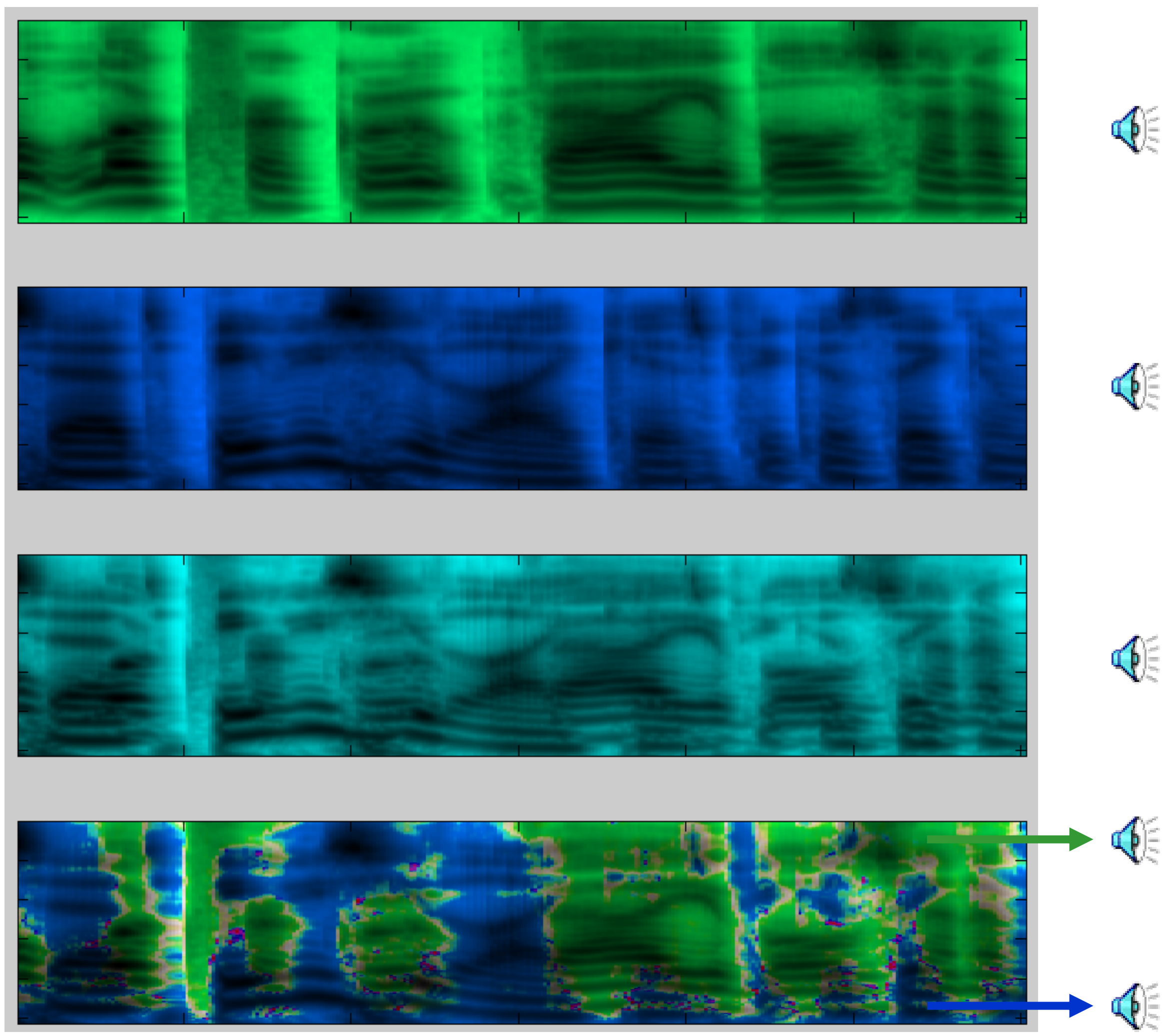
Listening to sparse information

Talker 1

Talker 2

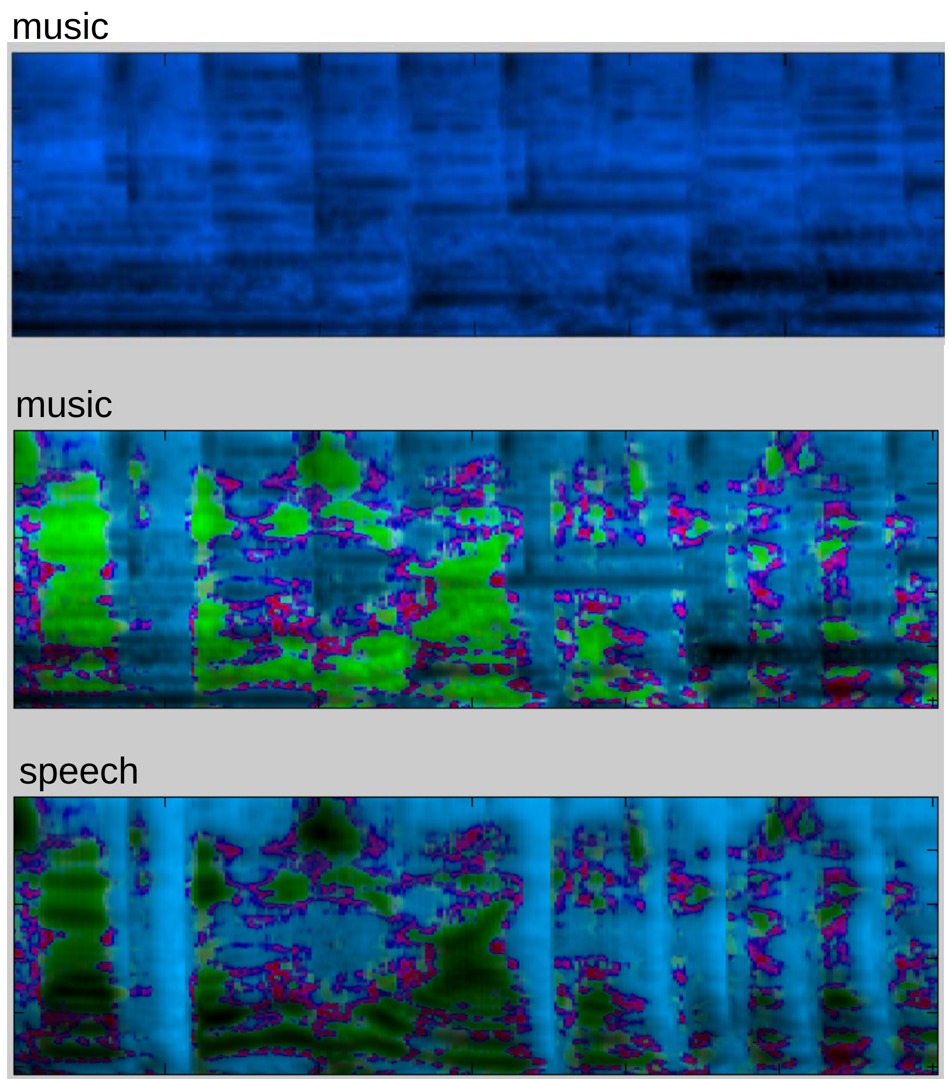
Mix@0dB

One or other
talker dominant



With added noise

Sparse-sampling of music



Green = speech
shaped regions

Summary of possible ingredients

- O Organisational cues in target speech
- O Organisational cues in background
- O Models for target speech
- O Models for background
- O Energetic masking
- O Informational masking
- O Stationarity of background
- O Source independence
- O Sparsity and redundancy

Auditory scene analysis

Speech perception

Speech intelligibility

Signal processing/robust ASR

Statistics, information theory, machine learning

Understanding the contribution of each ingredient: a multispeaker babble continuum

How many people to invite to the cocktail party?

Task

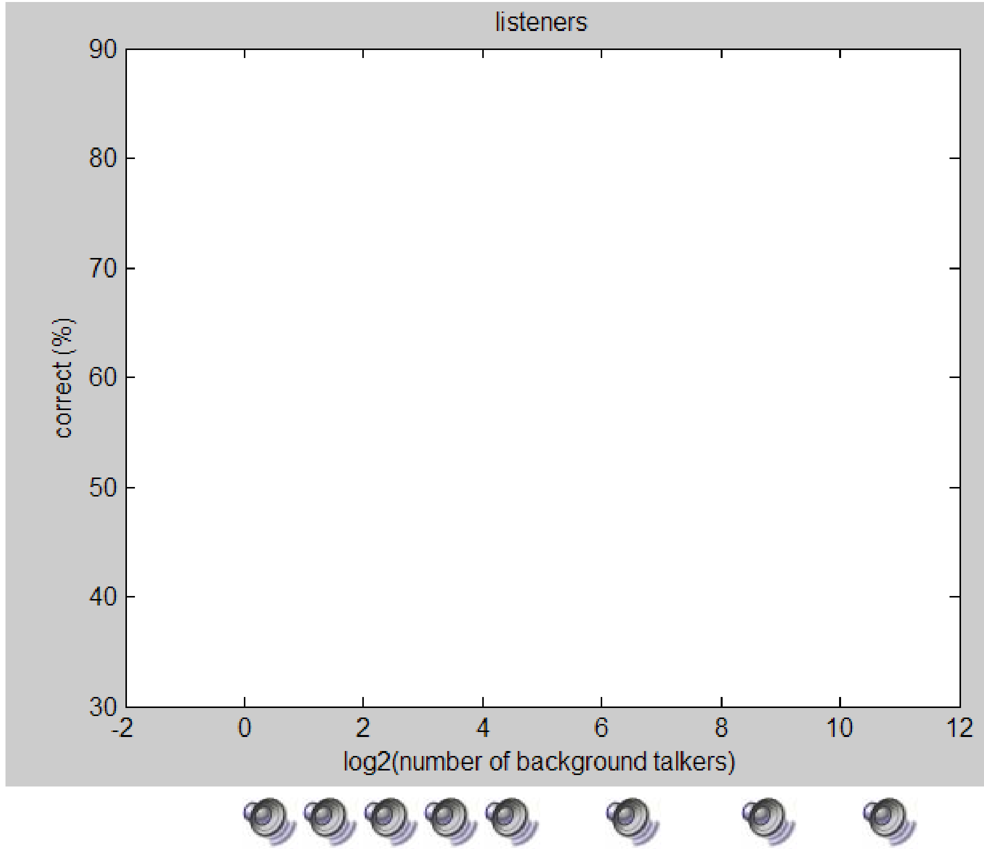
identify VCVs in N-speaker babble noise, for various N

As N tends to infinity

- **Increase** in energetic masking
- **Increase** in sparsity as spectro-temporal dips are filled
- Background grouping cues become **less effective**
- Background schemas become **less useful**

But:

- Babble becomes less speech-like leading to a **decrease** in informational masking?
- Signal becomes **more stationary**



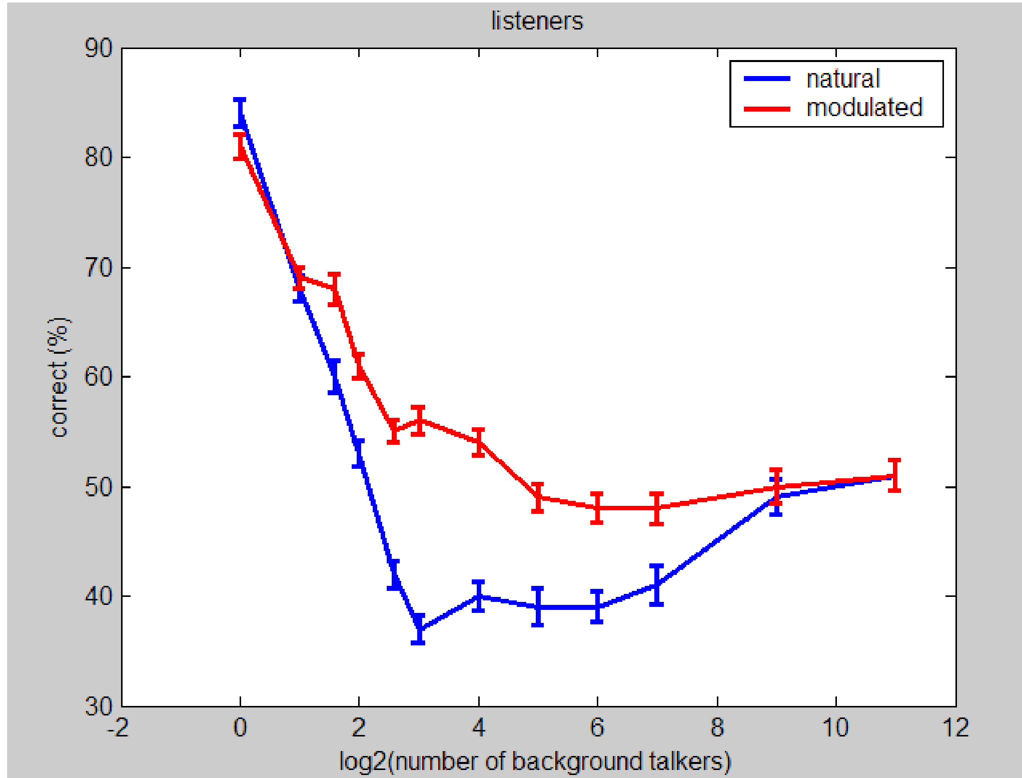
Factoring of ingredients

Example

Compare n-speaker babble-modulated speech-shaped noise with n-speaker babble to reveal contribution of informational masking

Issues

- Energetic masking produced by speech-modulated noise is **not** identical to that produced by natural speech
- Informational masking is a catch-all concept
- In general, difficult to isolate each ingredient experimentally since they are not really independent



cf: Bronkhorst & Plomp (1995)

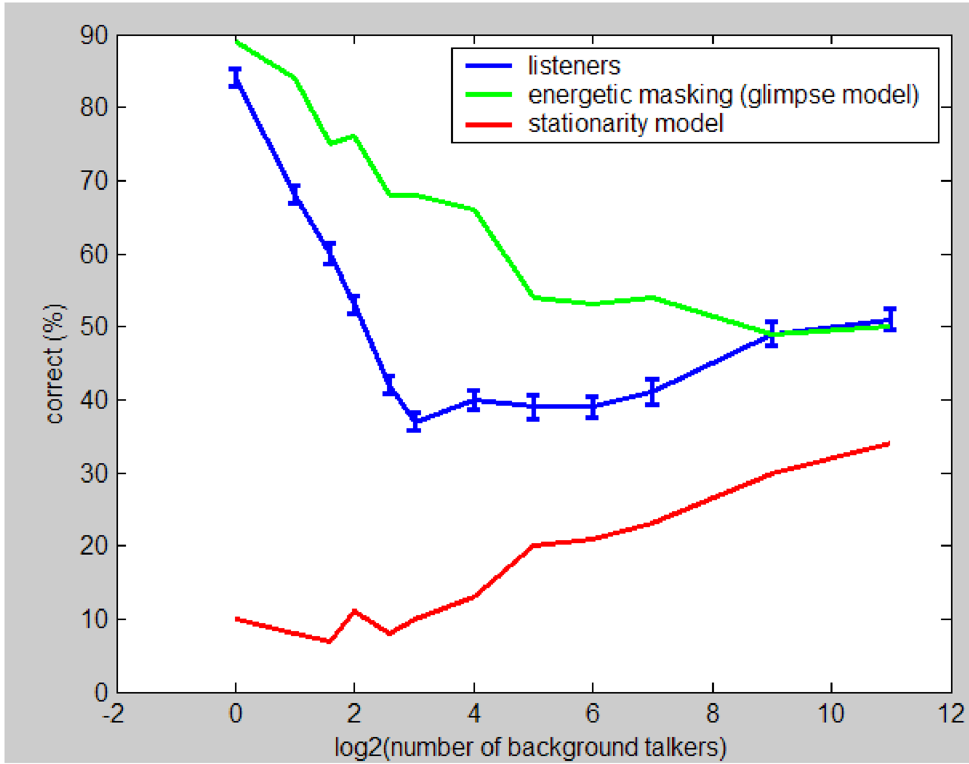
Modelling of ingredients

Examples

- Model of energetic masking
- Ideal spectral subtraction model given stationary estimate of interfering spectrum

Issues

- Not obvious how to construct and constrain models for all factors eg speech schemas
- Not clear how to combine models (a combined EM+stationarity model does **not** produce a dip)



XI: Glimpsing

- O Organisational cues in target speech
- O Organisational cues in background
- O Models for target speech
- O Models for background
- O Energetic masking
- O Informational masking
- O Stationarity of background
- O Source independence
- O Sparsity and redundancy

Principles:

(i) Sparsity permits listeners to glimpse clean views of the target source

(ii) Such glimpses can be quite large, suggesting that they may be detectable by the *local* application of primitive organisational cues

(iii) Models for the speech target help to integrate glimpses *sequentially*

Precursors: multiple looks (Viemeister & Wakefield, 1991; Hant & Alwan, 2003); double vowels -- Culling & Darwin (1994); dip listening (Peters et al, 1998); vowel identification (de Cheveigné & Kawahara, 1999); G & T (Assmann & Summerfield, 2004)

Model: Cooke (2003)

Issues

- Sufficiency of glimpses
- Glimpses detection
- Integration of glimpses

Sufficiency of glimpses

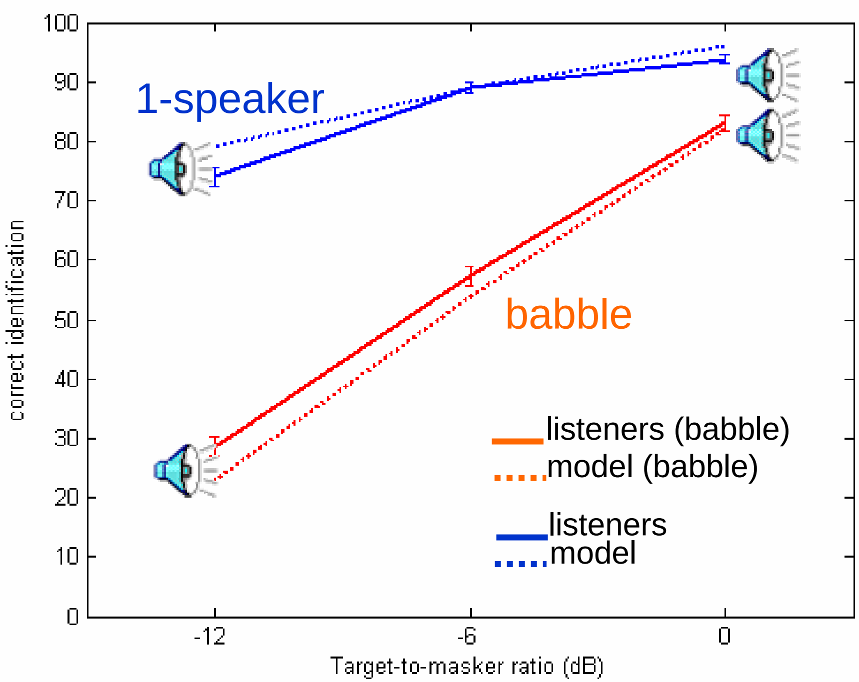
Procedure: use a computational model of speech perception and restrict its input to glimpses

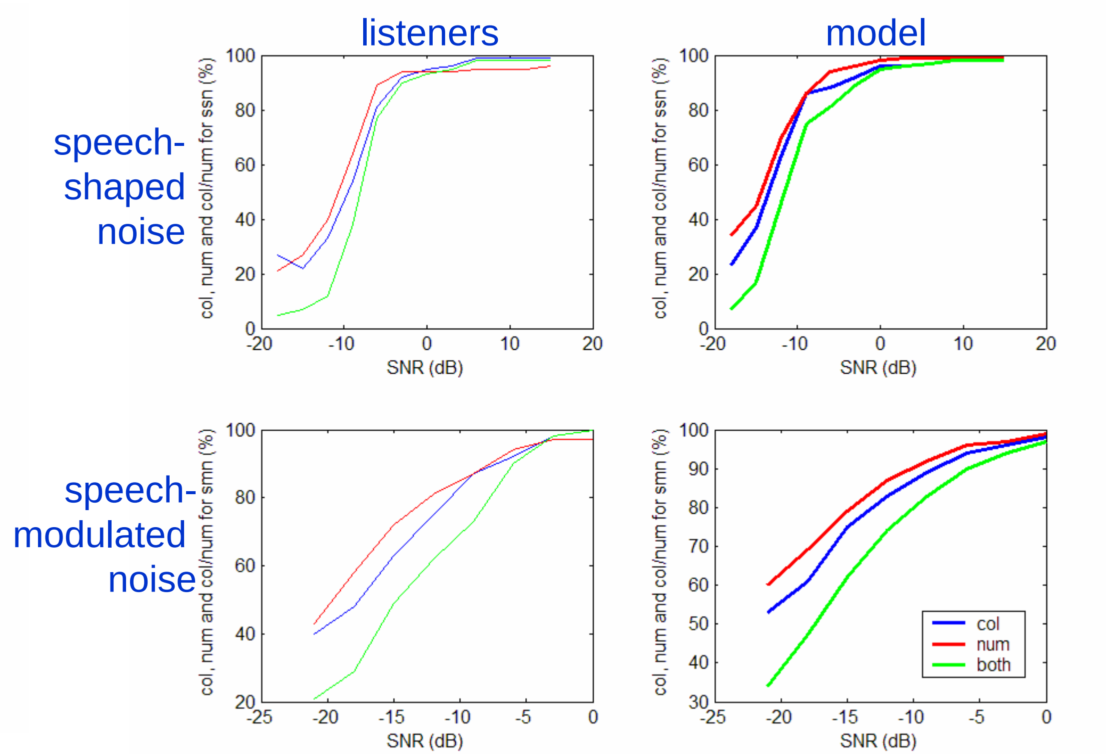
Task 1

- VCV intelligibility in noise
- Background ‘noise’ is N-speaker babble for N=1 and 8

Task 2

- CRM sentences
- Listeners’ data from Brungart (2001)



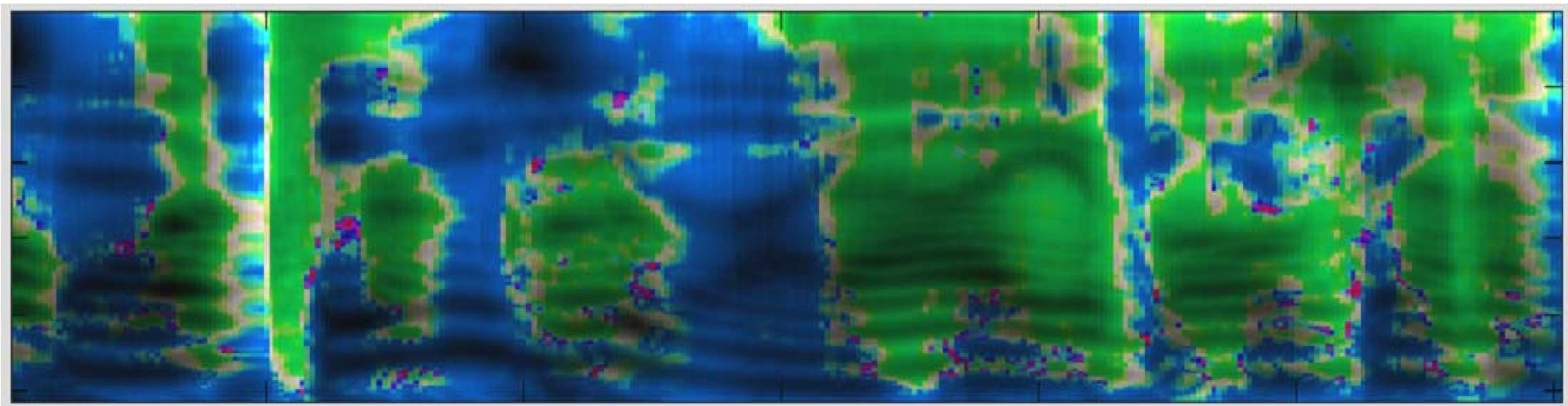


Glimpse detection via LASA?

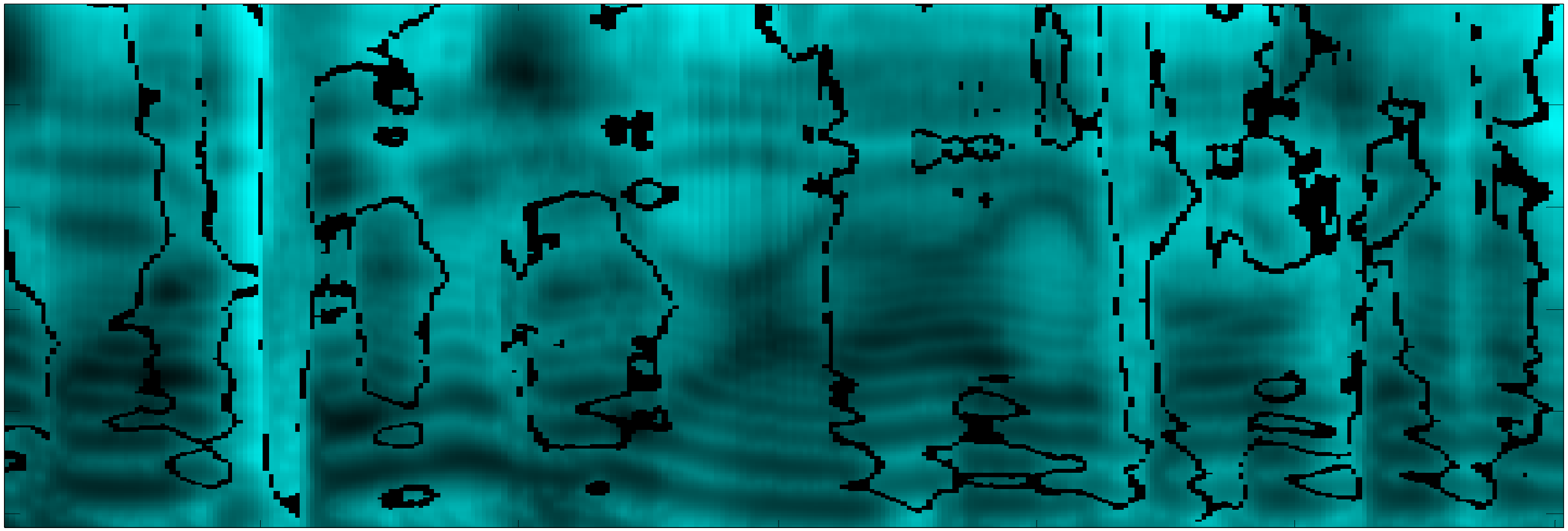
Role of auditory scene analysis may be ‘limited’ to

(i) Local organisation of the scene (harmonicity, common AM, etc)

(ii) Provision of a weak prior over glimpses (location, pitch continuity, etc



Glimpse integration:
the blind, multiple, partial jigsaws problem



Glimpse decoding solution: Barker, Cooke & Ellis (2004)

Some early results (1)

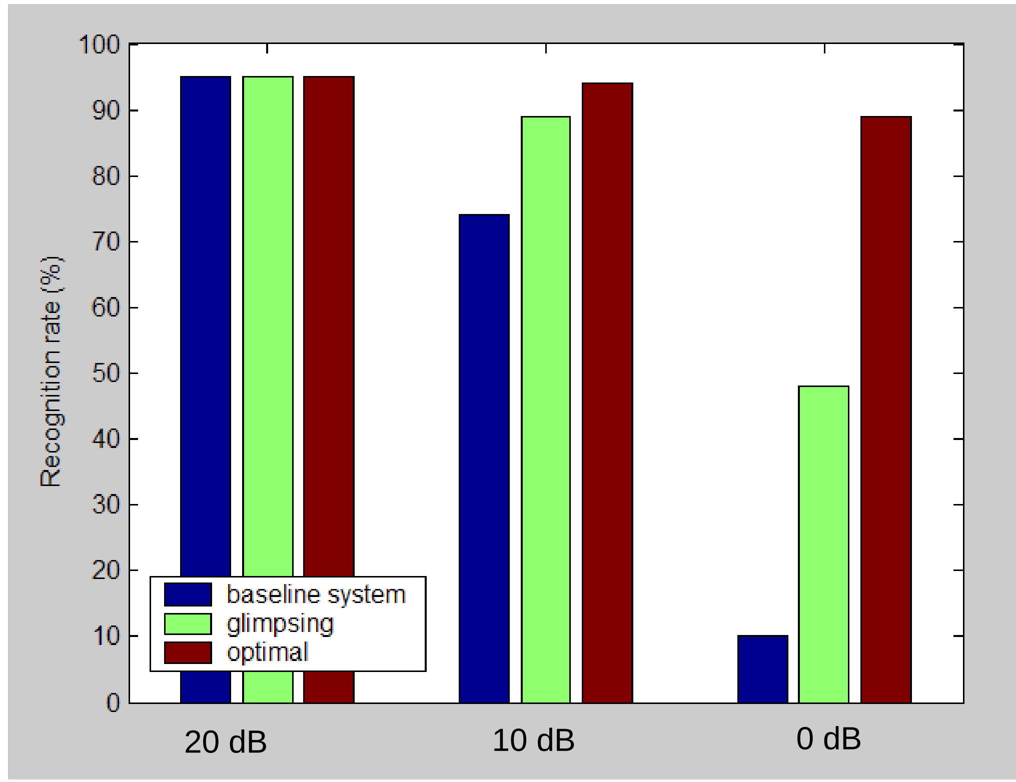
AURORA task: recognition of digit sequences in a background of factory noise backgrounds (stationary background + hammer blows, machine noise etc)

Key:

Optimal = best possible performance
using a glimpsing strategy

Glimpsing = automatically-determined
glimpses

Baseline = MFCC + CMN



Source: Barker, Cooke & Ellis (2004)

Some early results (2)

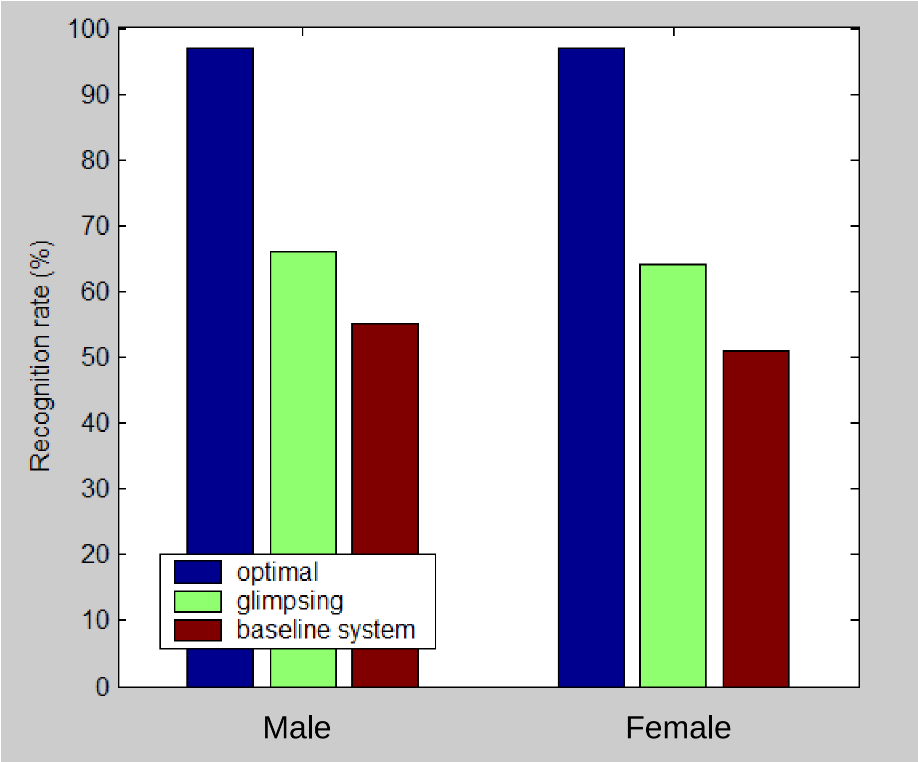
Recognition of digit sequences in a background of a competing talker (who is also speaking digits)

Key:

Optimal = best possible performance using a glimpsing strategy

Glimpsing = automatically-determined glimpses (using only harmonicity so far)

Baseline = MFCC + CMN



Source: Coy & Barker, in progress

Summary

- Different perspectives (ASA, speech is special, intelligibility, robust ASR, information theoretic, …) give rise to many possible *non-independent* ingredients of a solution to the speech separation problem

- Experimentally, difficult to tease them apart

- Listeners probably exploit many ingredients, but most theoretical and modelling accounts are based around one or two only ('silver bullet') – ASA is an exception

- The glimpsing account differs from traditional ASA:
  - Emphasis on local rather than global organisation
  - Information derived from glimpses can act as a weak global prior
  - Emphasis is on identification from sparse data rather than on separation