Separation of sound sources by humans and machines

Martin Cooke

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

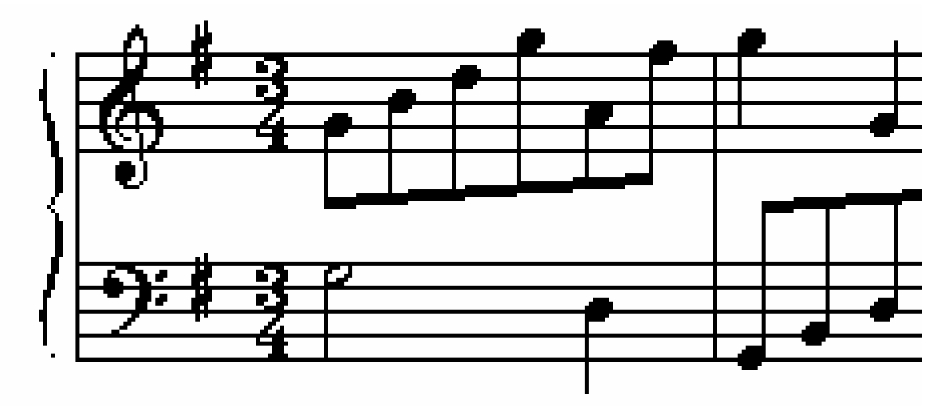
For a longer (160 slides) version of this tutorial, see my NIPS tutorial available at

http://www.dcs.shef.ac.uk/~martin/nips.ppt

Part I:

The auditory scene analysis problem

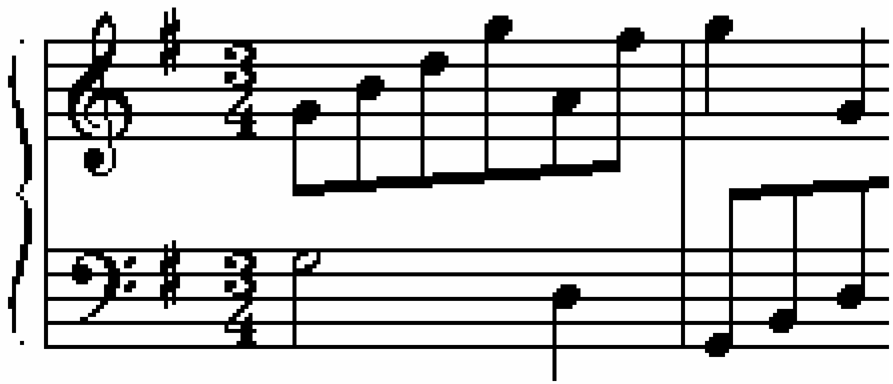
The scope of auditory perception



Source: Compay Segundo "Ahora me da pena"

The scope of auditory perception

Now try to identify each instrument/voice as it comes in and follow it for a while

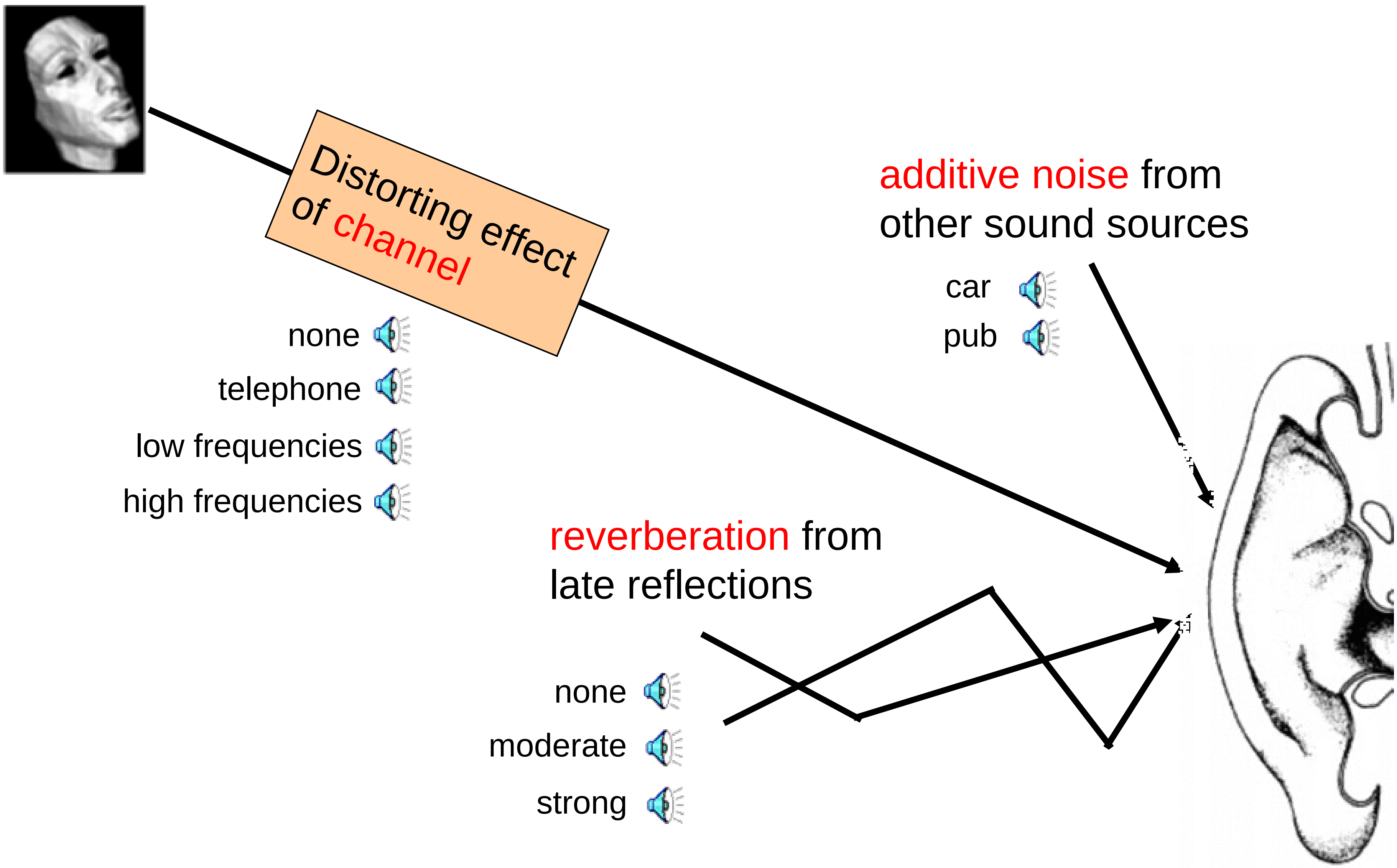


Source: Compay Segundo “Ahora me da pena”

Auditory perception answers these questions:

What/who?	Type of acoustic source eg talker, instrument, car engine, *Eg for speech*: message content, talker identity, age, gender, linguistic origin, mood, state of health, …
Where?	Location: left, right, up down Distance: promixity Environment: bathroom, concert hall, open space?
How many?	1, 2, more
Transmission channel?	Telephone, radio, …

Issues facing everyday speech communication (and associated technology)



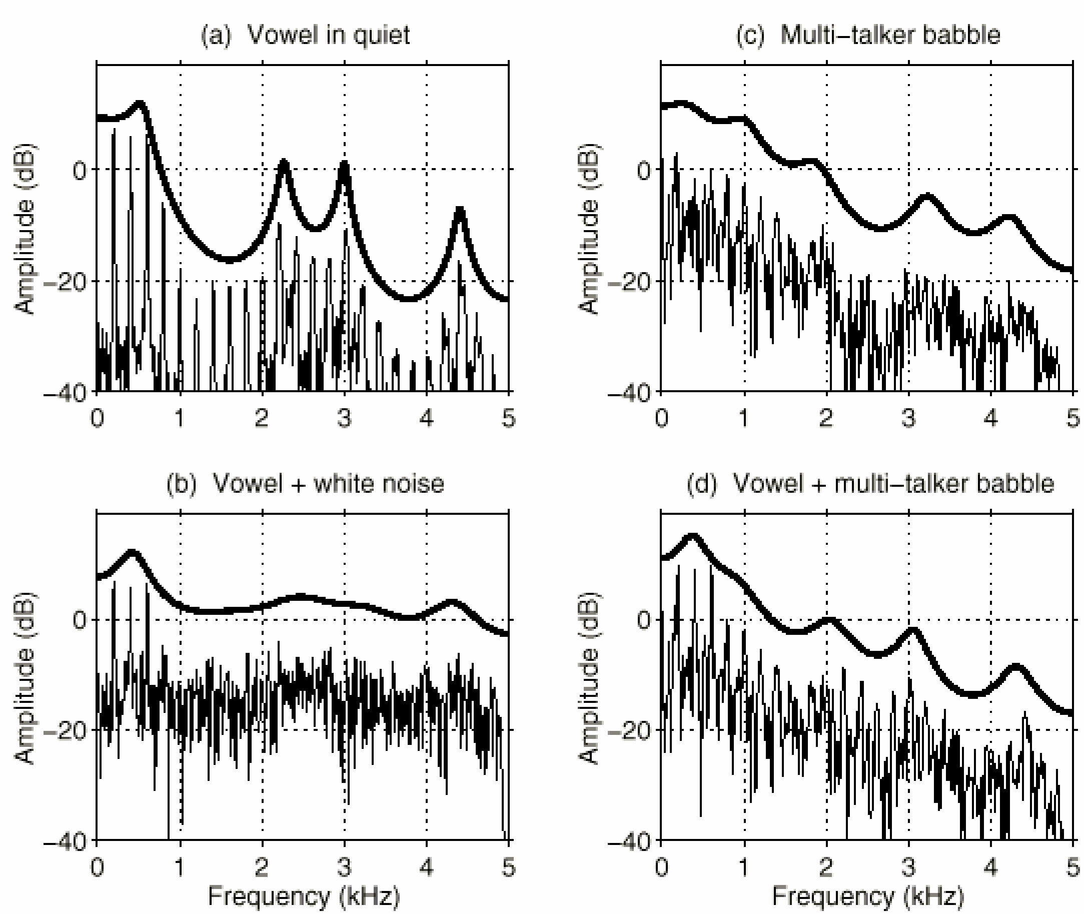
Effects of additive noise on speech

Fourier amplitude and LPC smoothed spectra of the vowel in “head”

white noise or multi-talker babble mixed at 0 dB SNR

in white noise, spectral contrast reduced, harmonicity obscured

in babble, less reduction of spectral contrast, better preservation of harmonics

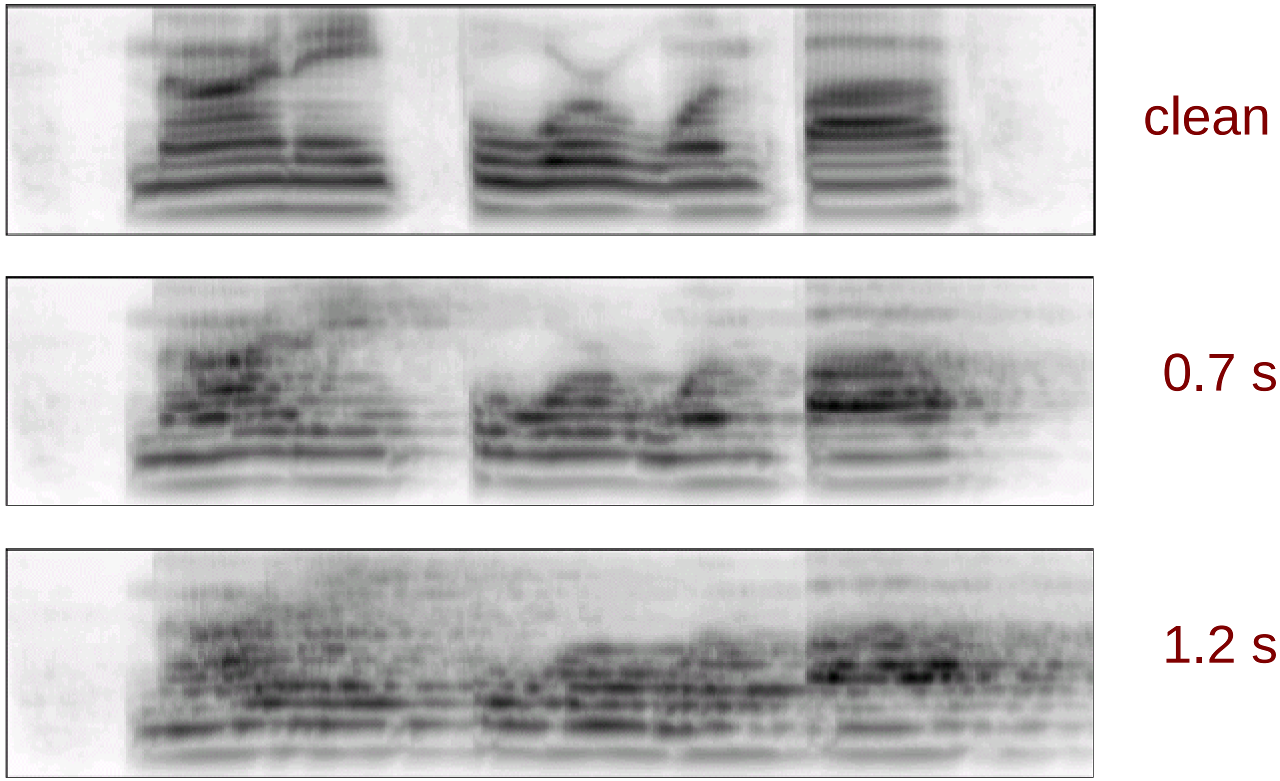


Source: Assmann & Summerfield (in press)

Effects of reverberation on speech

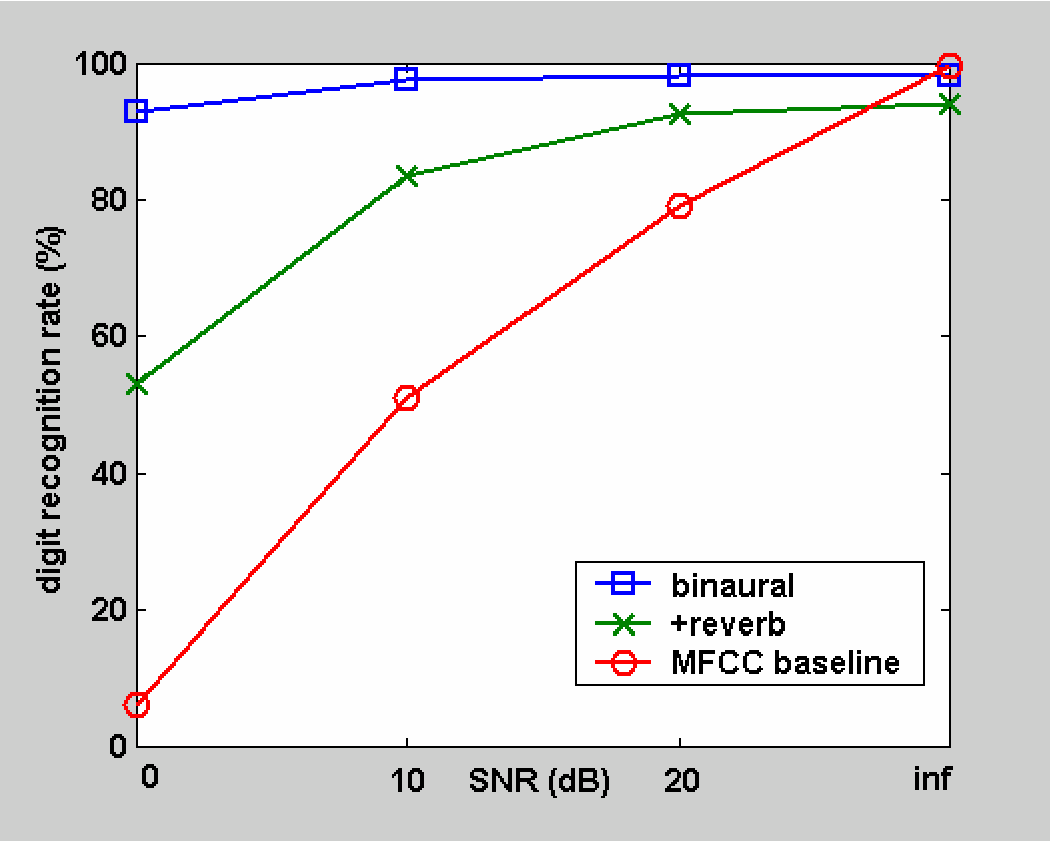
From Assmann & Summerfield (in press)

- Fills gaps associated with vocal tract closure
- Blurs onsets & offsets, reducing durational cues
- Extends noise bursts
- Flattens formant transitions in diphthongs & glides
- Removes evidence of amplitude modulation at pitch rate
- Preserves vowels



Source:synthetic data from Kalle Palomaki

Effect of reverb on ASR performance



Source: Palomaki, Brown & Wang (2004)

- AURORA task
- Missing data processing to handle reverberation
- Reverberation significantly reduces the binaural advantage

Effects of noise on speakers

normal

with broadband noise

Speech Adjustments	Source
increase in vocal intensity (about 5 dB increase in speech for every 10 dB increase in noise level)	Dreher and O'Neill (1957)
decrease in speaking rate	Hanley and Steer (1949)
increase in average F_0	Summers et al. (1988)
increase in segment durations	Pisoni et al. (1985)
reduction in spectral tilt (boost in high frequency components)	Summers et al. (1988)
increase in F1 and F2 frequency (inconsistent across talkers)	Summers et al. (1988) Junqua and Anglade (1990) Young et al. (1993)

Table 5.2. Summary of changes in the acoustic properties of speech produced in background noise (Lombard speech) compared to speech produced in quiet.

Source: Assmann & Summerfield (2004)

Auditory scene analysis

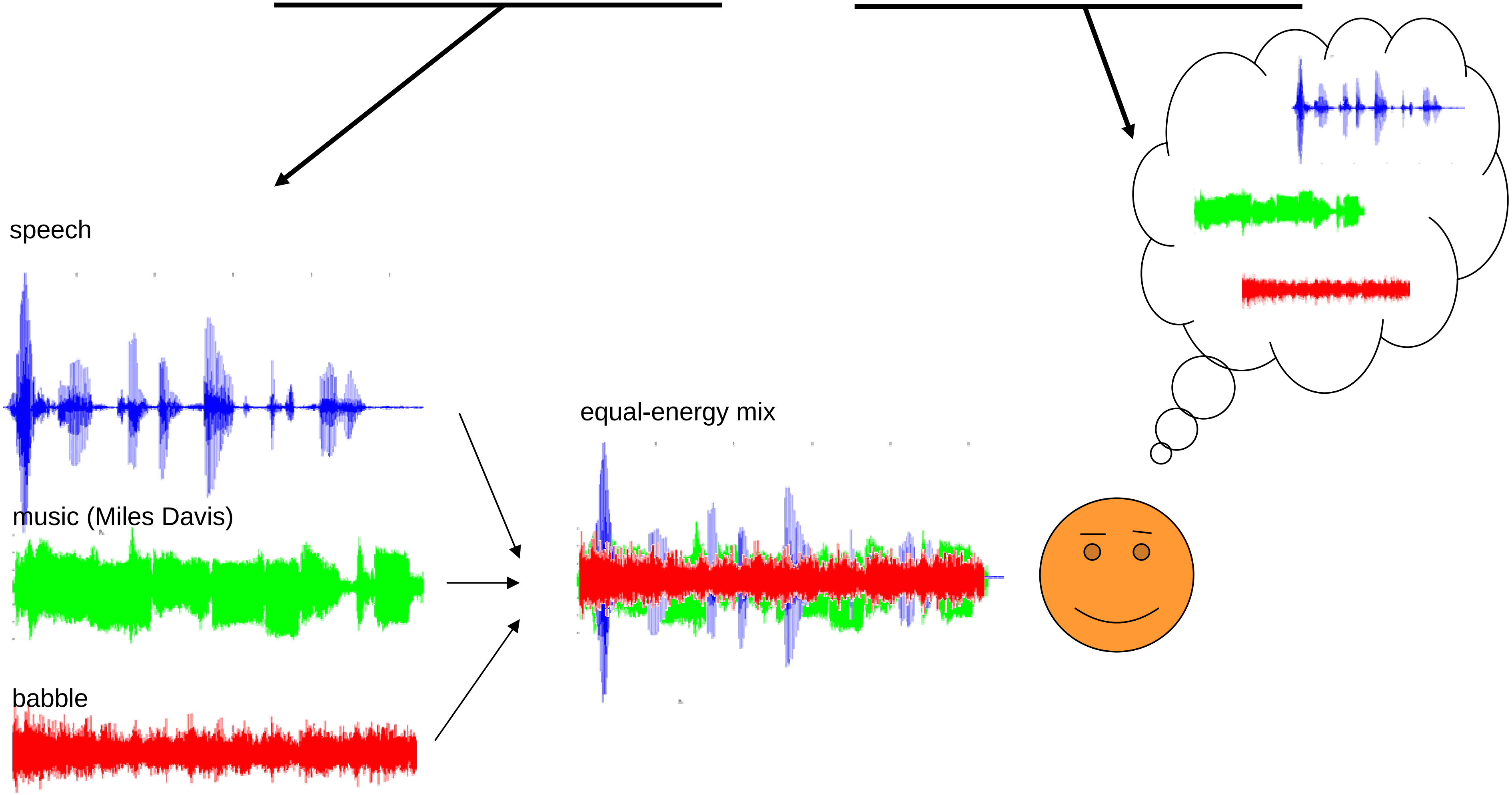
Key idea

acoustic signals are littered with cues which allow our ears and brain to form separated perceptual representations ('auditory streams') for each individual source

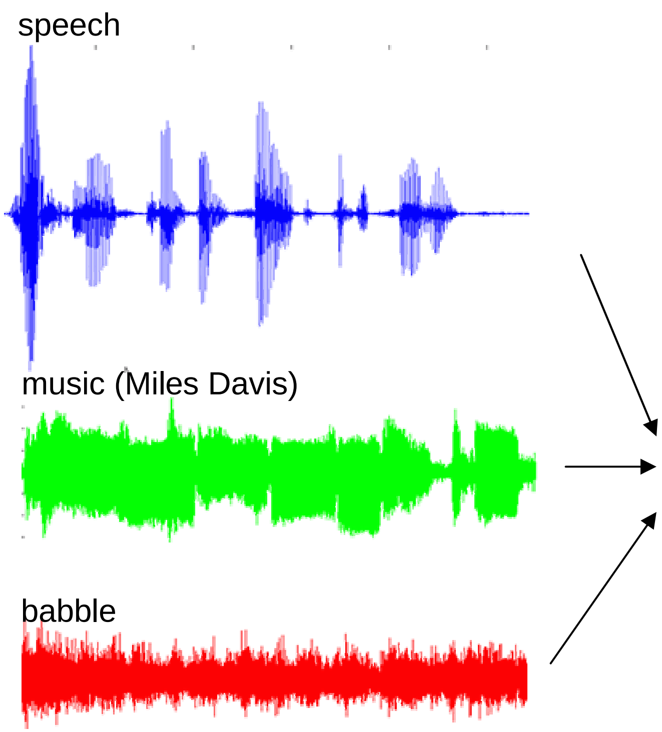
Bregman (1990) *Auditory Scene Analysis*, MIT Press

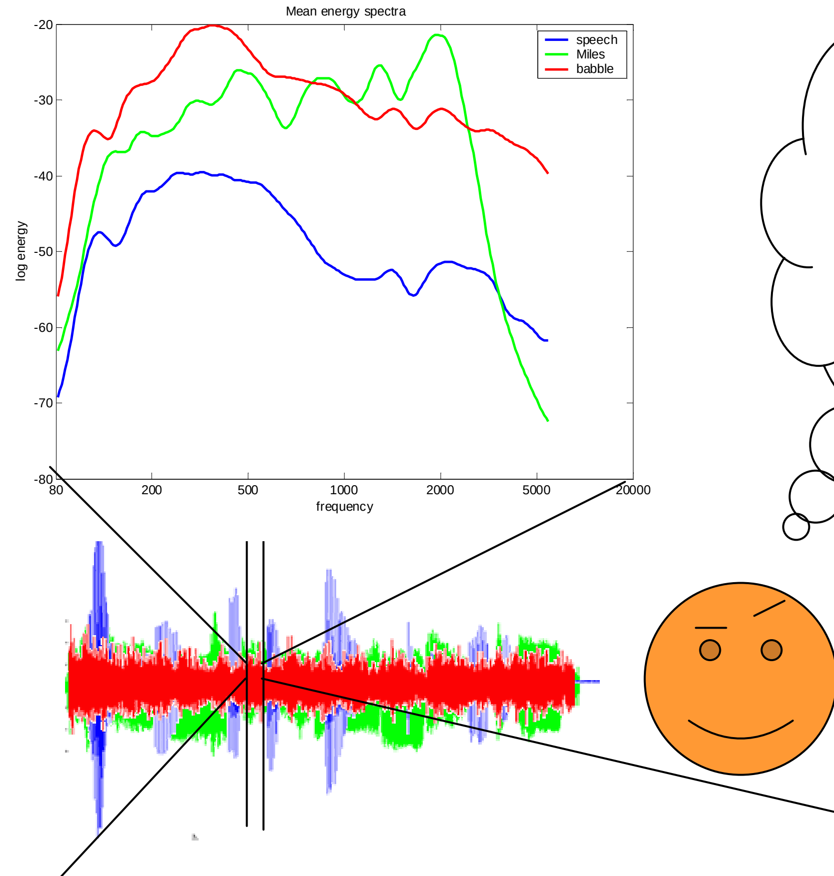
Cooke (1991) PhD developed a system for *Computational Auditory Scene Analysis*

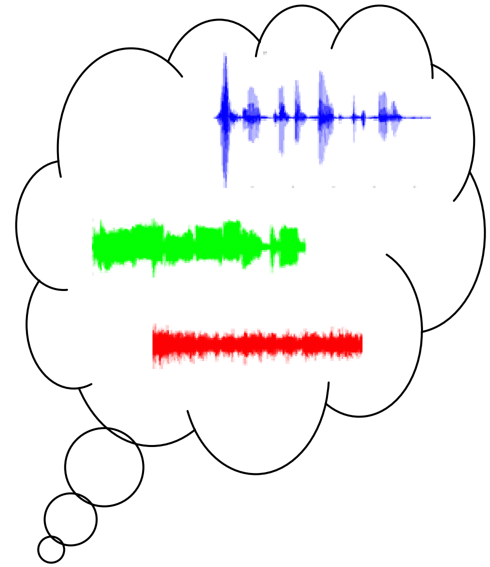
Acoustic sources and auditory streams

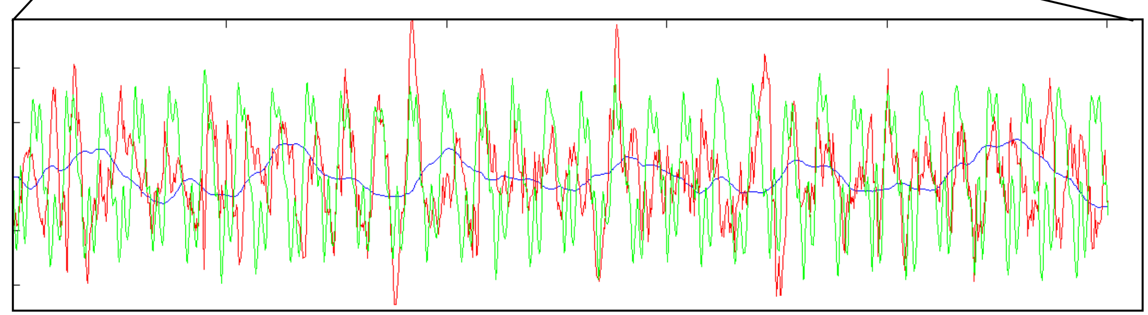


Acoustic sources and auditory streams







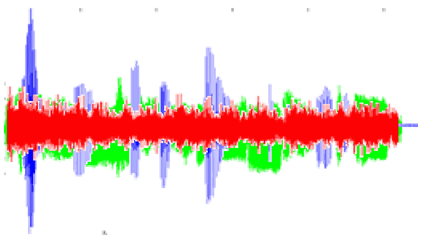


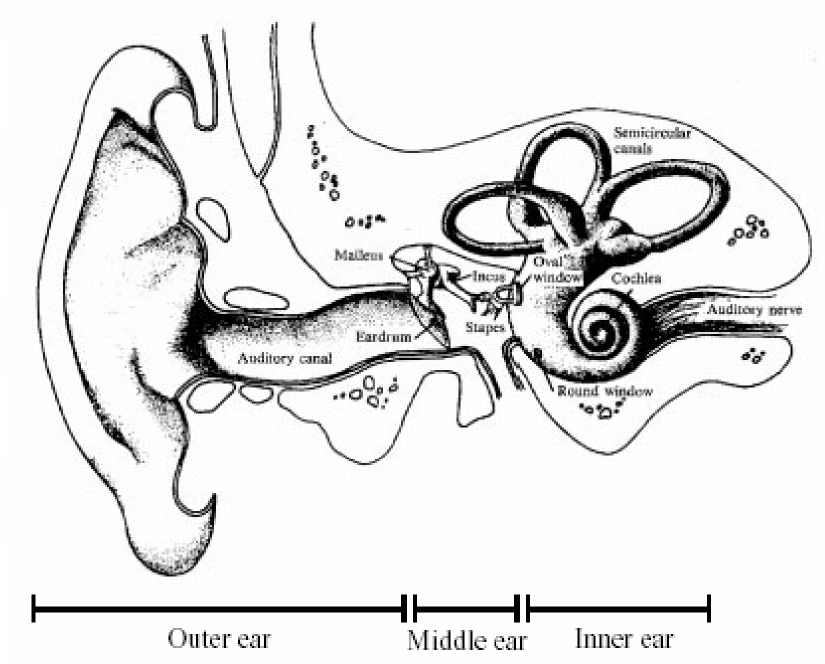
A problem: sources overlap in the time and frequency domains

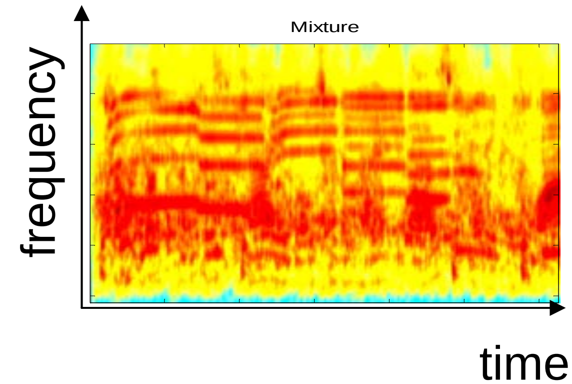
Part II:

Models of early processes in hearing

Preliminary: the spectro-temporal excitation pattern







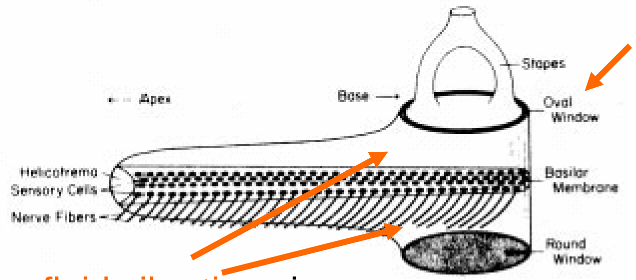
Auditory nerve

peripheral
auditory system

Sound: from air to auditory nerve

1. Sound enters the outer ear as air pressure variations about atmospheric mean …

2. … and is converted to mechanical vibration of the oval window by the ossicles of the middle ear

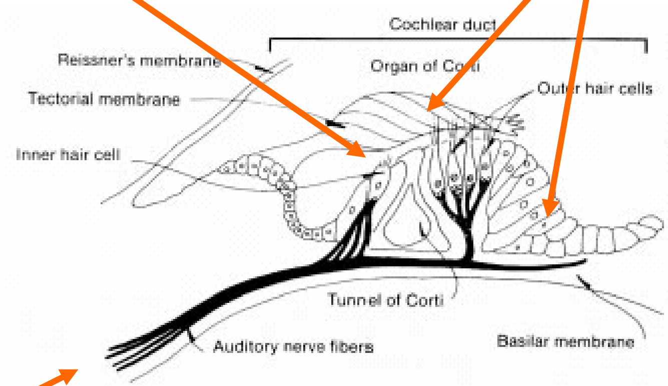


3. … causing fluid vibrations in the incompressible cochlear liquids…

4. … giving rise to shearing movements between the basilar and tectorial membranes …

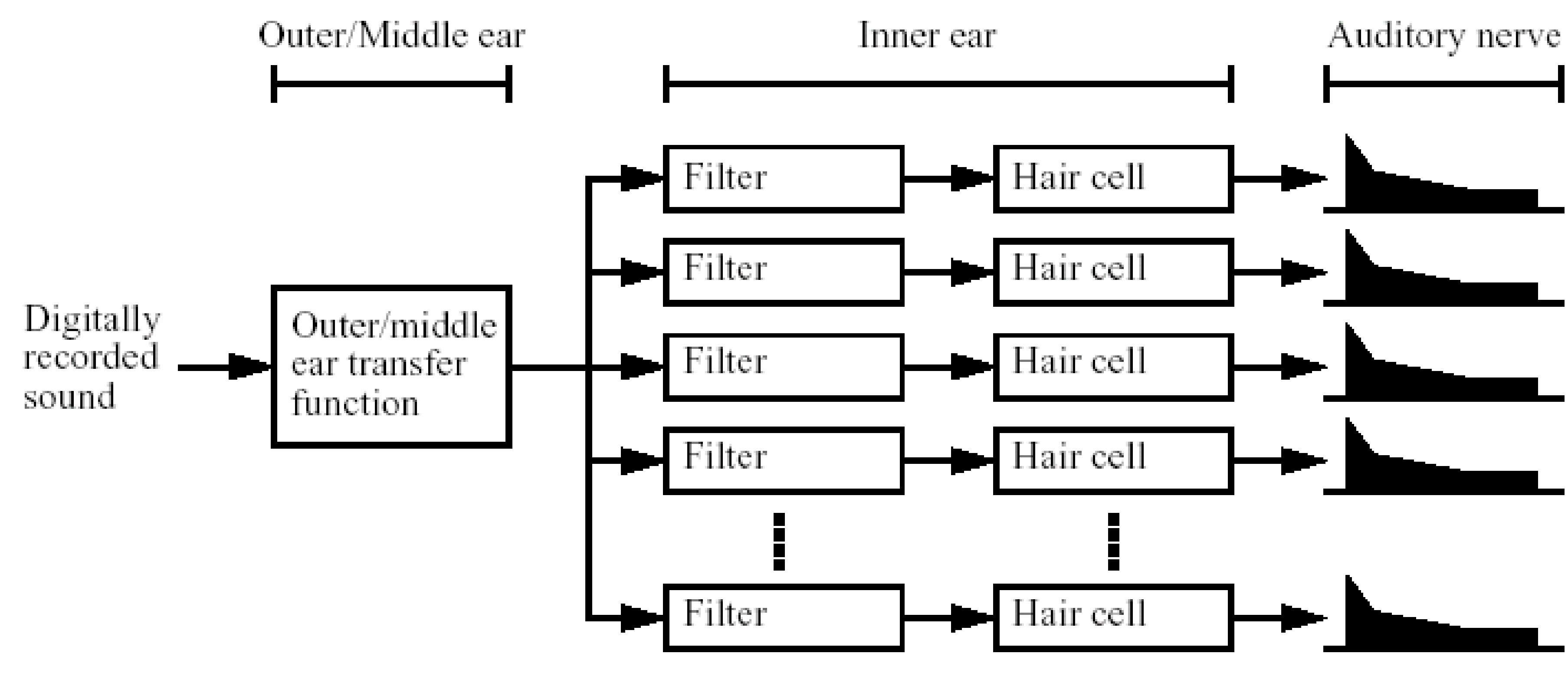
5. … which are detected by mechanical deflections of stereocilia of the inner hair cells …

6. … modulating the release of chemical neurotransmitter …



7. … which builds up and eventually produces an electrical impulse in an auditory nerve fibre

Typical model structure

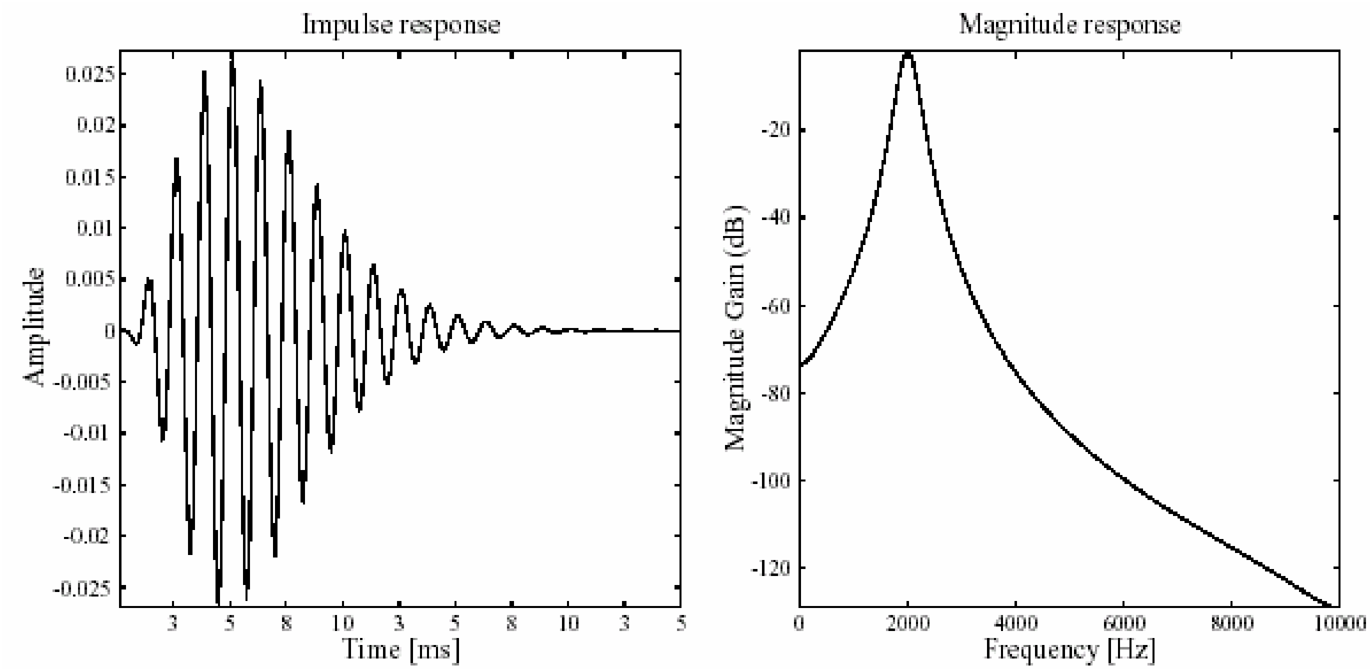


Cochlear filtering model

The *gammatone* function approximates physiologically-recorded impulse responses

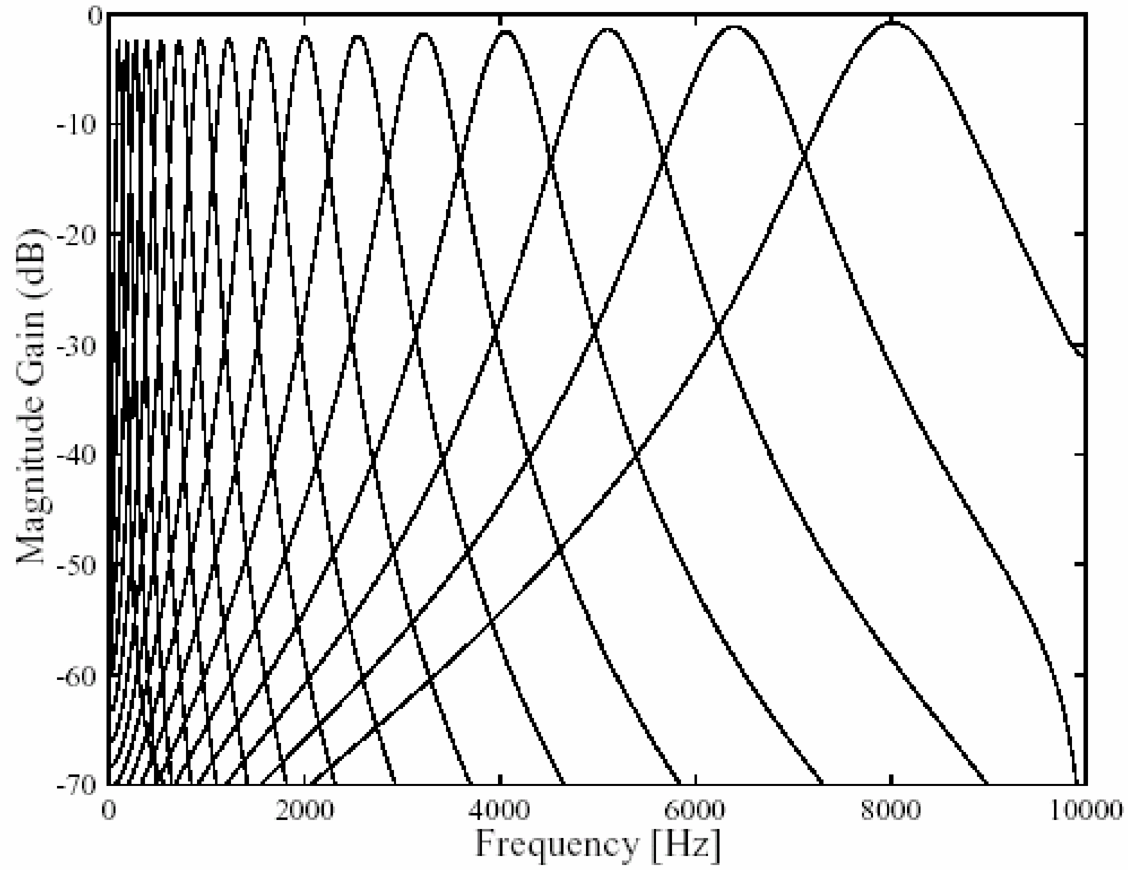
$$g(t) = t^{n-1}\exp(-2\pi bt)\cos(2\pi f_0 t + \phi)$$

- n = filter order (4)
- b = bandwidth
- f_0 = centre frequency
- ϕ = phase



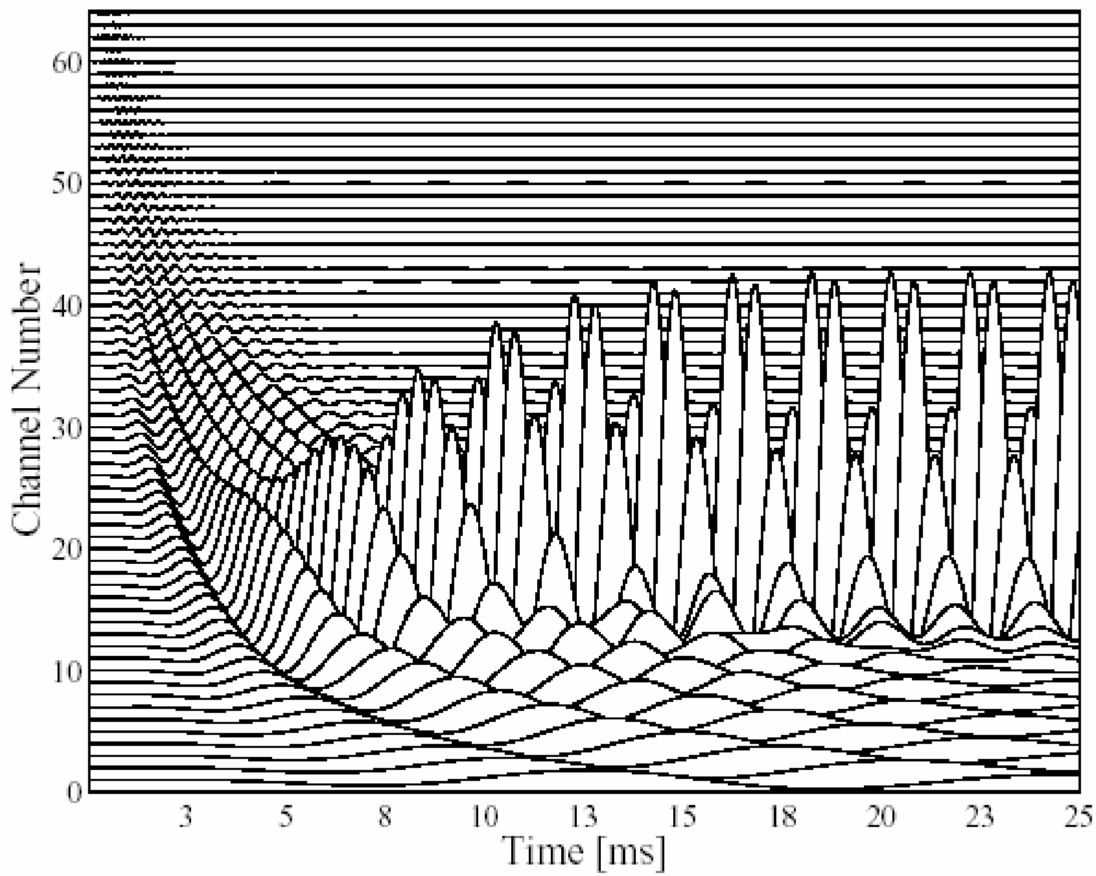
Bank of gammatones

- Each position on the basilar membrane is simulated by a single gammatone filter with appropriate centre frequency and bandwidth
- 32 filters are generally sufficient to cover the range 50-8 kHz
- Note variation in bandwidth with frequency (unlike Fourier analysis)



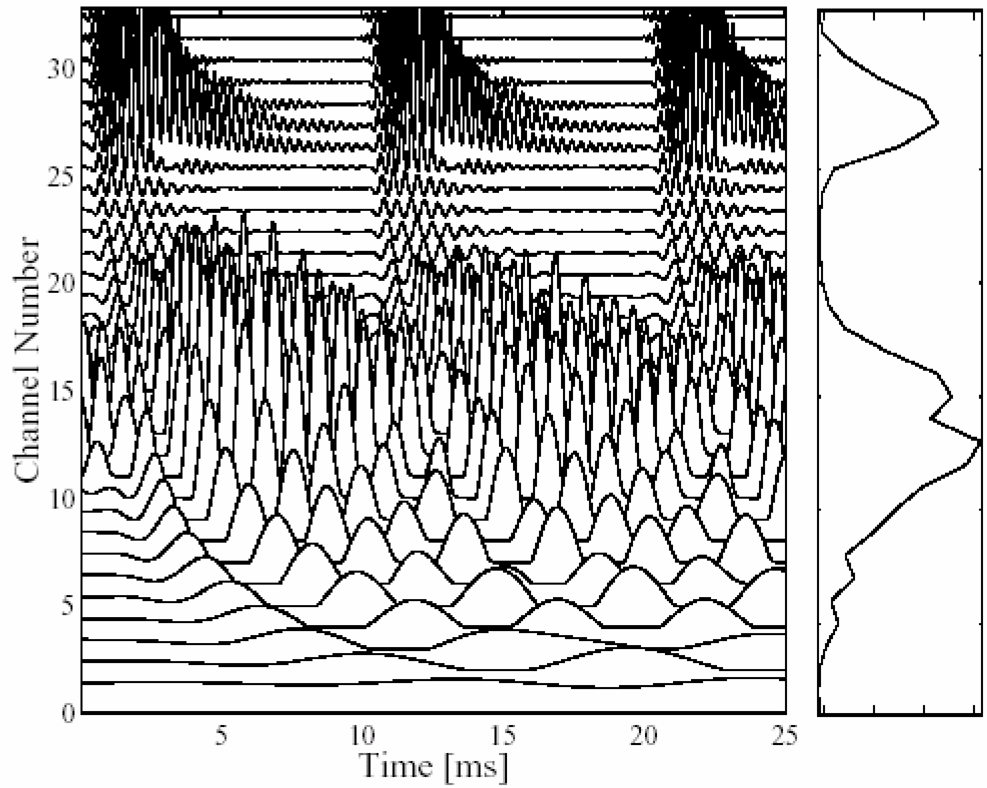
Response to a pure tone

- Many channels respond, but those closest to tone frequency respond most strongly (*place coding*)
- The interval between successive peaks also encodes the tone frequency (*temporal coding*)
- Note propagation delay along the membrane model

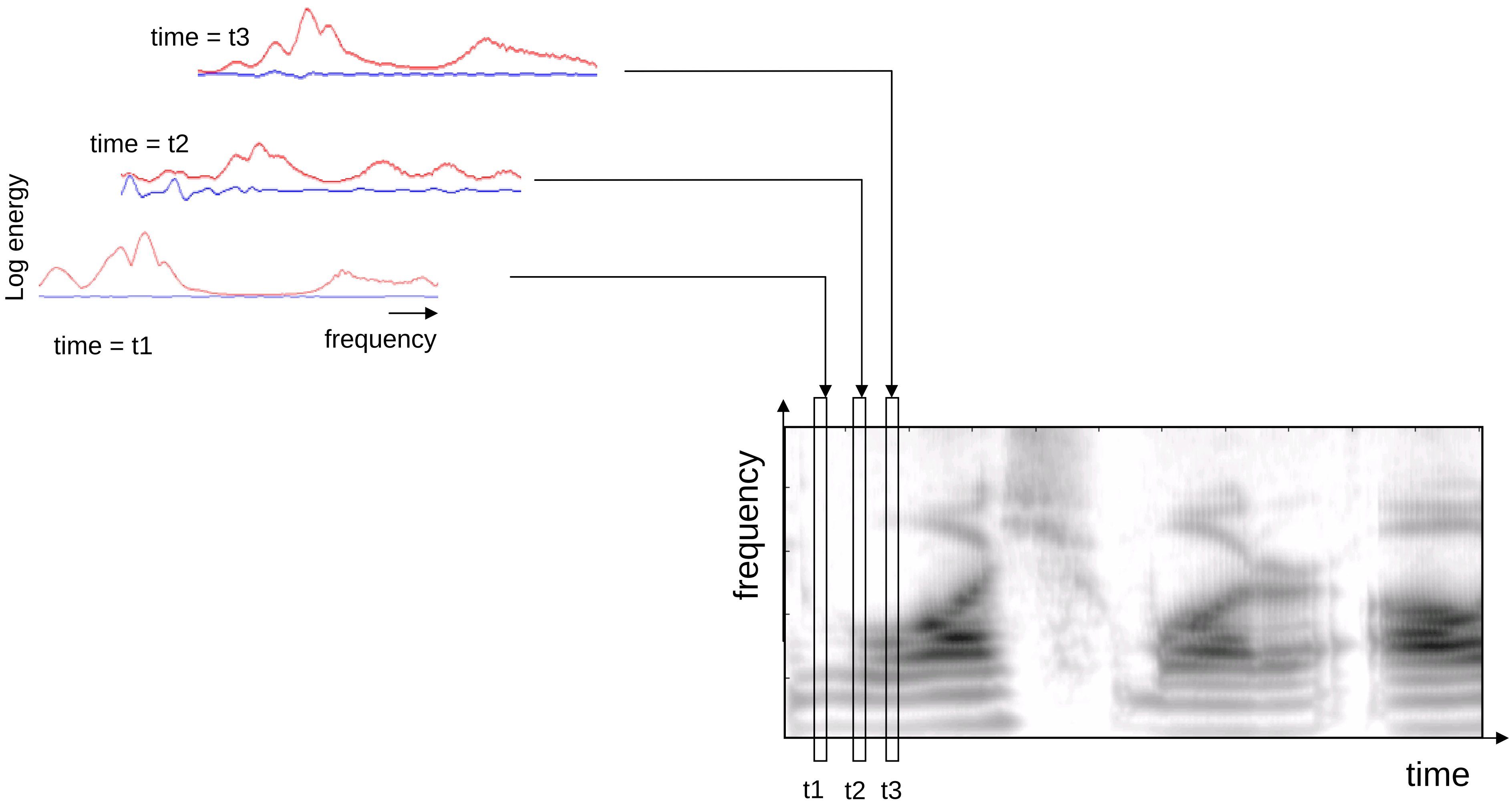


Example output for vowel “ah”

- Phase-locking to individual vowel harmonics at low frequencies, where the filter bandwidth is narrow enough to resolve them
- Amplitude-modulated response at high frequencies, caused by beating (interaction of unresolved harmonics) in the wider auditory filters
- Summed response across time gives an auditory spectrum



Constructing the spectro-temporal excitation pattern



Part III:

Cues for separating sources

Illustration of potential cues in excitation patterns for speech

"… wash water all … "

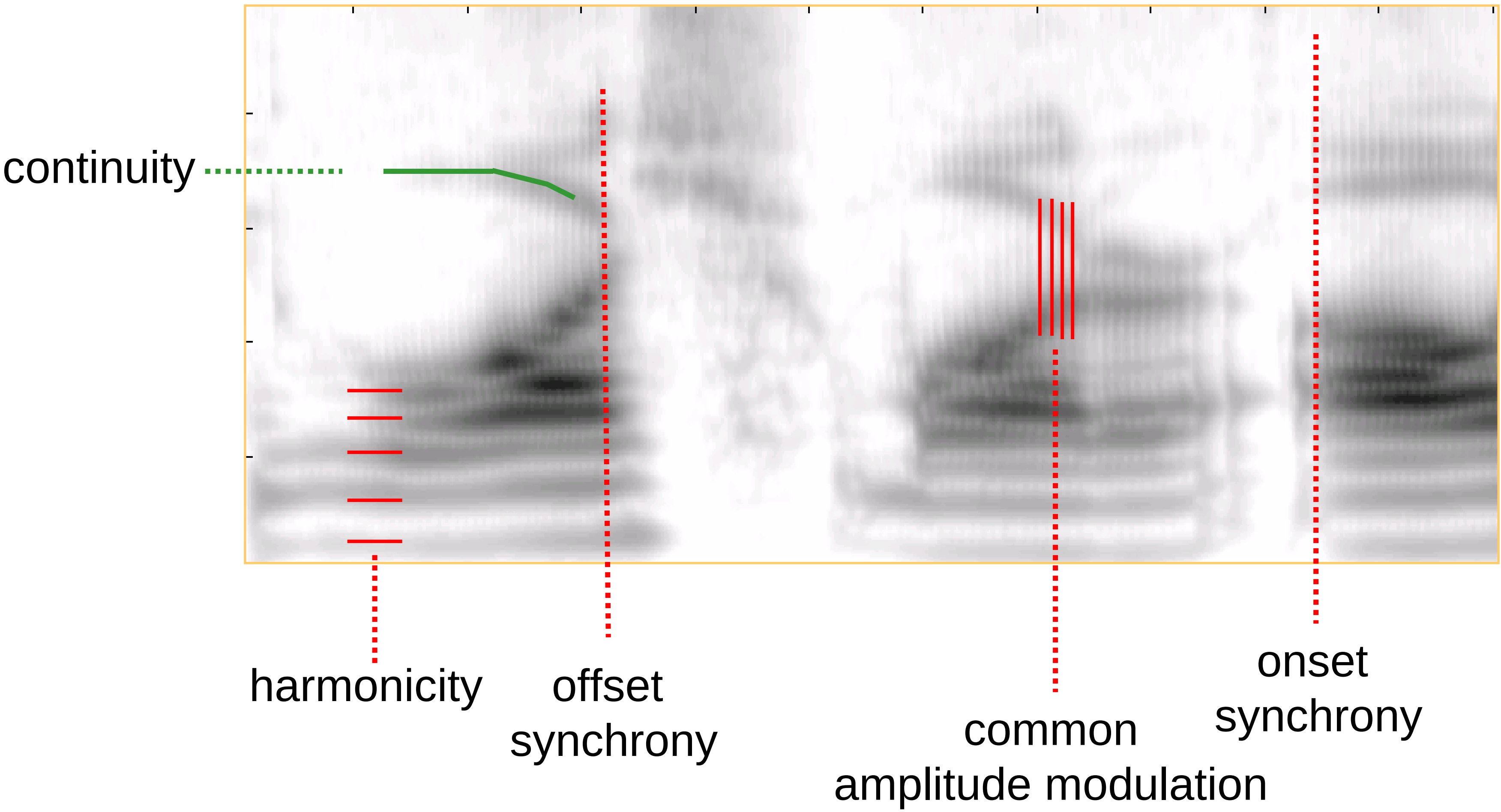
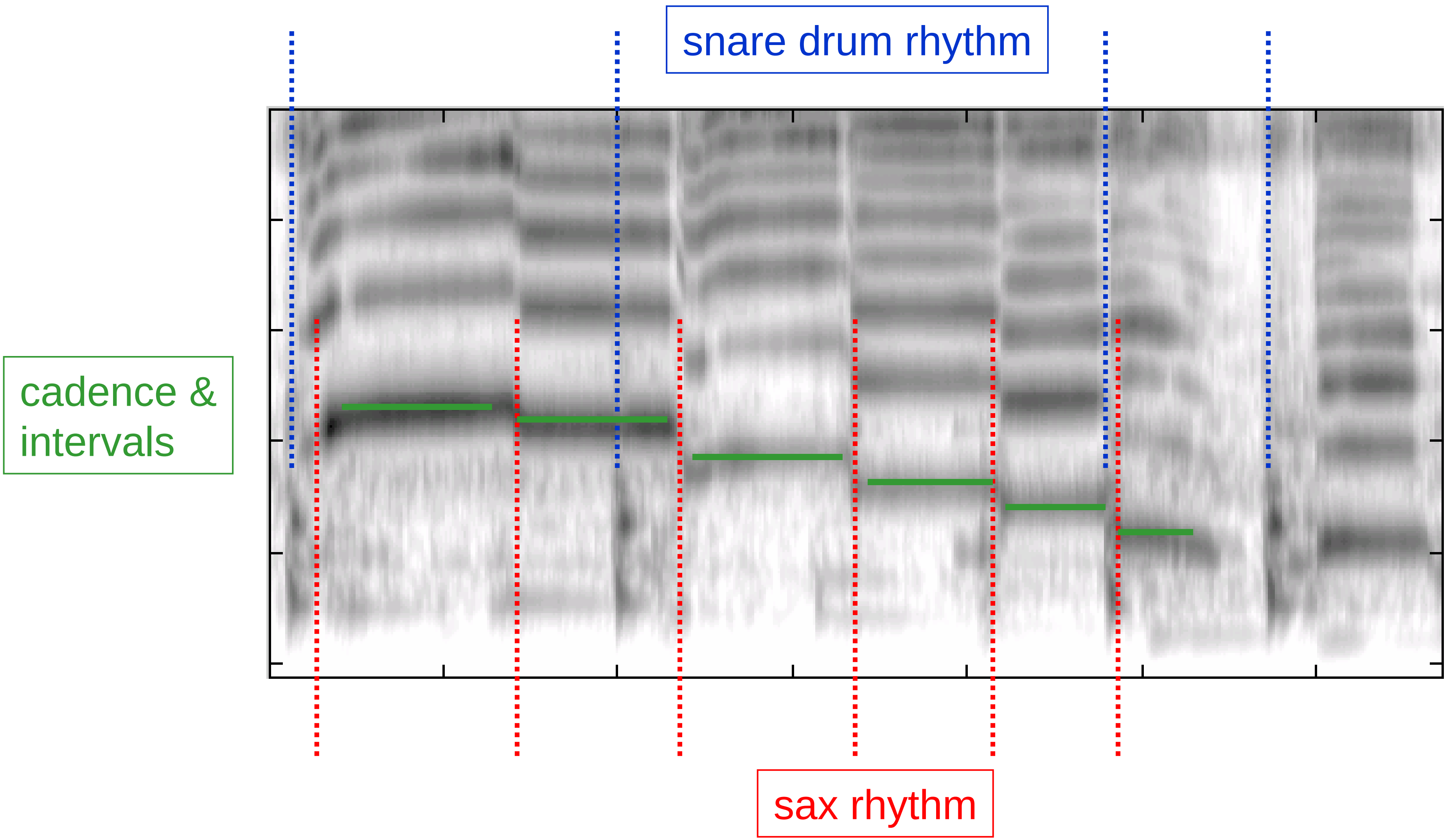
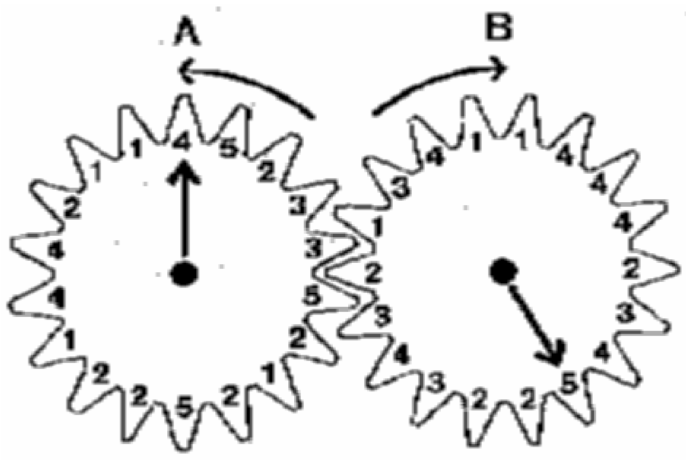


Illustration of potential cues in music



Ugandan xylophone music

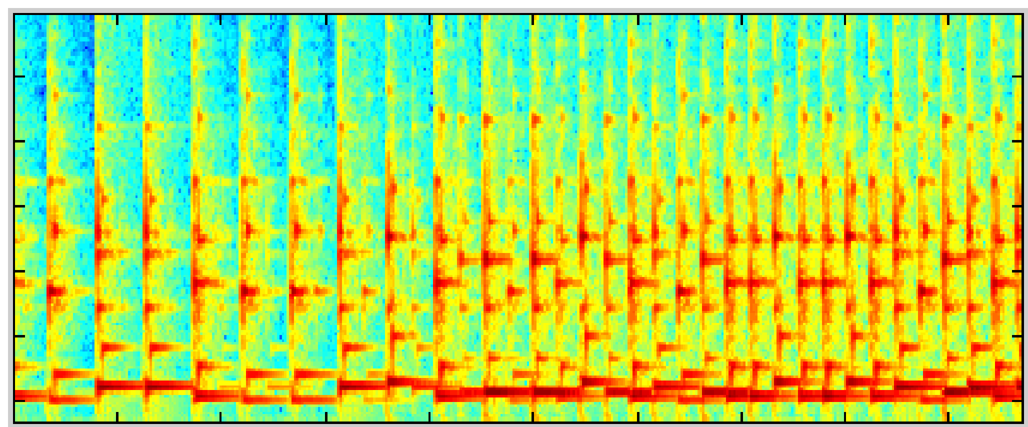
2 players alternate so their notes interleave (pentatonic scale)



- notes occupy different pitch range
- notes occupy same pitch range
- notes have same pitch range but different timbre

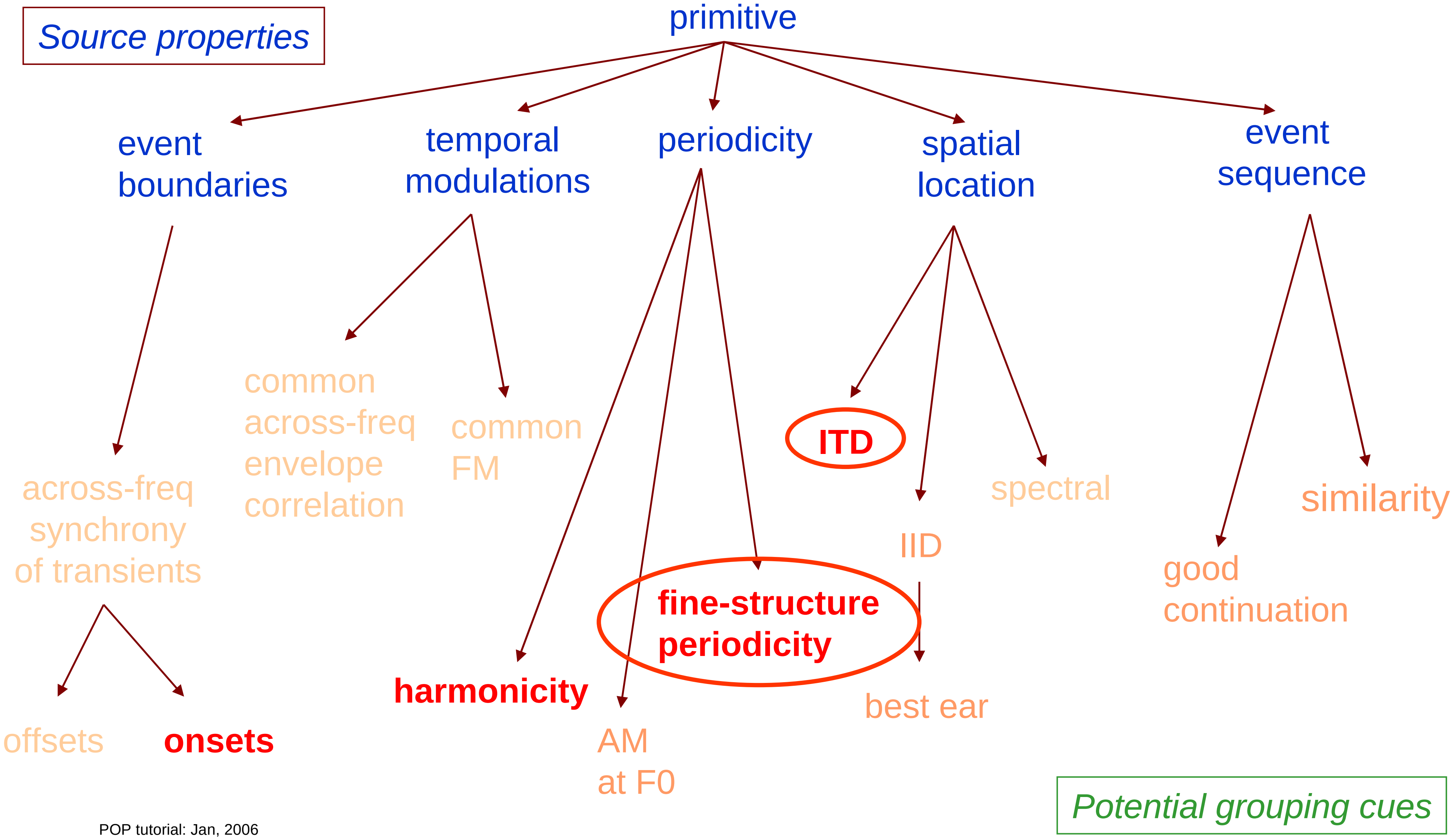
Excitation patterns for same pitch

↓ 2nd player joins in



Source: Bregman & Ahad (1995); original demo by Wegner

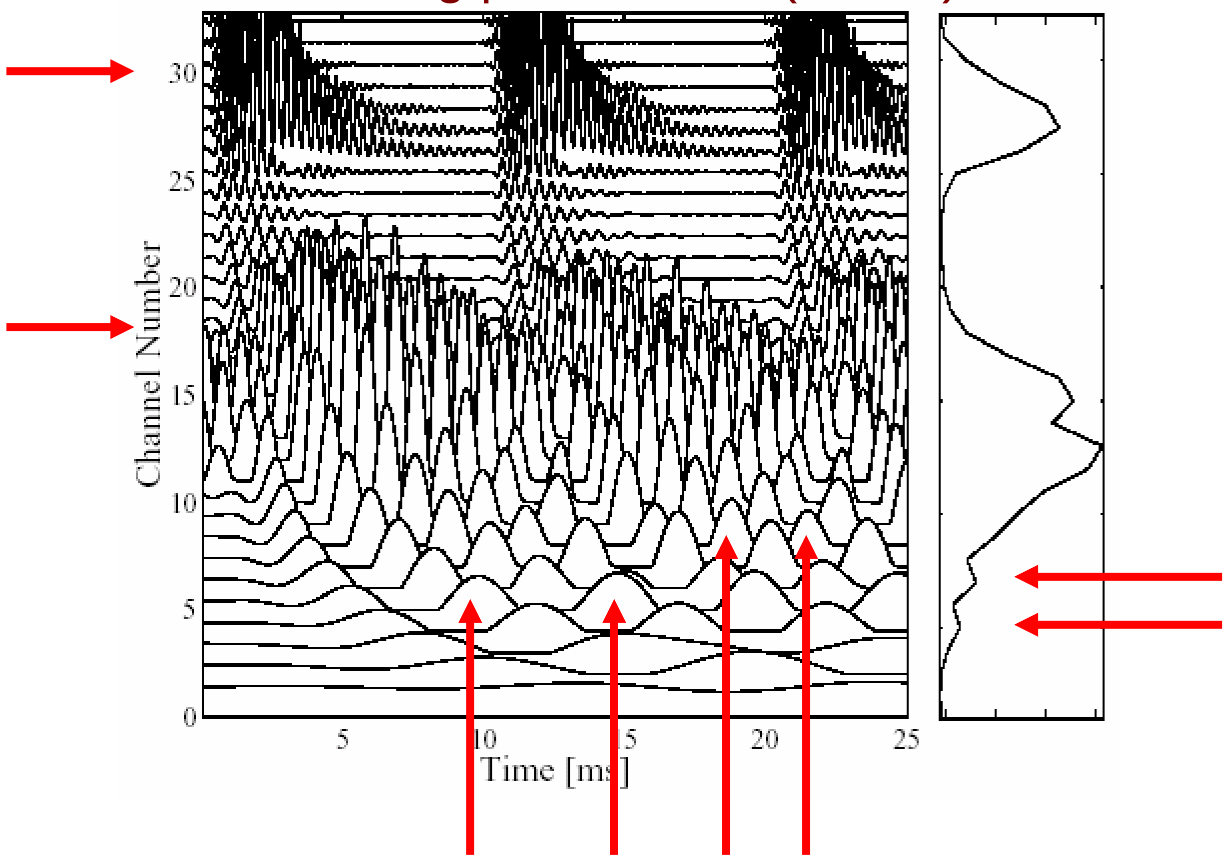
Cues for CASA?



Periodicity cues in auditory nerve

nerve firing probabilities (model)

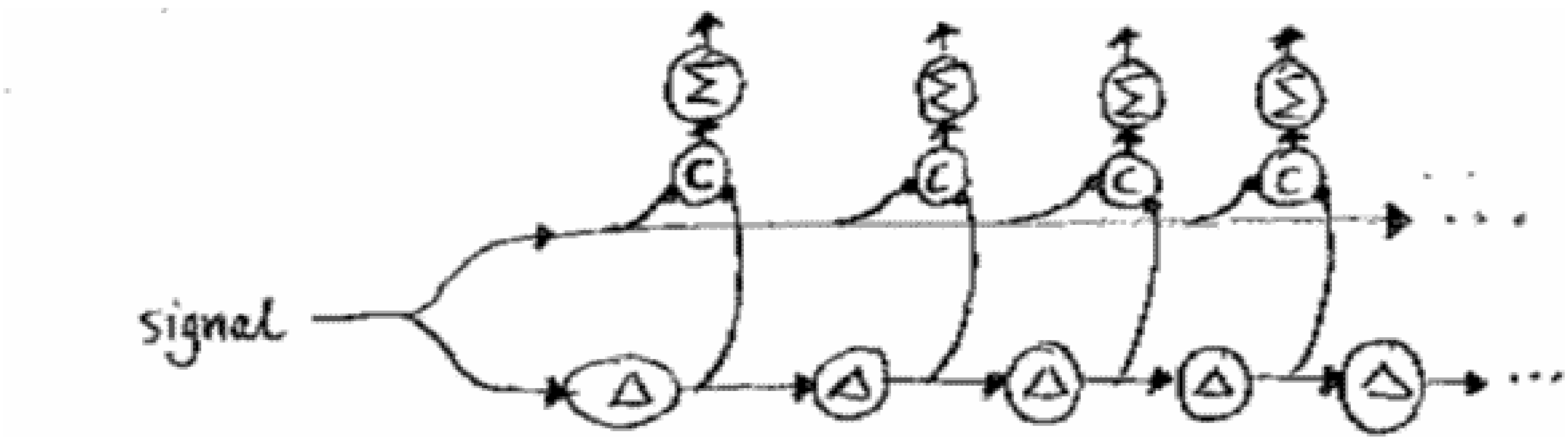
Common amplitude modulation



harmonicity

Fine structure periodicities
(common intervals)

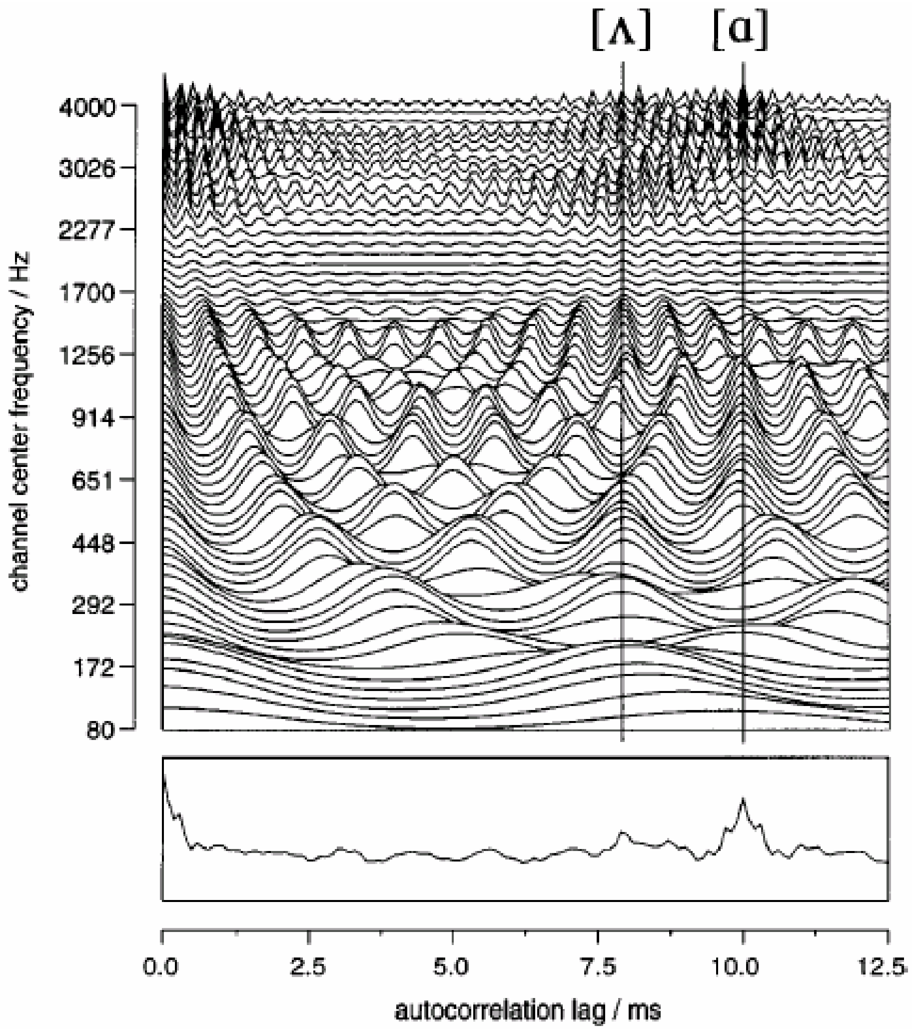
Neural autocoincidence



Licklider (1951)

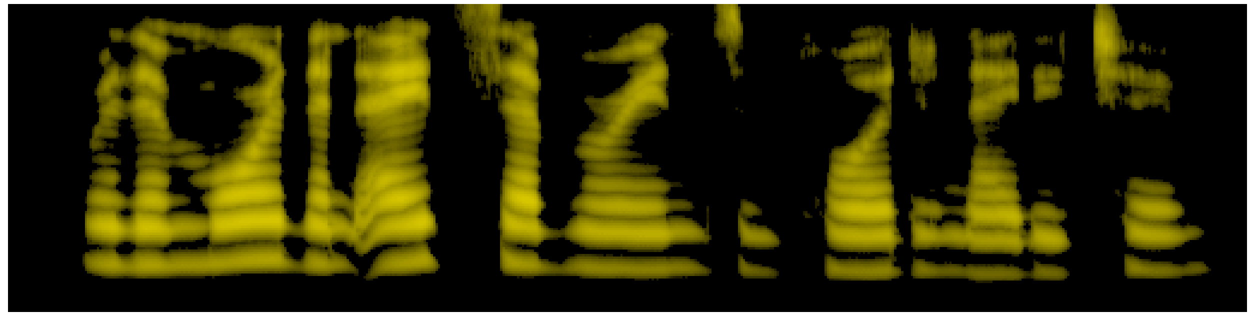
Autocorrelogram

- Short-term autocorrelation of the output of each channel of the auditory periphery model
- Low frequency channels respond to individual harmonics, showing AC peaks at the period of the closest harmonic and at multiples of that period
- A summary autocorrelogram is formed by summing responses across frequency
- Peaks from harmonics of the same F0 show constructive interference



Autocorrelogram (ACG) & summary ACG of a double vowel, showing F0s

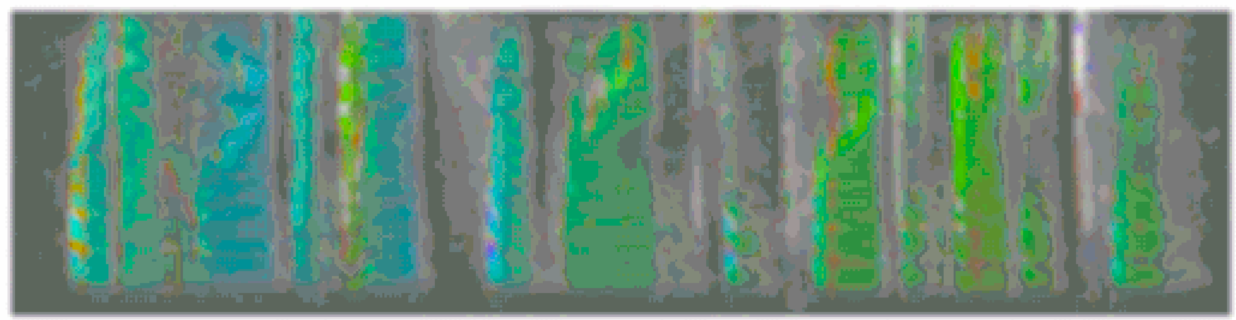
Visualising grouping cues



‘Pitch’ spectrograms

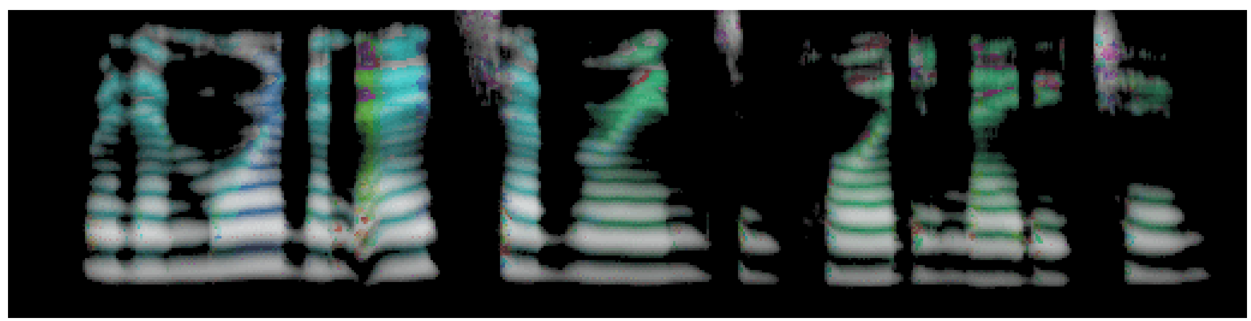
- Energy (value)
- Pitch (hue)
- Pitch strength (saturation)

A



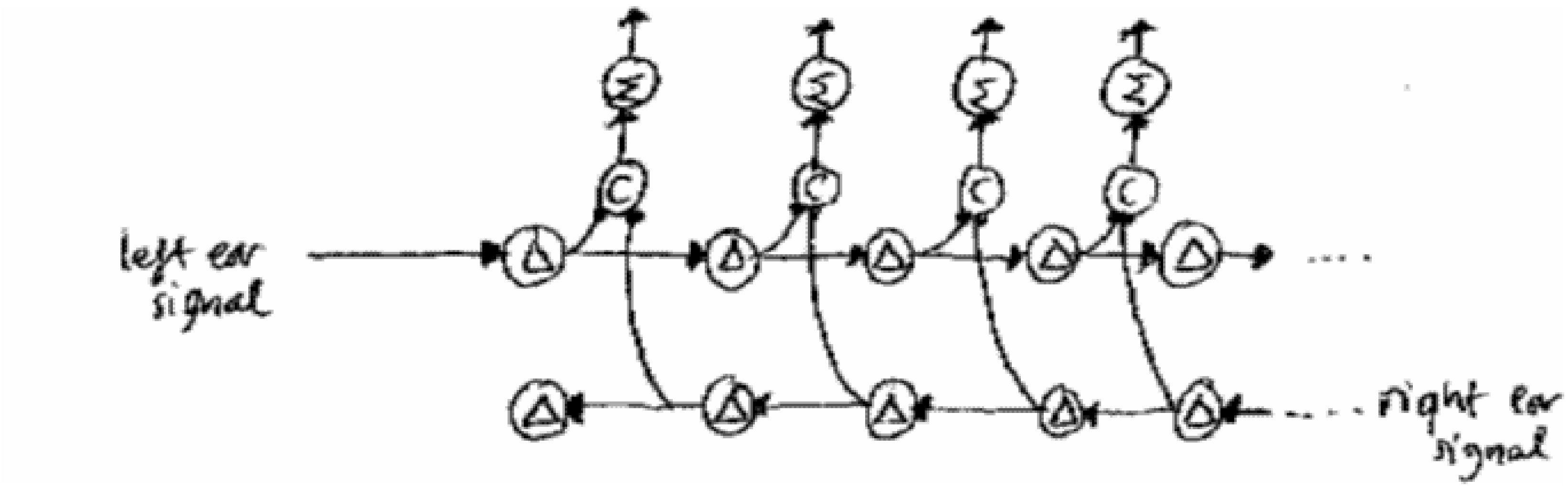
A: pitch & strength from location & height of dominant ACG peak

B



B: … from dominant amplitude modulation freq & depth

Neural cross-coincidence

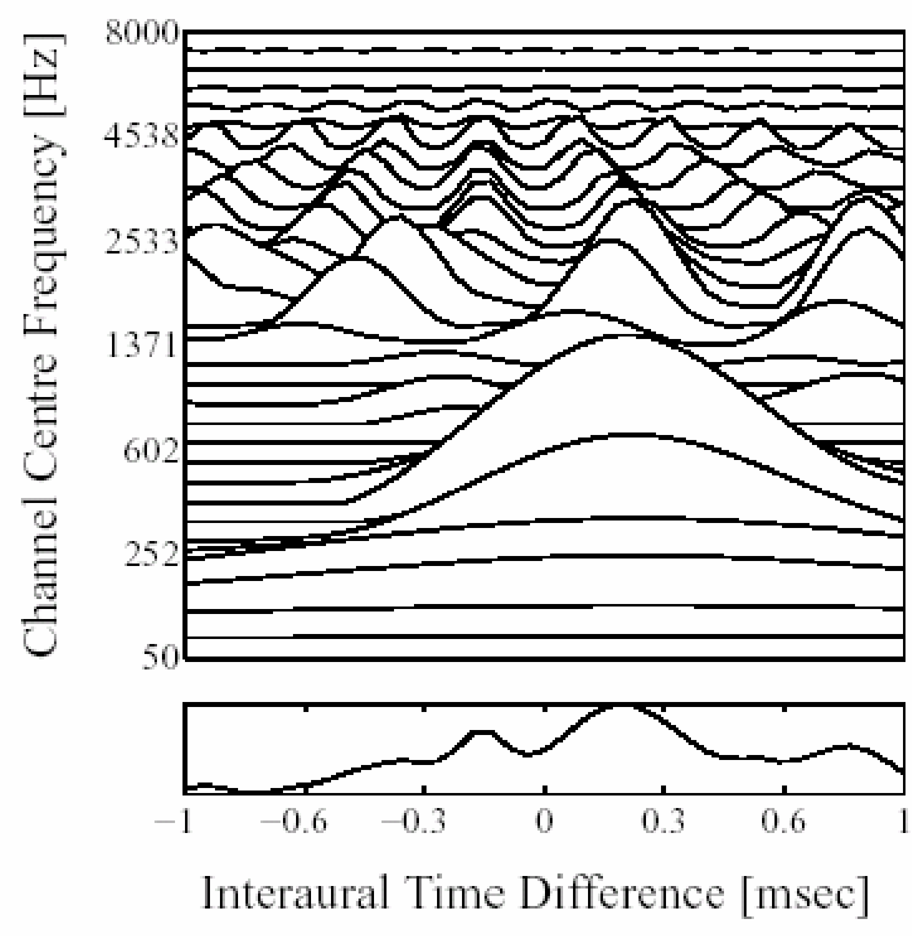


Jeffress (1948)

Cross-correlogram

- One frame (30 ms) of a mixture of spatialised male and female speech, located at -20 and +20 degrees azimuth
- Ideally, a CC should show a spines at delays corresponding to the ITDs of each sound source
- summary CC emphasises such delays, reducing problems due to false peaks

- remaining problem:
  - Multiple peaks at high frequencies where wavelengths are shorter than ear separation; effect is to limit the number of sources localisable to 2 (humans=6)



Source:Palomaki, Brown & Wang (2004)

Contribution of other factors to the perception of everyday speech in masking noise

Technique: how much extra masking (in dBs) can listeners tolerate to reach a criterion level of intelligibility? Each extra dB provides 5-10% increase in intelligibility.

4 dB or more	linguistic entropy (low vs high predictable sentences)
Up to 5 dB	intensity differences
Up to 8 dB	single competing speaker vs many talkers noise
Up to 8 dB	binaural cues

1 dB	location cues in reverberant environments
3 dB	binaural cues to location
4 dB	improved SNR at closest ear

Up to 15 dB	visual cues eg lipreading

Sources: Sumby & Pollack, 1954; Rooij & Plomp, 1991; Brungart, 2001; Festen & Plomp, 1990; Bronkhorst & Plomp, 1992

The case of two radios

Suppose the task is to monitor two radio channels simultaneously, monaurally. At what level should you set the volumes of each channel?

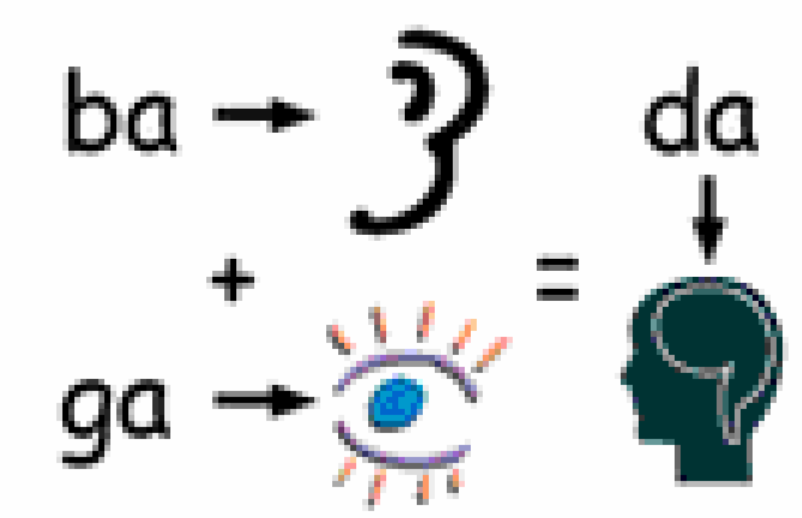
Worst strategy: same volume

Best strategy: around 9 dB apart (improvement of 5 dB SRT; Brungart, 2001)

Vision influences what we hear …

McGurk effect

[video demo]



Source: McGurk & McDonald (1976) "Hearing lips and seeing voices", *Nature*

… and sound influences what we see

Sound-induced visual rabbit

[DEMO]

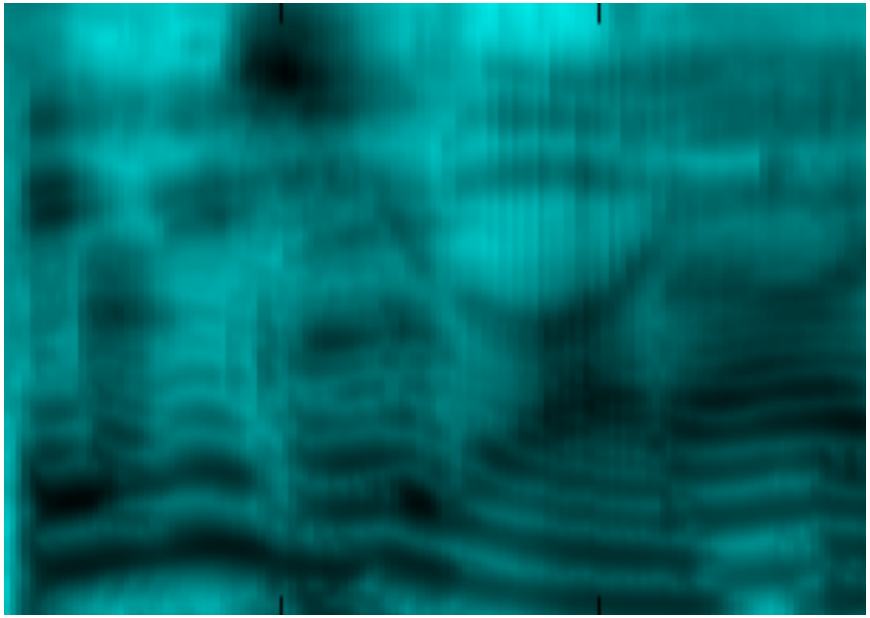
Source: http://neuro.caltech.edu/~kamitani/audiovisualRabbit/
Kamitani, Y & Shimojo, S (2001)

Part IV:

Summary of computational approaches to source separation

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

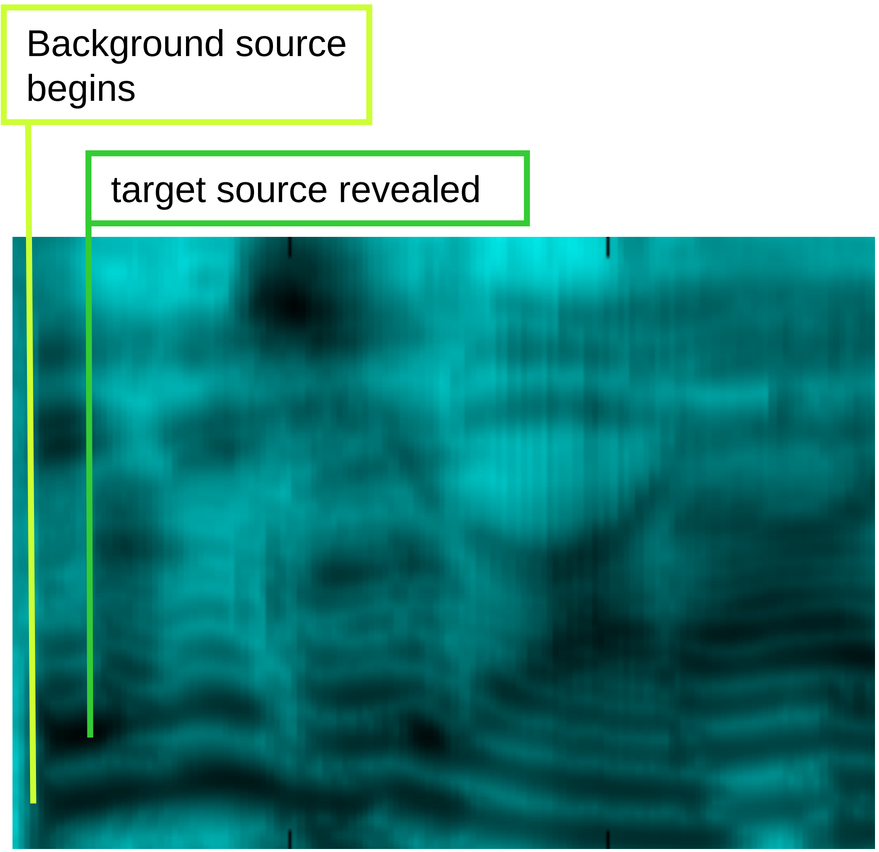
- How to combine cues
- Grouping is not all-or-nothing
- Different thresholds for different tasks (Darwin)
- No really successful model of sequential grouping

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

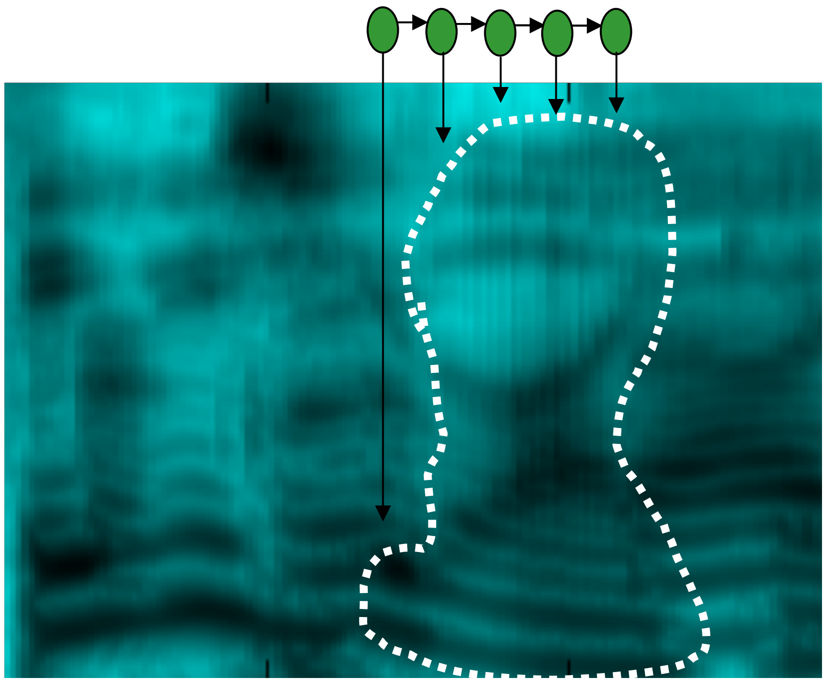
Models

(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)

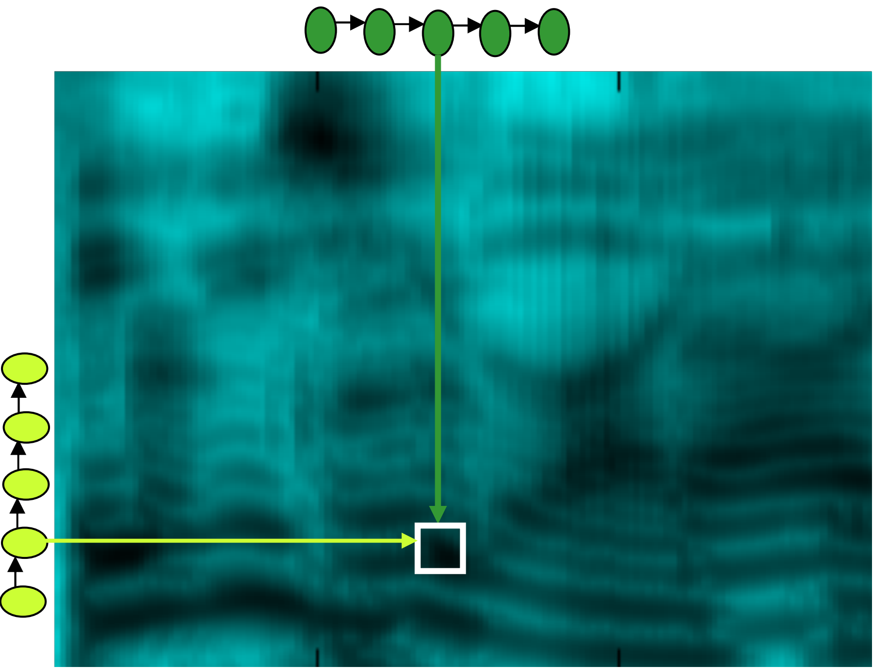
IV: Hard-core model-based explanation

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

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

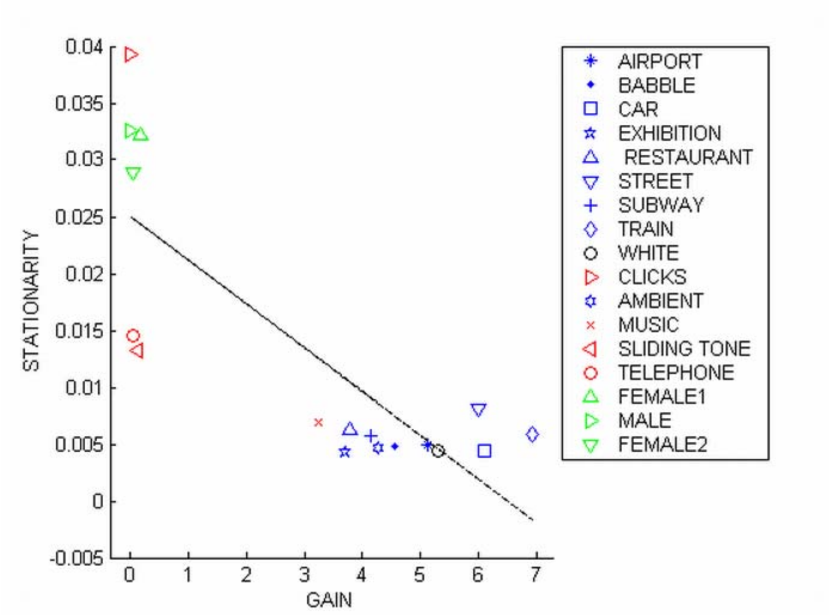
- O Organisational cues in target speech
- O Organisational cues in background
- O Models for target speech
- O Models for background
- O Energetic masking
- O Informational masking
- O **Stationarity of background**

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

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

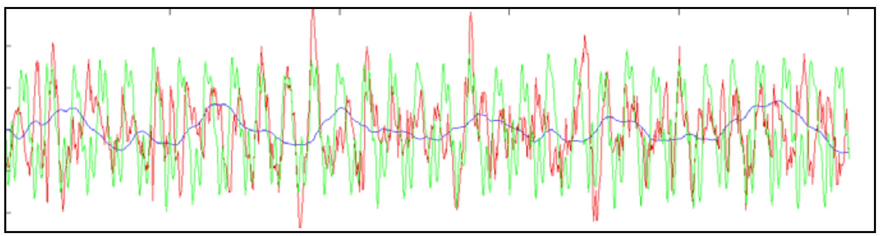
- detection and integration of sparse information in speech

The glimpsing hypothesis: listeners separate sources by exploiting brief regions where the target source is dominant

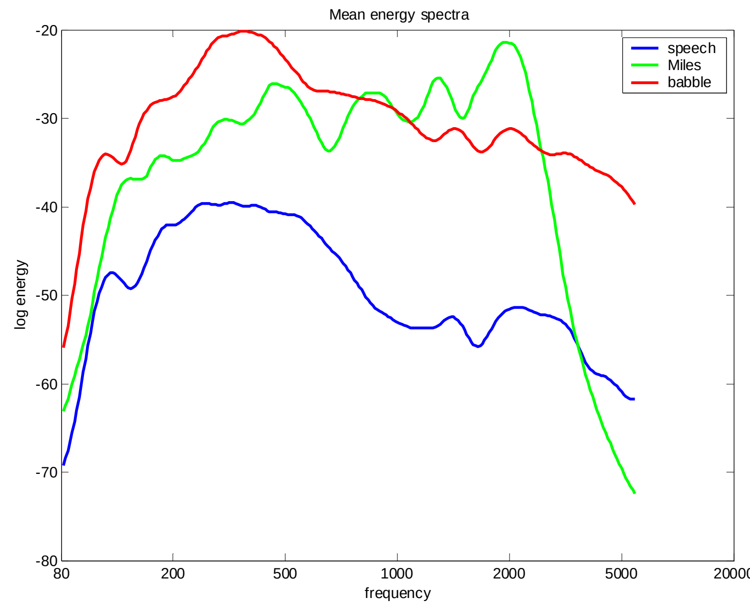
Cooke (2006) A *glimpsing model of speech perception in noise*, Journal of the Acoustical Society of America, to appear in March

Recall the overlap problem

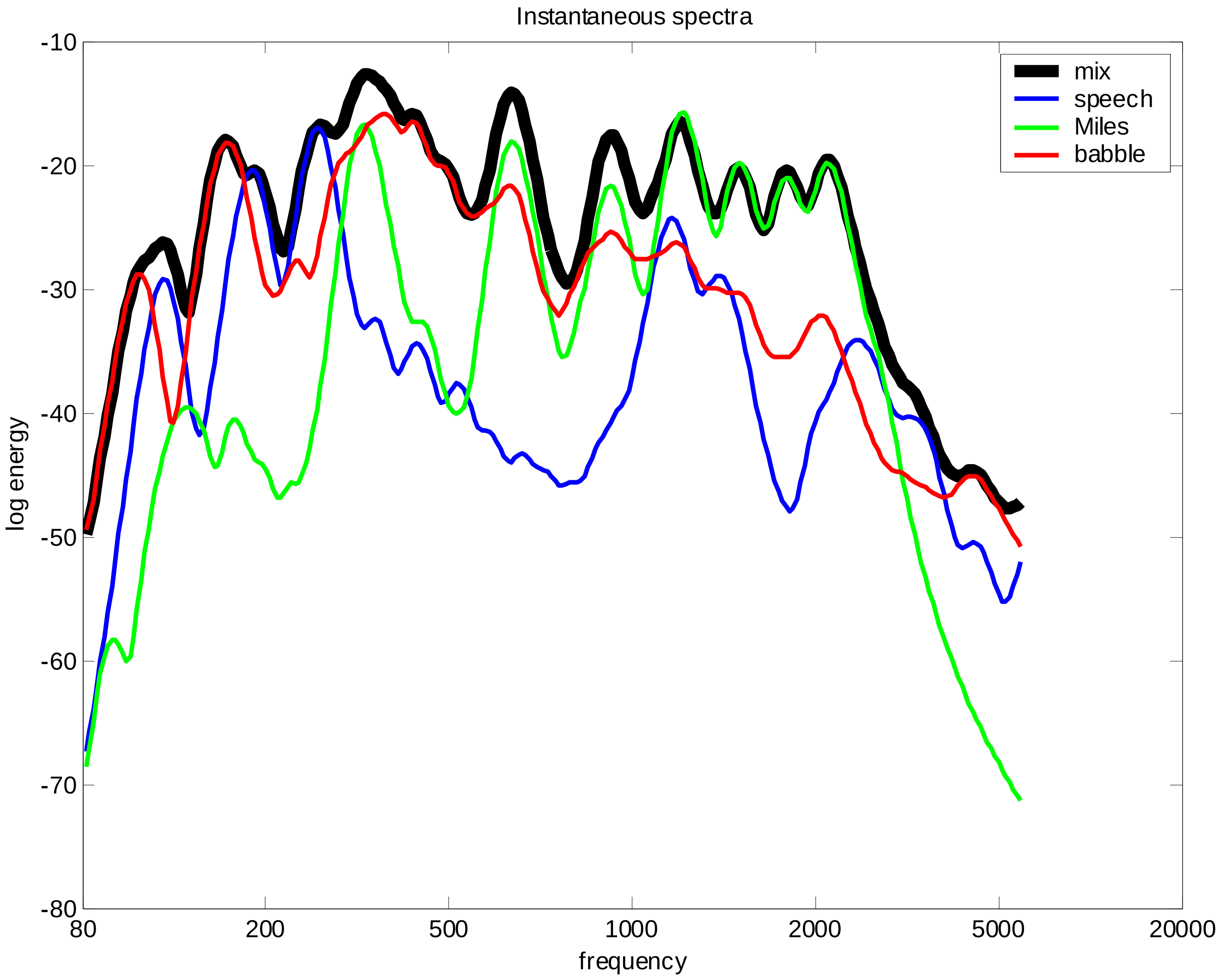
Time domain



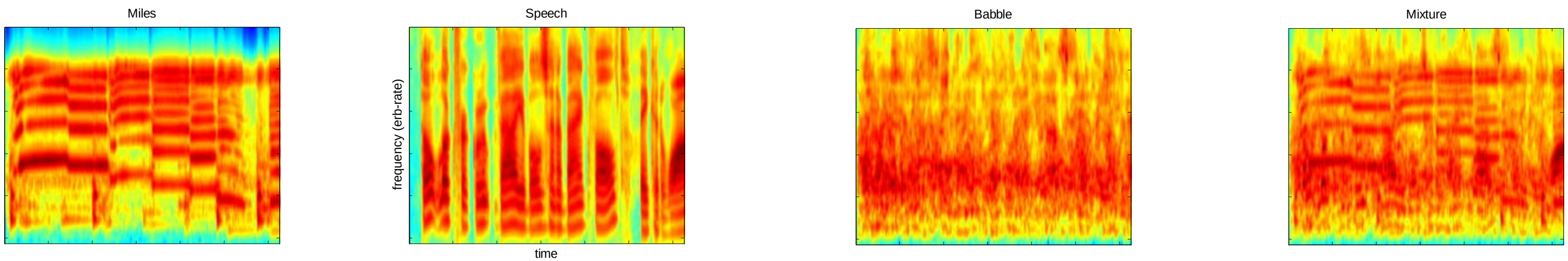
Frequency domain



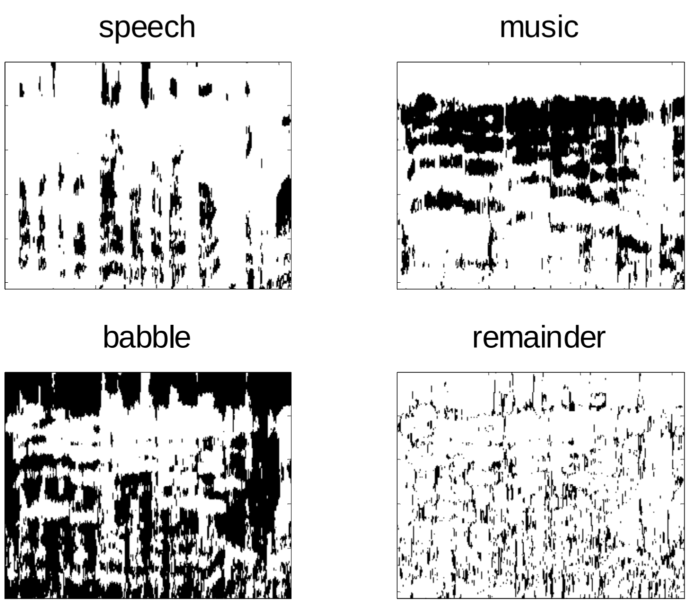
A single frequency slice



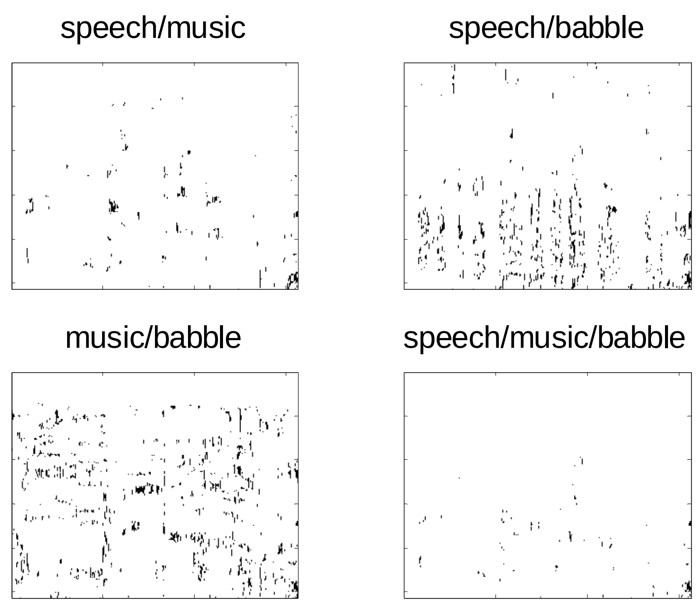
Sparse information in mixtures



Energy within 3 dB of value in mix



Energy within 3 dB of other source



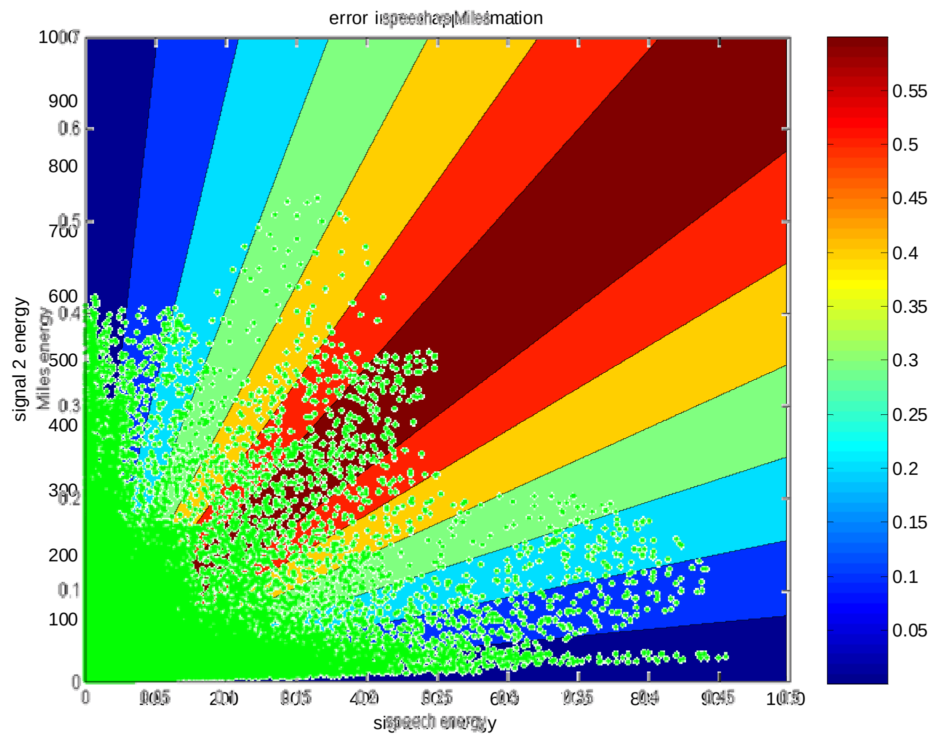
An explanation of the 'dominance effect'

1. As pointed out many times (eg Varga & Moore, 1991), the energy in dBs of a mixture is nearly equal to the dB energy of the most intense source in the mixture.

   log (x+y) ≅ max(log(x),log(y))

   This approximation is at its *worst* when the constituents are equally intense.

2. Two or more modulated sources rarely inject similar energies in the same frequency region at the same time



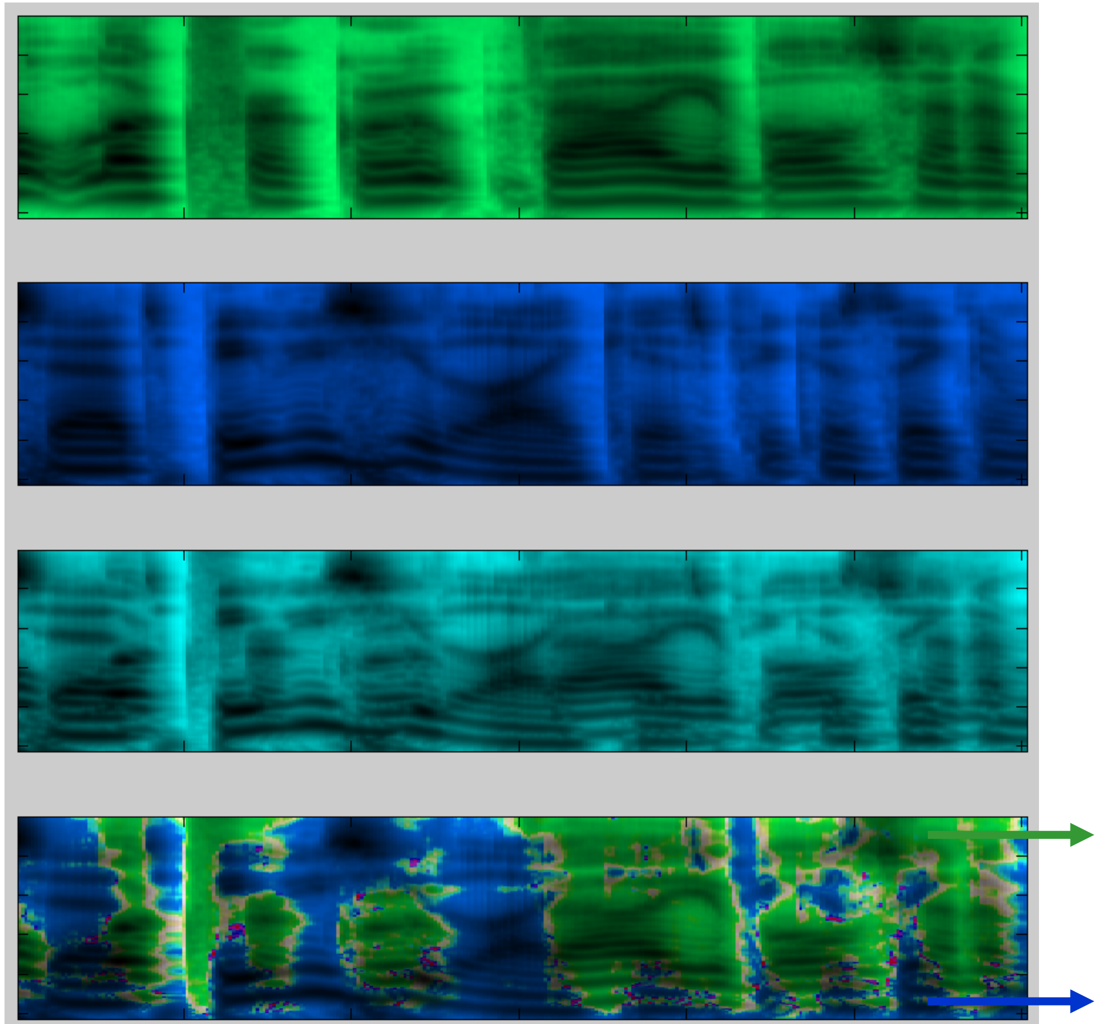
Listening to sparse information

Talker 1

Talker 2

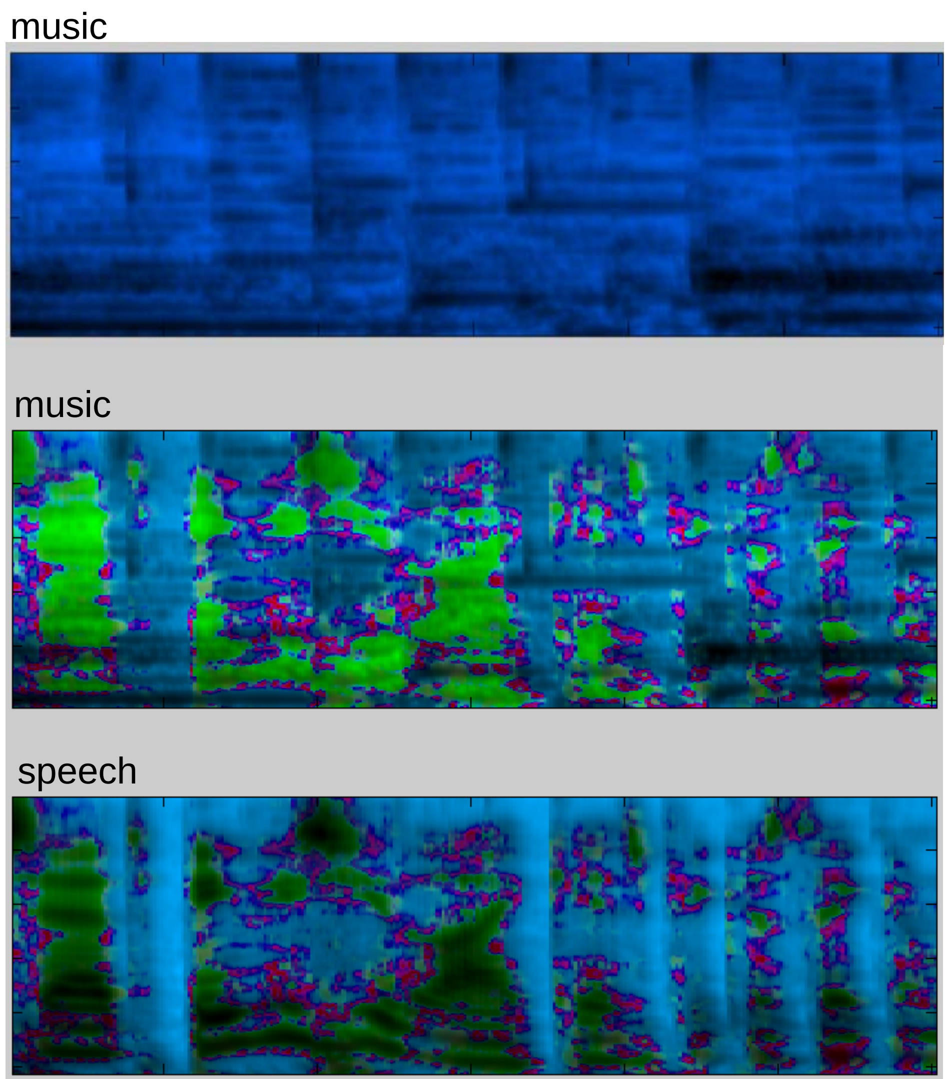
Mix@0dB

One or other
talker dominant



With added
noise

Sparse-sampling of music



Green = speech
shaped regions

Summary of possible ingredients for computational source separation

O Organisational cues in target source	**Auditory scene analysis**
O Organisational cues in background	
O Prior models for target	**Model-based separation**
O Models for background	
O Energetic masking	
O Informational masking	
O Stationarity of background	**Signal processing/robust ASR**
O Source independence	**Statistics, information theory, machine learning**
O Sparsity and redundancy	