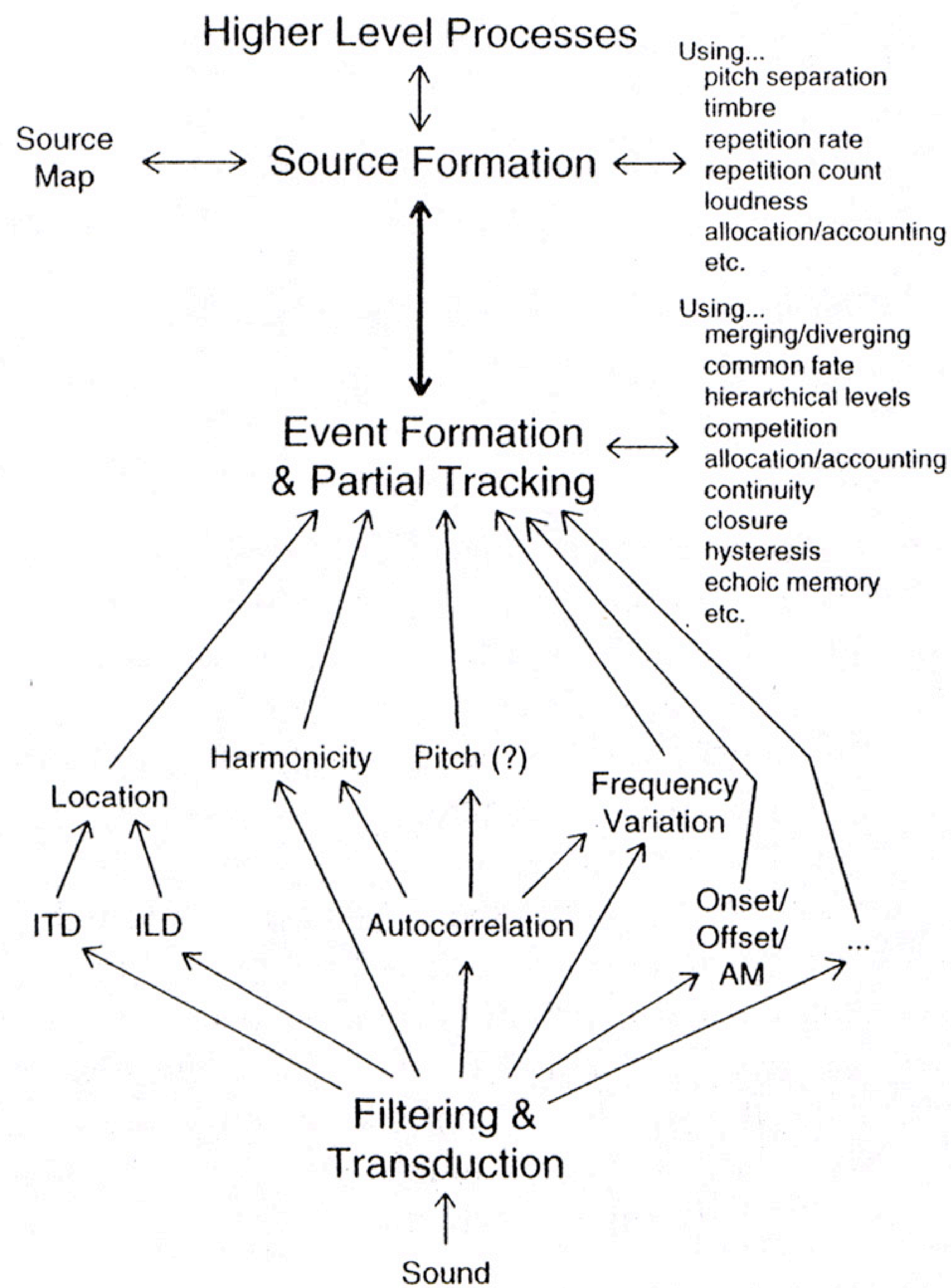
Computational Perception

15-485/785

Auditory Scene Analysis

A framework for auditory scene analysis

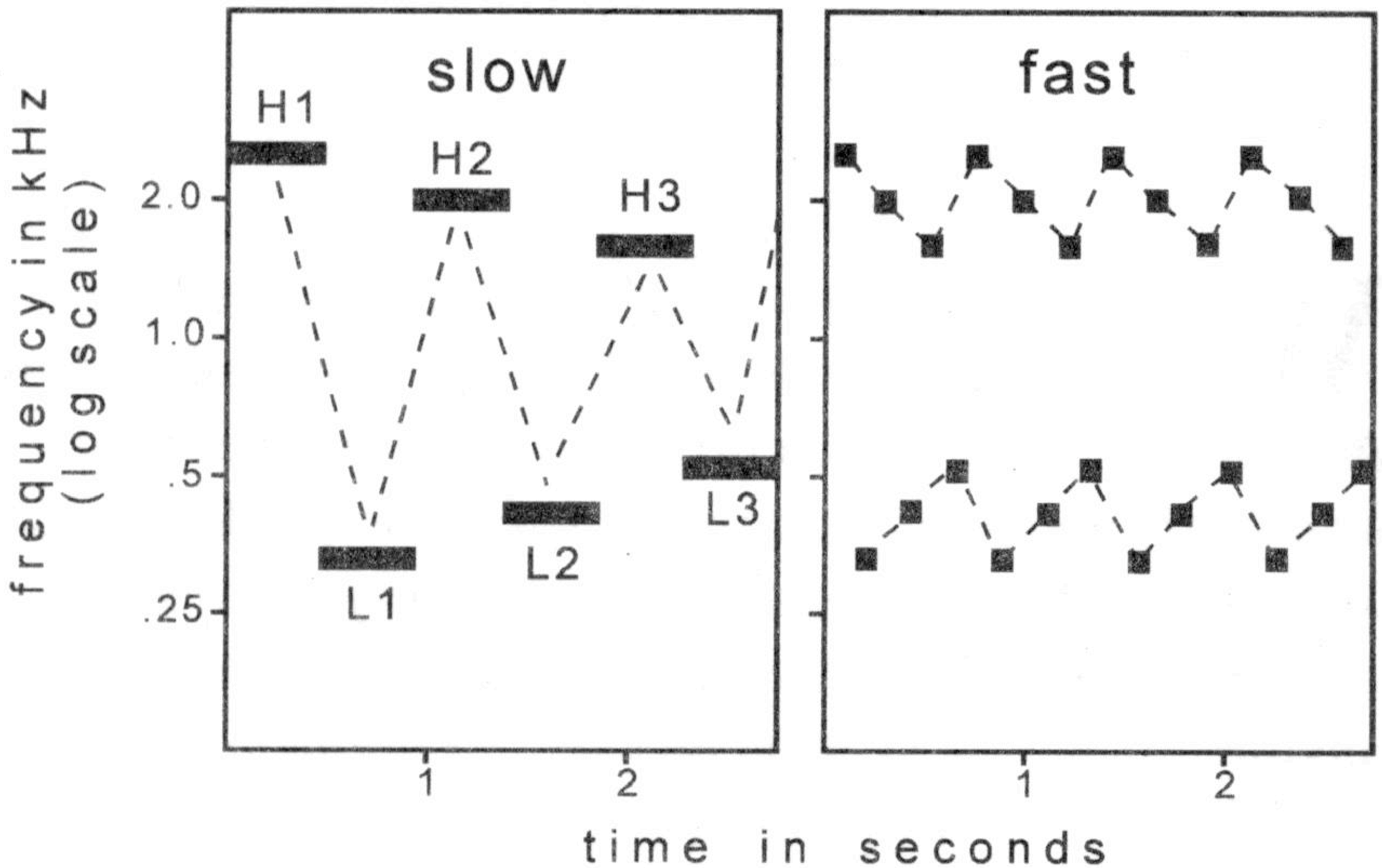


- Auditory scene analysis involves low and high level cues
- Low level acoustic cues are often result in spontaneous grouping
- High level cues can be under attentional control

Cues for auditory grouping

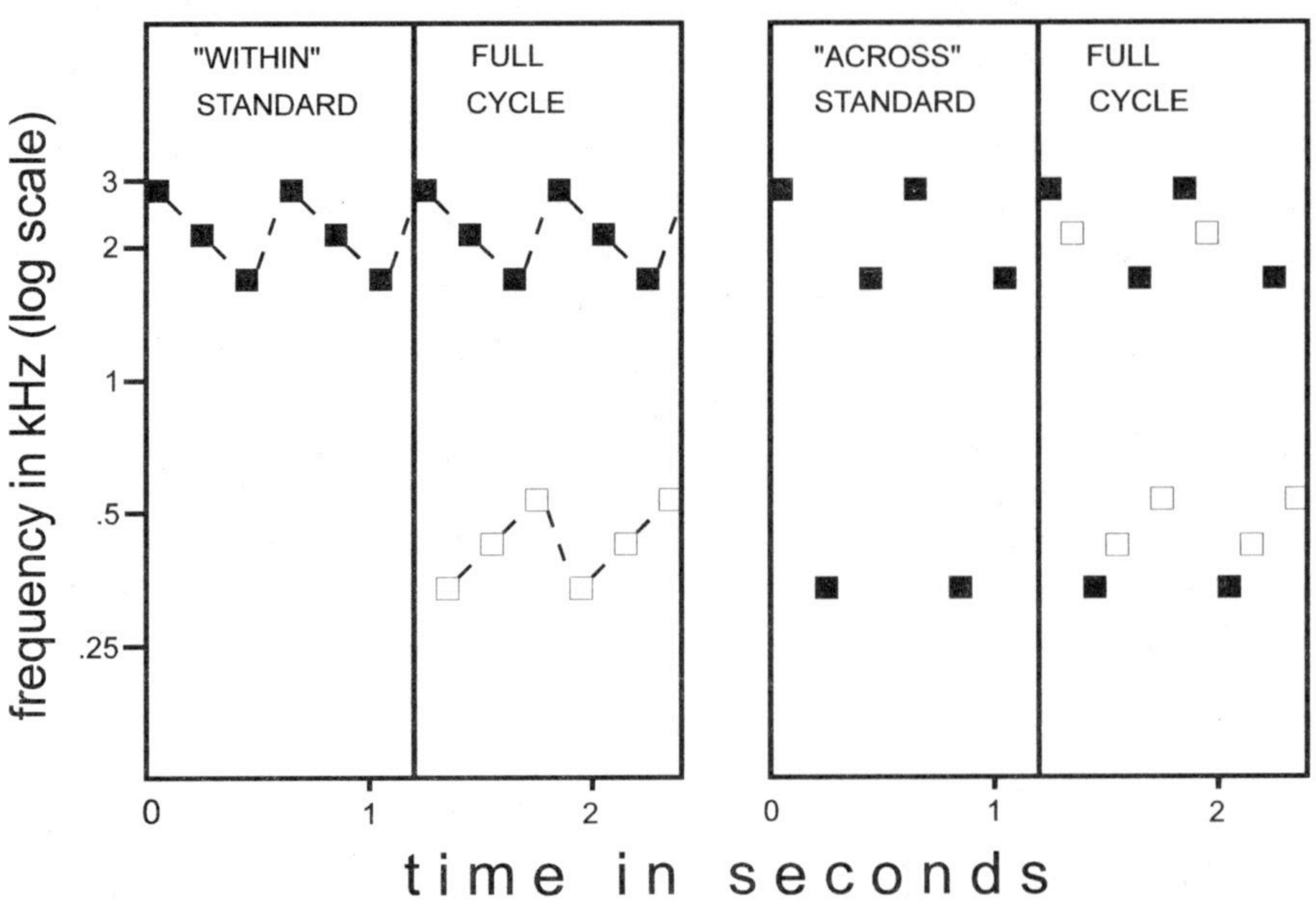
- temporal separation
- spectral separation
- harmonicity
- timbre
- temporal onsets and offsets
- temporal/spectral modulations
- spatial separation

Frequency and tempo cues for steam segregation



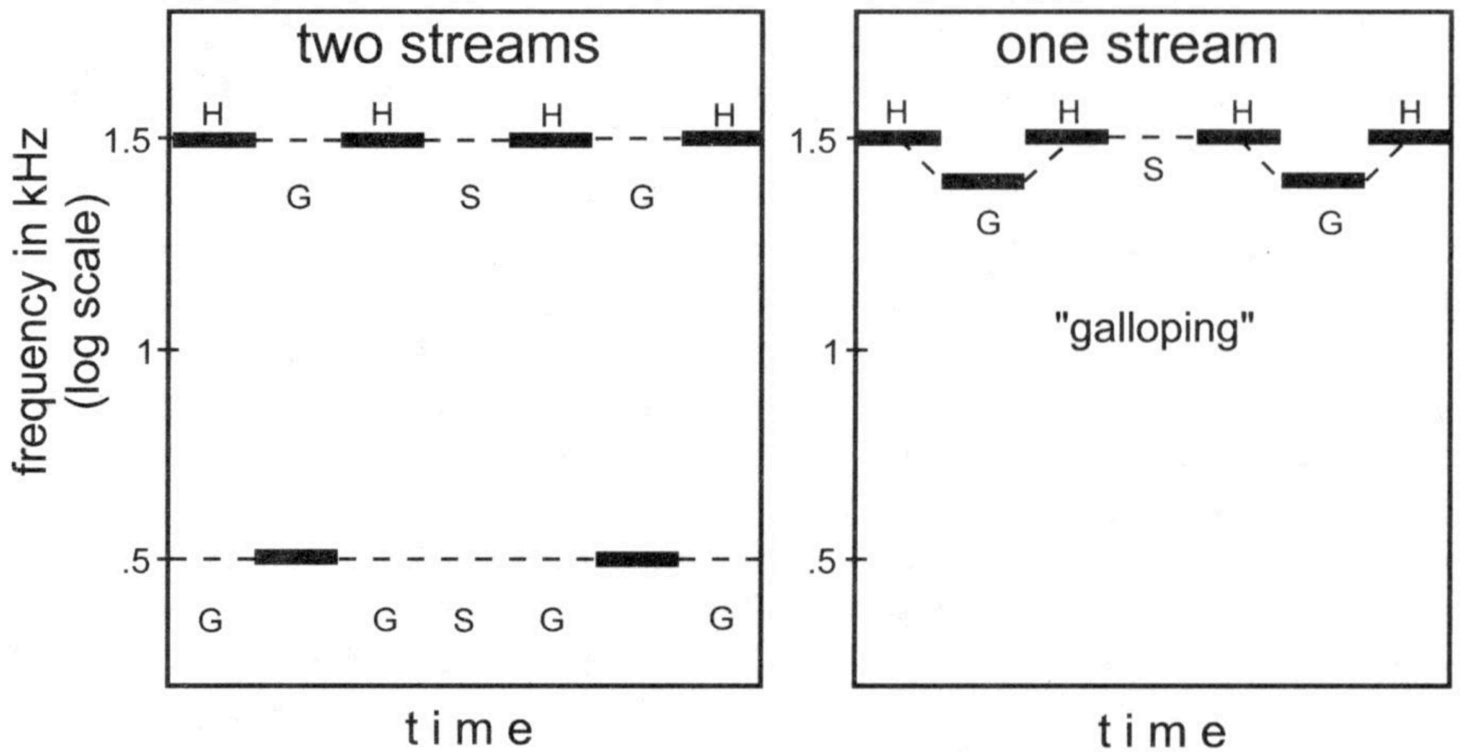
Bregman demo 1.

Perceptual streams can be masked



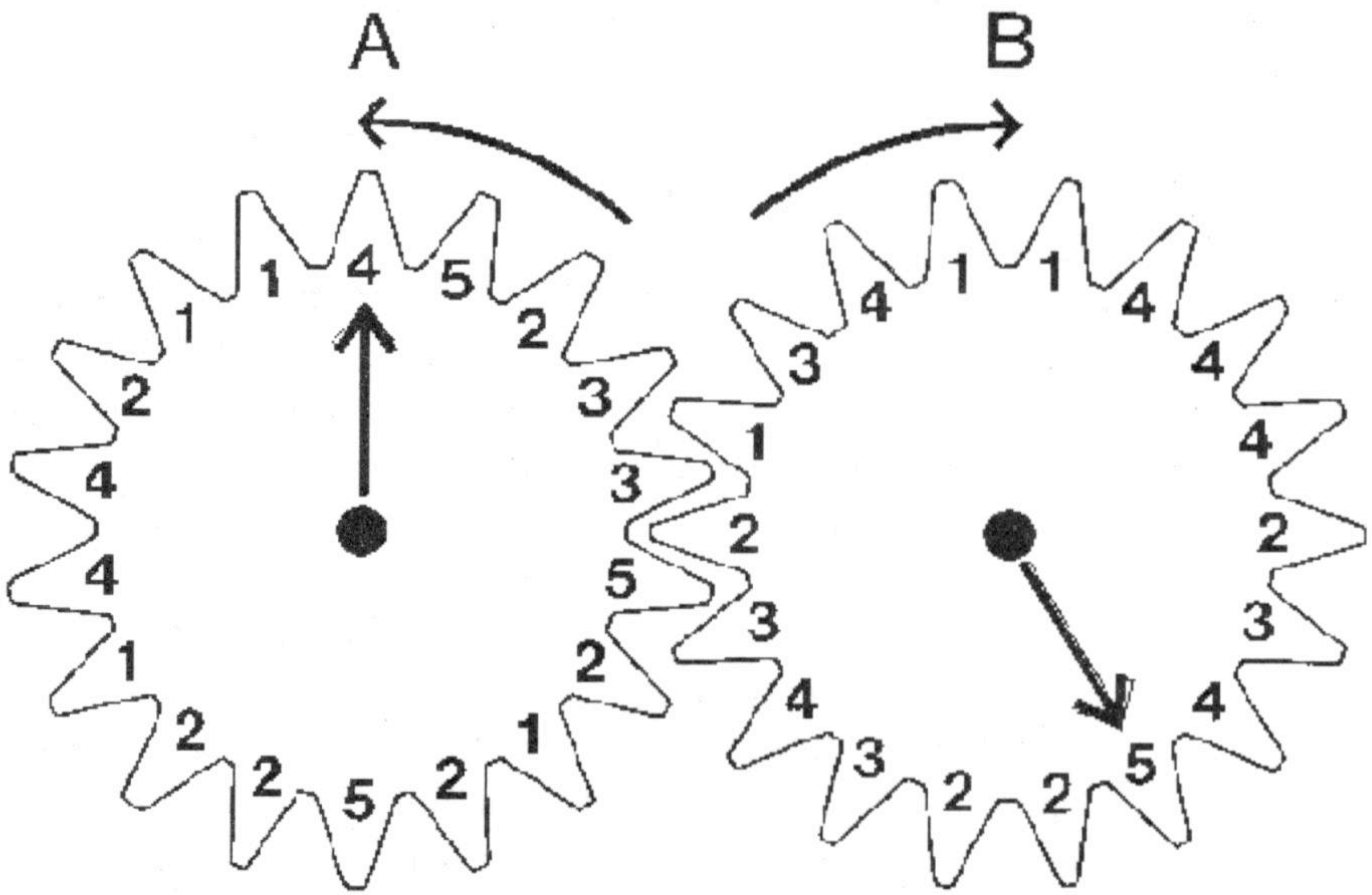
Bregman demo 2. “Auditory camouflage”

Stream segregation and rhythmic information



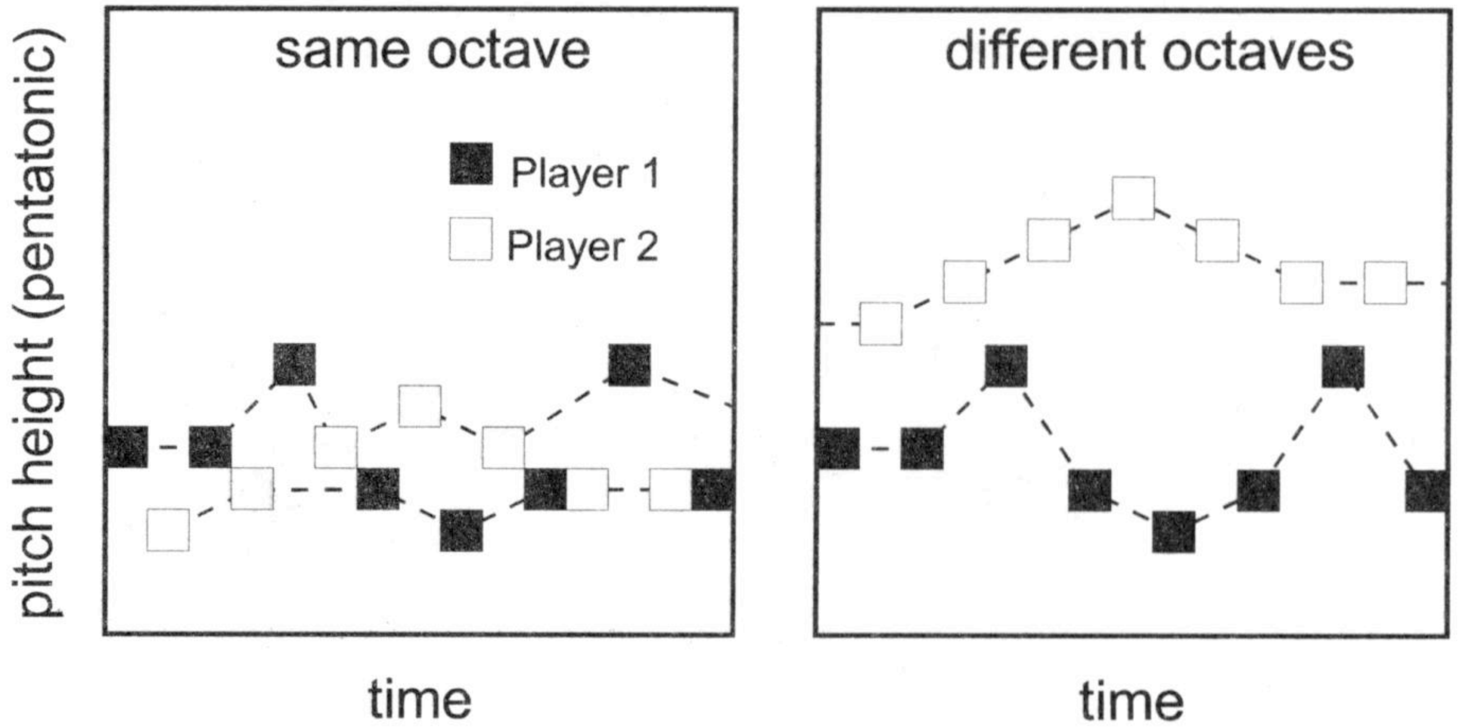
Bregman demo 3. The perception of rhythmic information depends on the segmentation of the auditory streams.

Streaming in African xylophone music



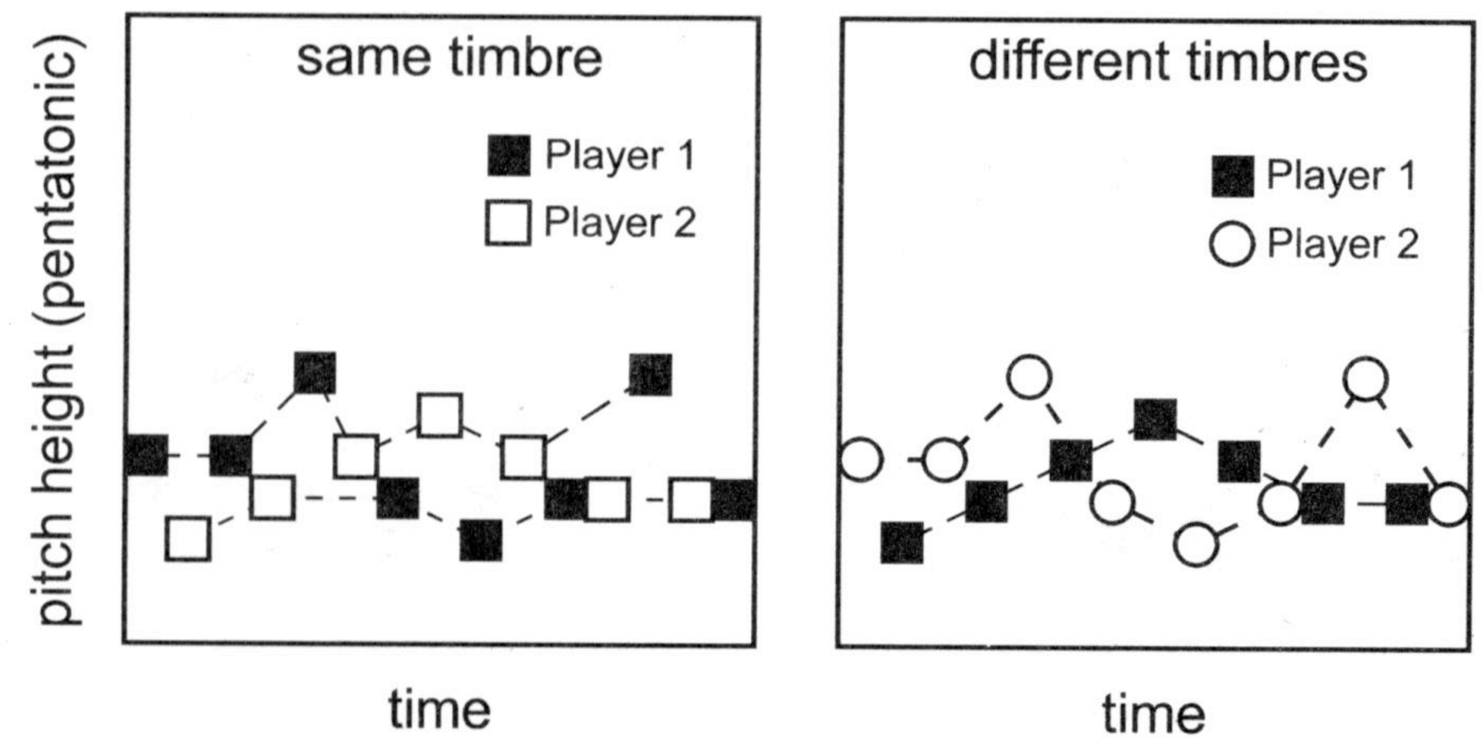
Bregman demo 7.

Effect of pitch range



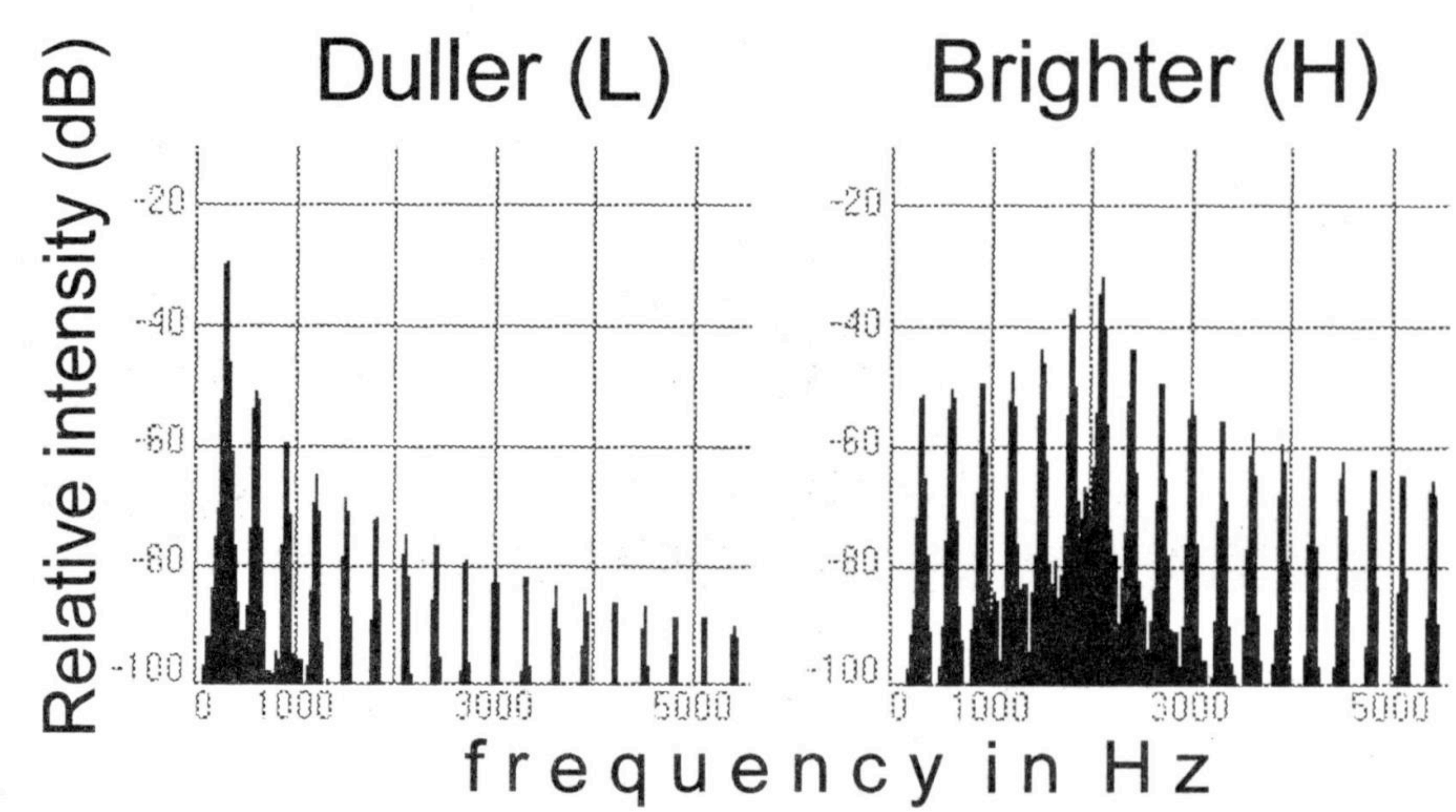
Bregman demo 8.

Effect of timbre



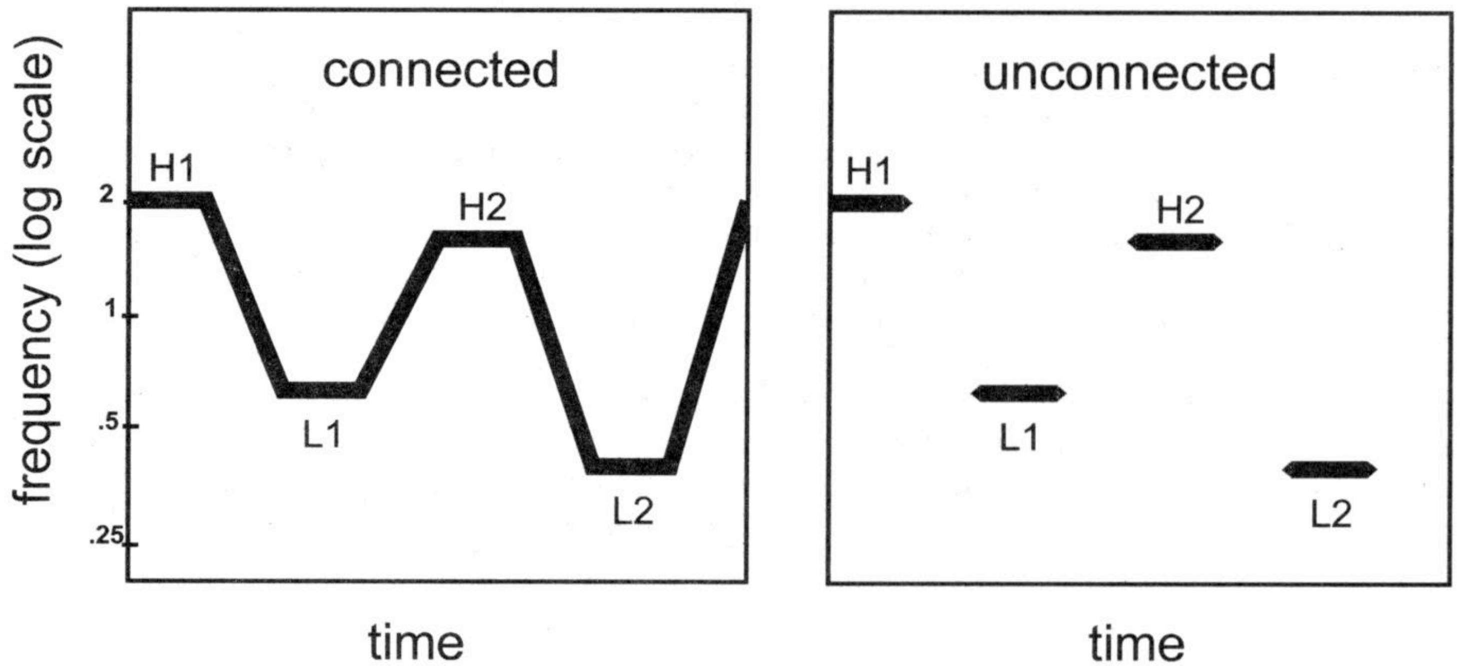
Bregman demo 9.

Stream segregation based on spectral peak position



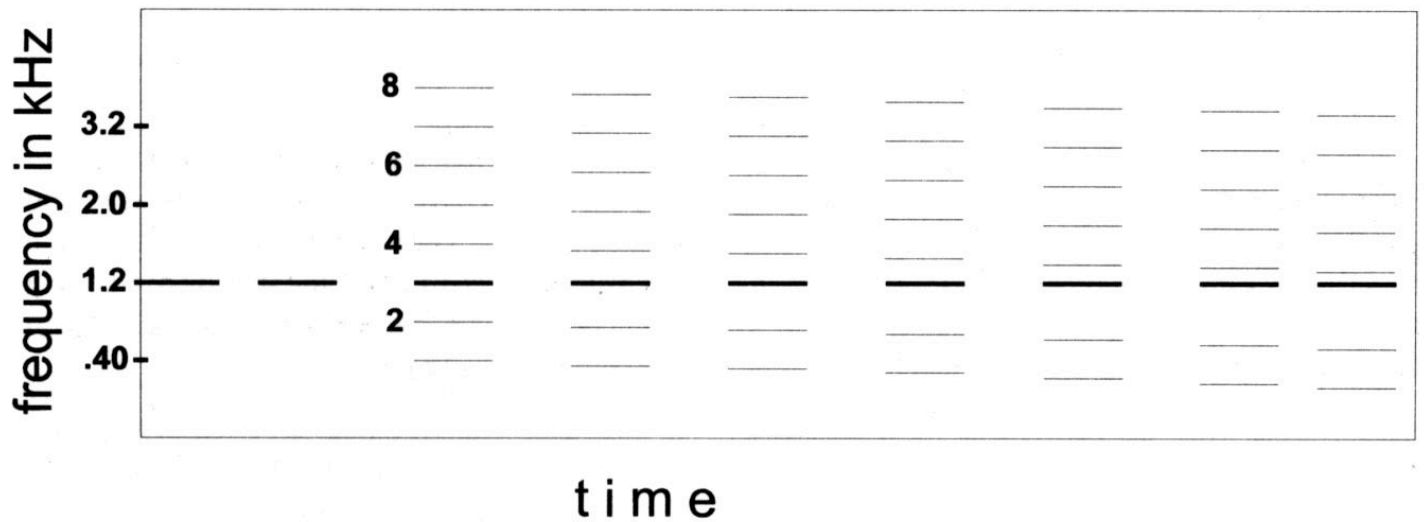
Bregman demo 10.

Effect of connectedness



Bregman demo 12.

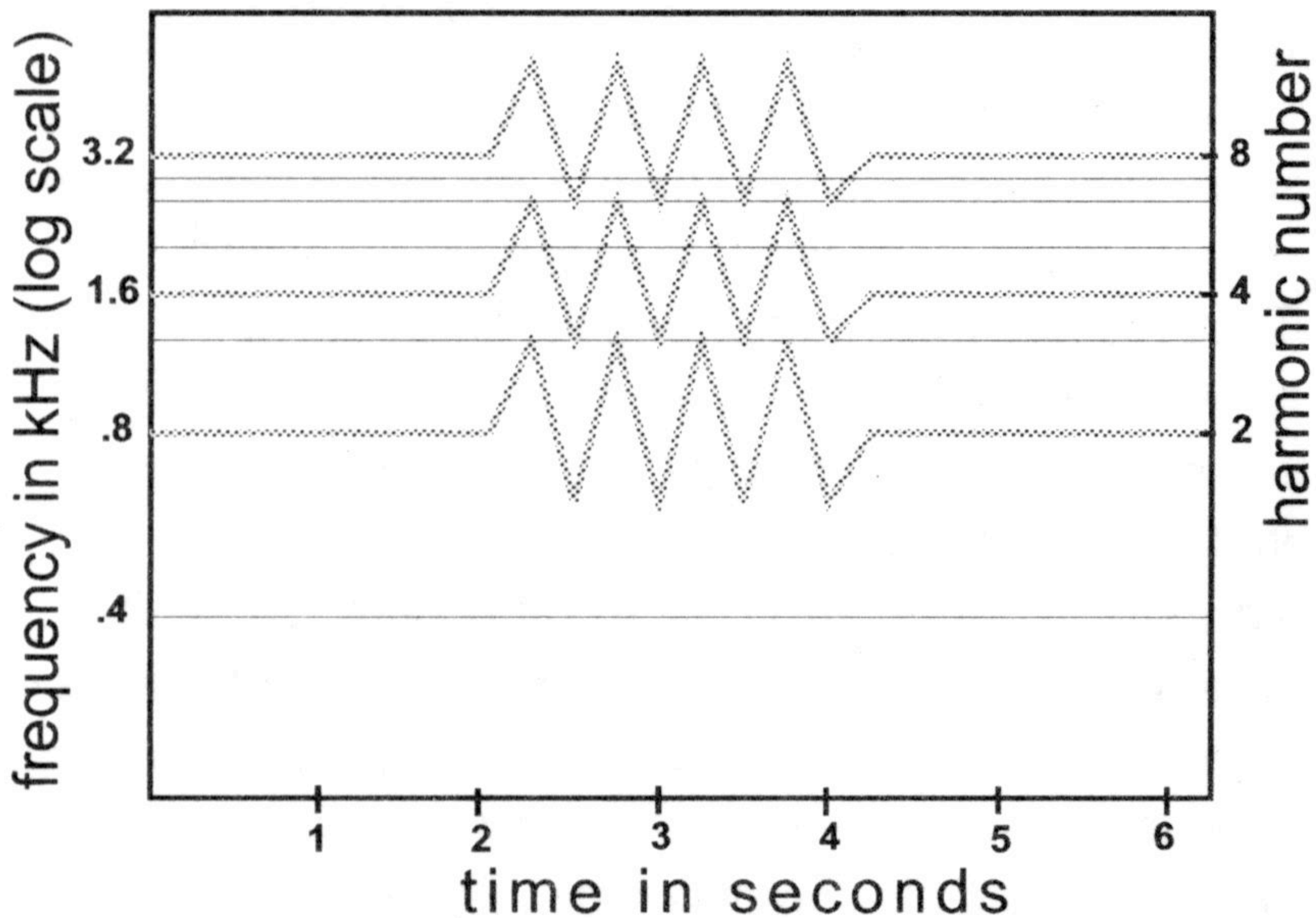
Grouping based on common fundamental frequency



Bregman demo 18.

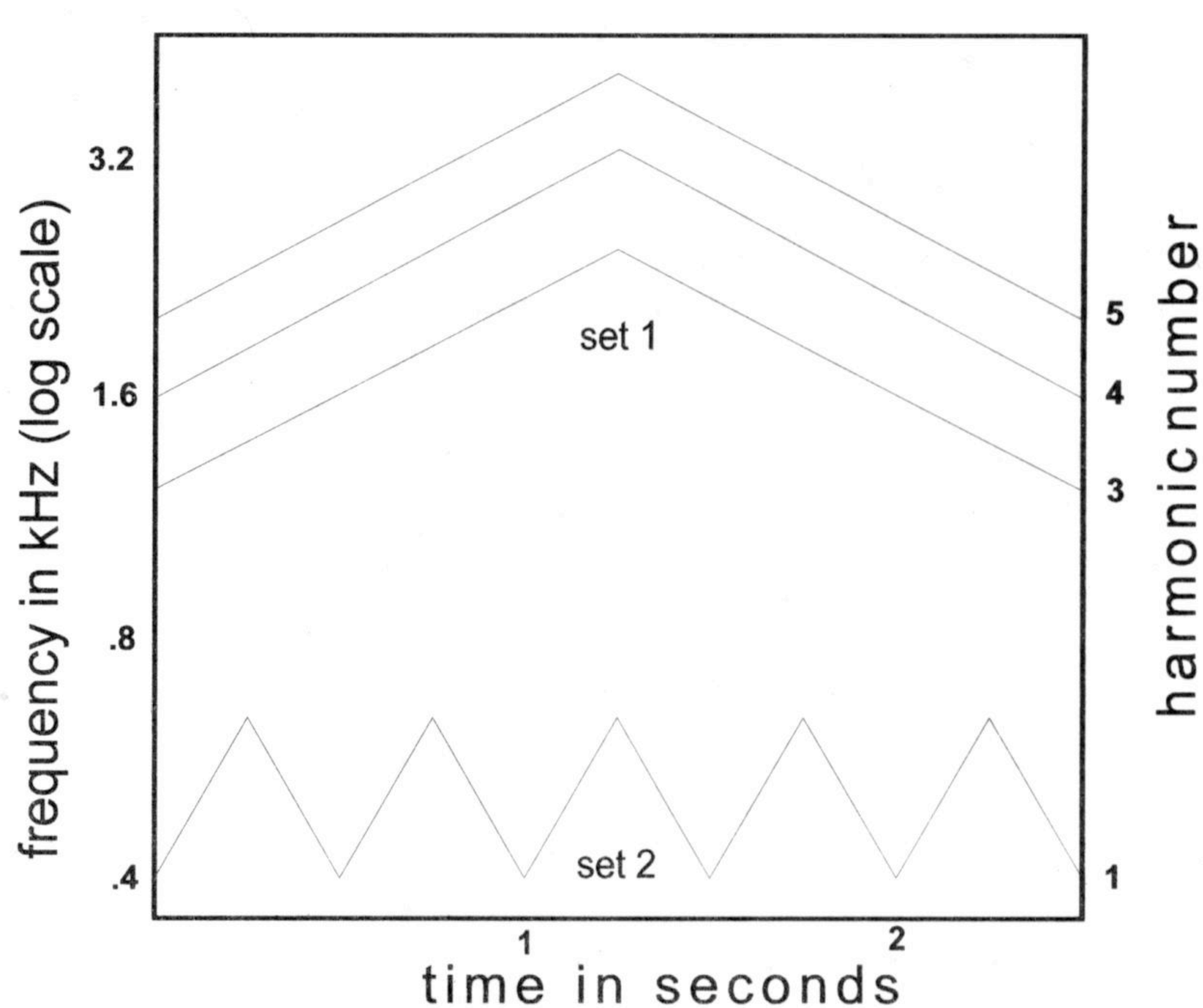
- Do the set of frequencies come from the same source?
- Those of a common fundametal a grouped together.

Fusion based on common frequency modulation



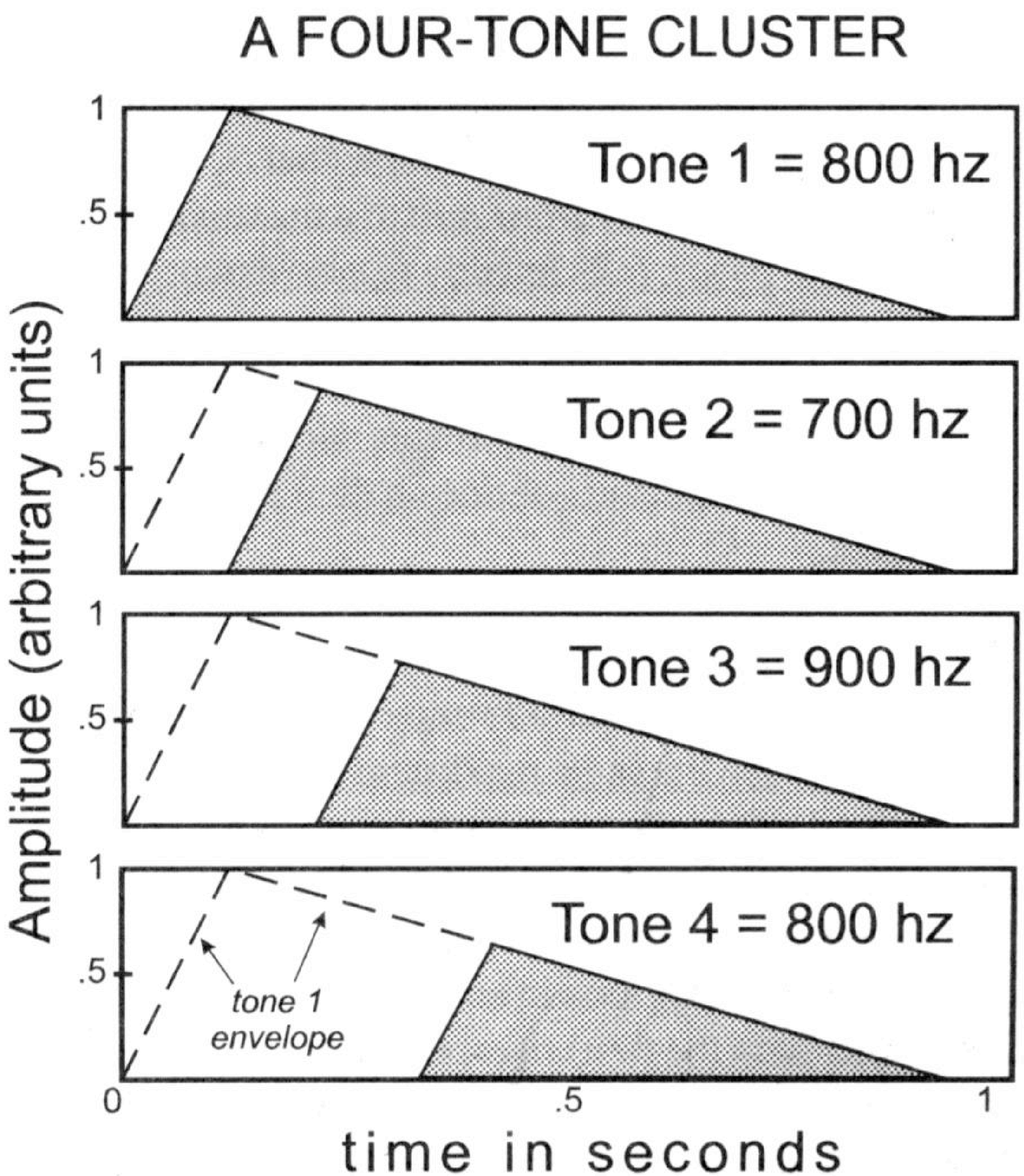
Bregman demo 19.

Fusion based on common frequency modulation



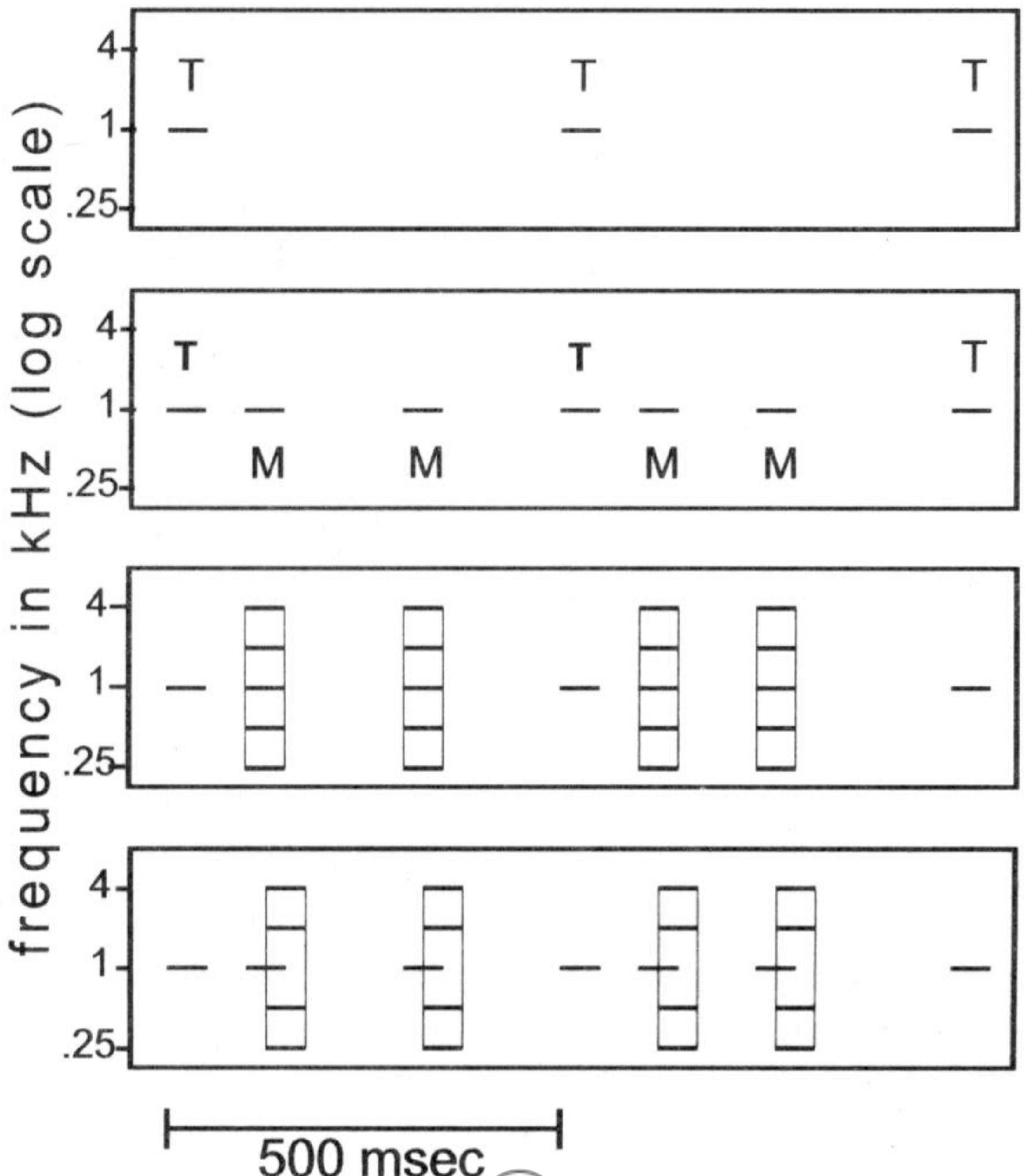
Bregman demo 20.

Onset rate affects segregation



Bregman demo 21.

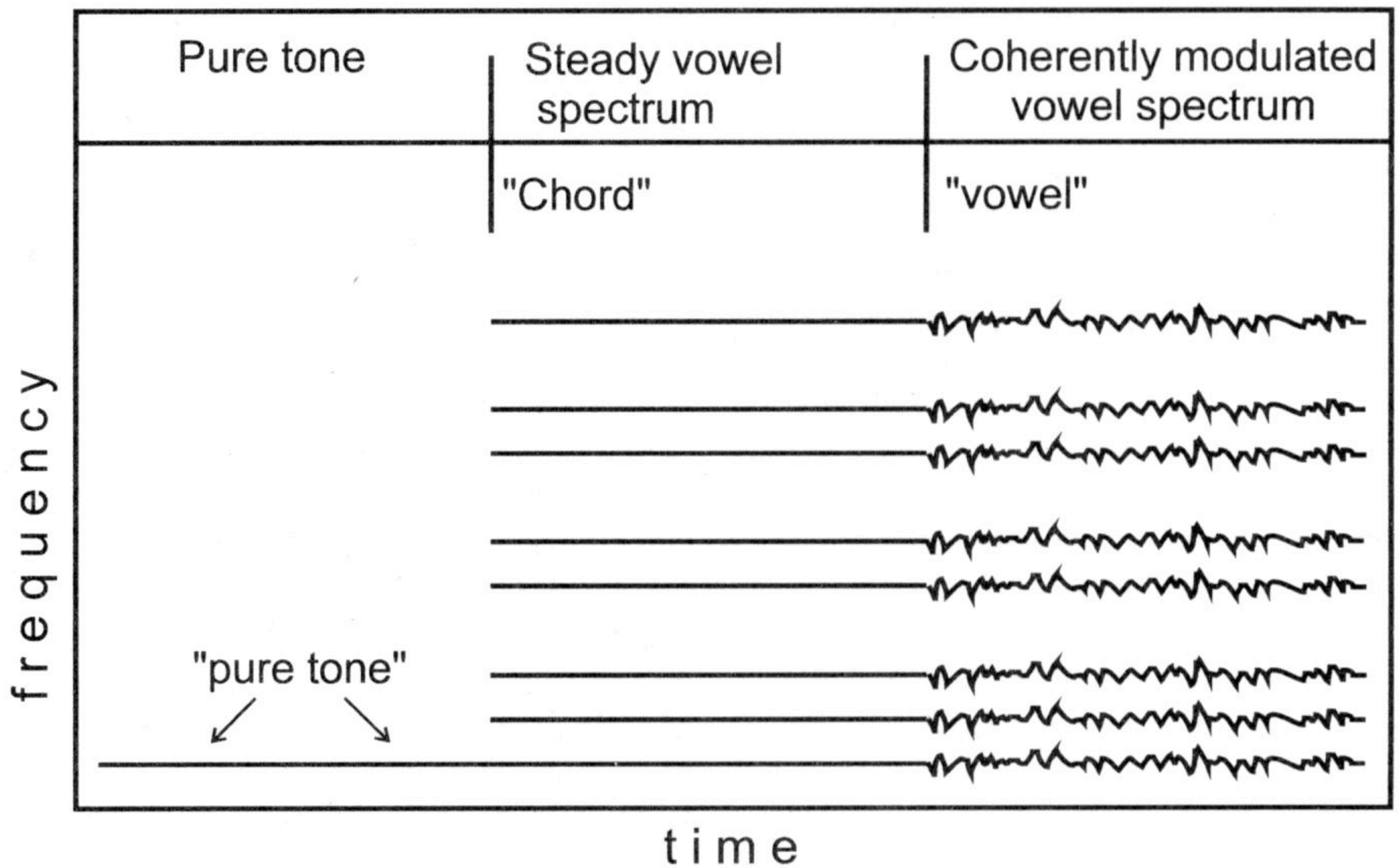
Rhythmic masking release



Bregman demo 22.

- This is an example of temporal grouping with overlapping frequency ranges.
- Frequecies outside the "critical band" of a target influence the ability to hear it

Micro modulation in voice perception



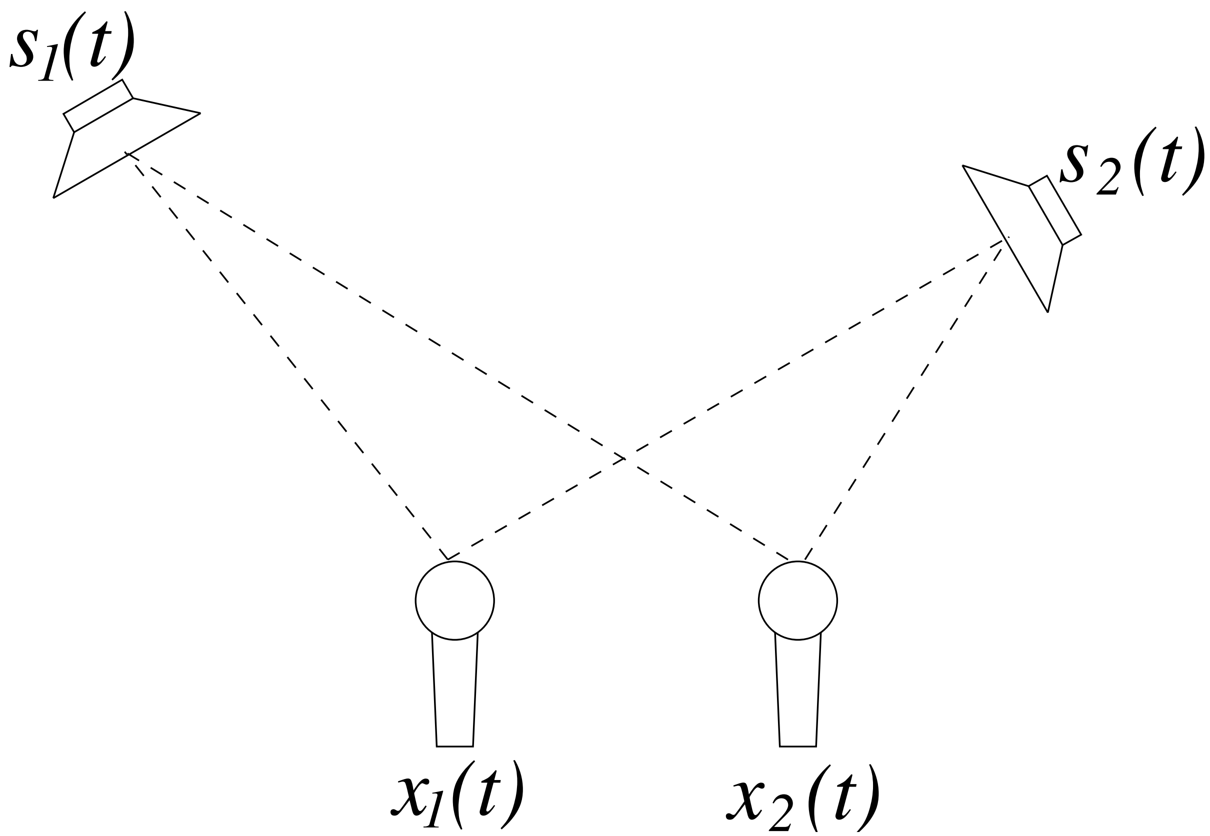
Bregman demo 24.

- Normal speech vowels contain small, random fluctuations
- the pure tone plus harmonic has weak vowel quality
- addition of micro-modulation enhances the perception of the sound as a singing vowel

Blind source separation

Modeling the cocktail party problem

Suppose we have two speakers (sources), $s_1(t)$ and $s_2(t)$ and two microphones (mixtures), $x_1(t)$ and $x_2(t)$:



The general problem is called *blind source separation*. We only observe the mixtures and are "blind" to the sources.

How do we model this in general?

A general formulation

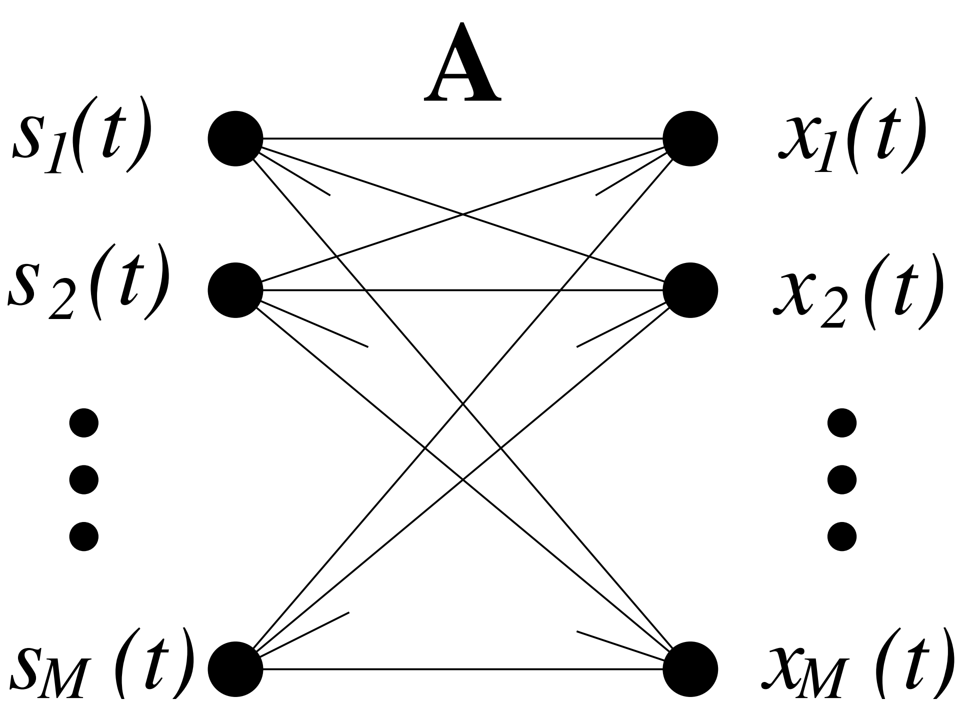
Suppose we have M sources

$$s_1(t), \ldots, s_M(t)$$

and M mixtures

$$x_1(t), \ldots, x_2(t)$$

This can represented diagramatically as:



What's the simplest mathematical model?

A general formulation

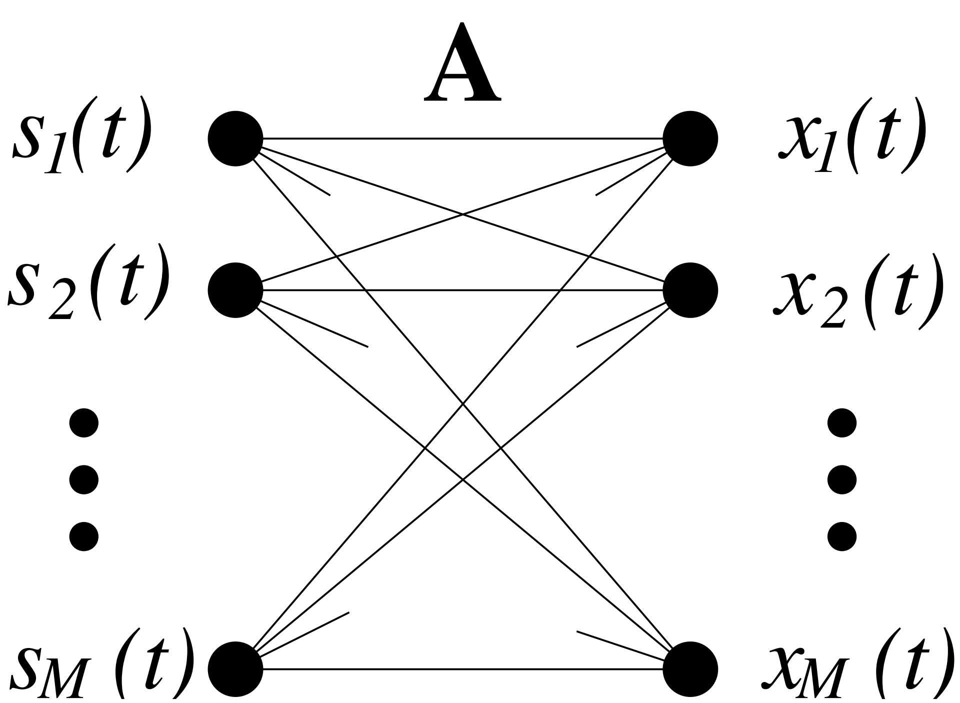
Suppose we have M sources

$$s_1(t), \ldots, s_M(t)$$

and M mixtures

$$x_1(t), \ldots, x_2(t)$$

This can represented diagramatically as:



What's the simplest mathematical model?

Assume linear, instantaneous mixing:

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t)$$

A general formulation

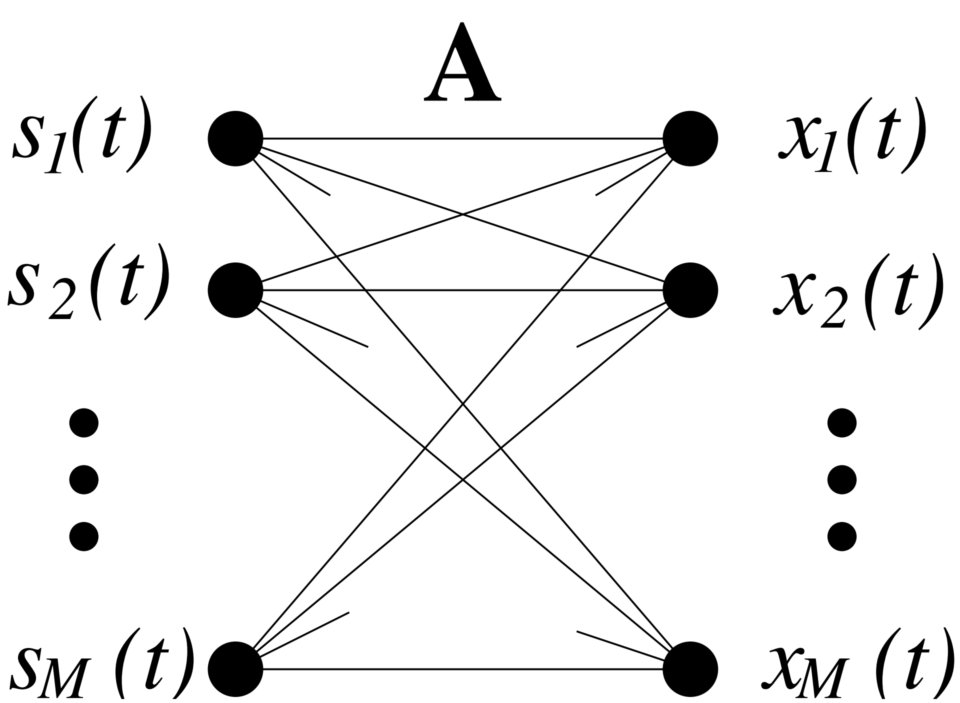
Suppose we have M sources

$$s_1(t), \ldots, s_M(t)$$

and M mixtures

$$x_1(t), \ldots, x_2(t)$$

This can represented diagramatically as:



What's the simplest mathematical model?

Assume linear, instantaneous mixing:

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t)$$

A is a called the *mixing matrix*.

What does this model ignore?

A general formulation

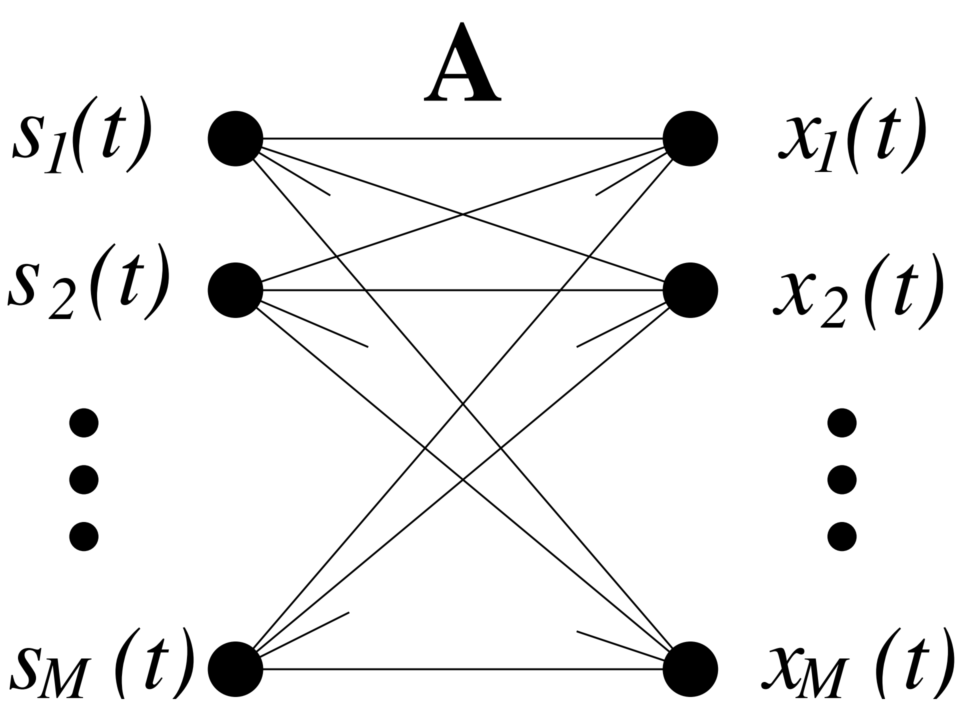
Suppose we have M sources

$$s_1(t), \ldots, s_M(t)$$

and M mixtures

$$x_1(t), \ldots, x_2(t)$$

This can represented diagramatically as:



What's the simplest mathematical model?

Assume linear, instantaneous mixing:

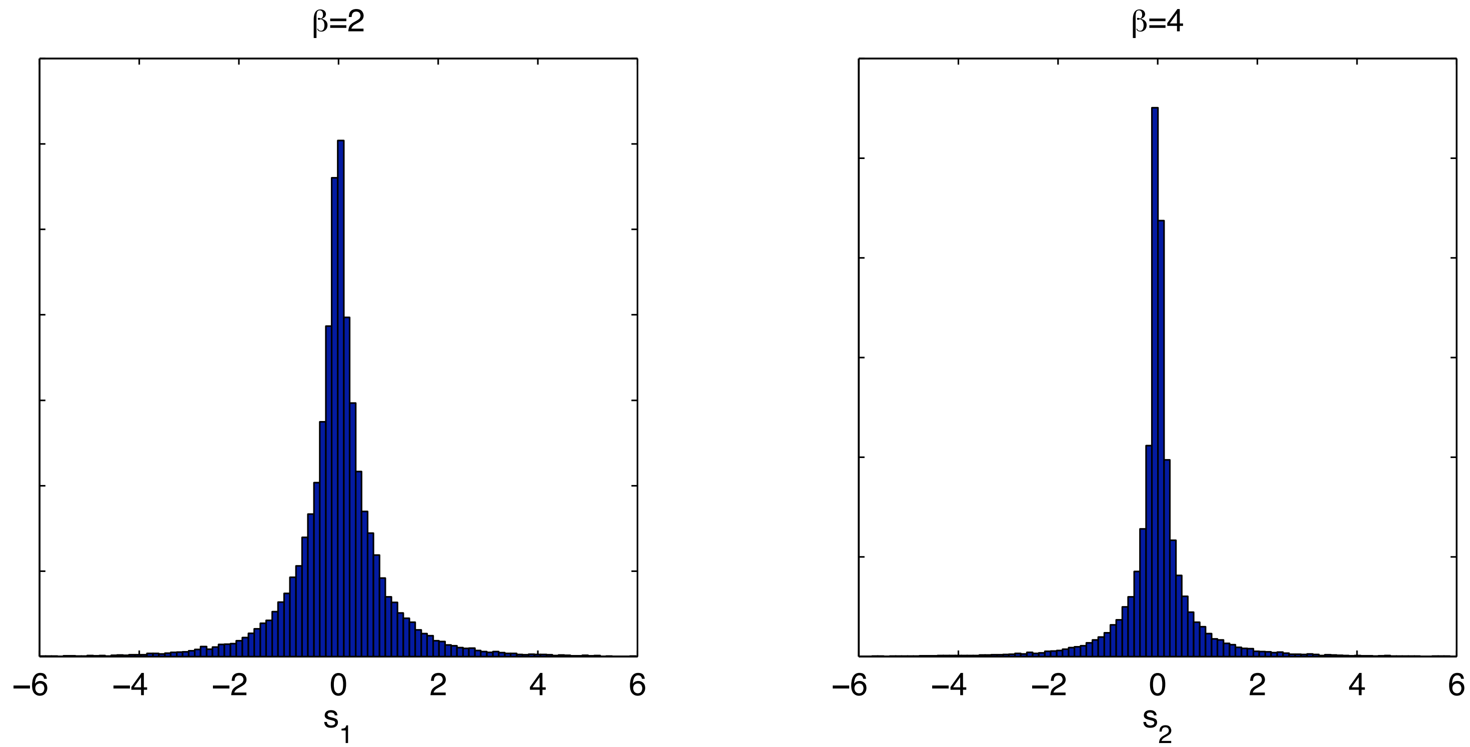
$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t)$$

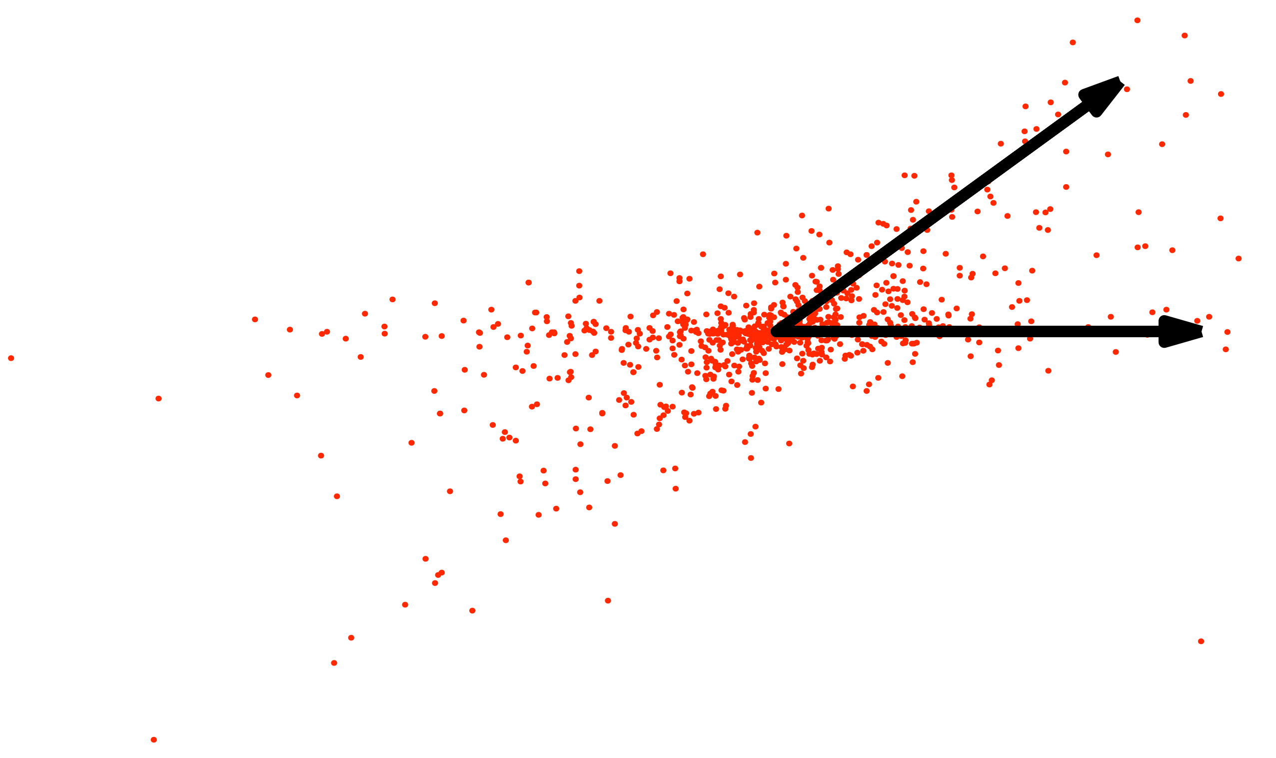
A is a called the *mixing matrix*.

What does this model ignore?

- room acoustics, reverberation, echos
- filtering, noise
- might have more than two sounds
- sounds might not come from a single source
- sound sources could change location

The distribution of sample points



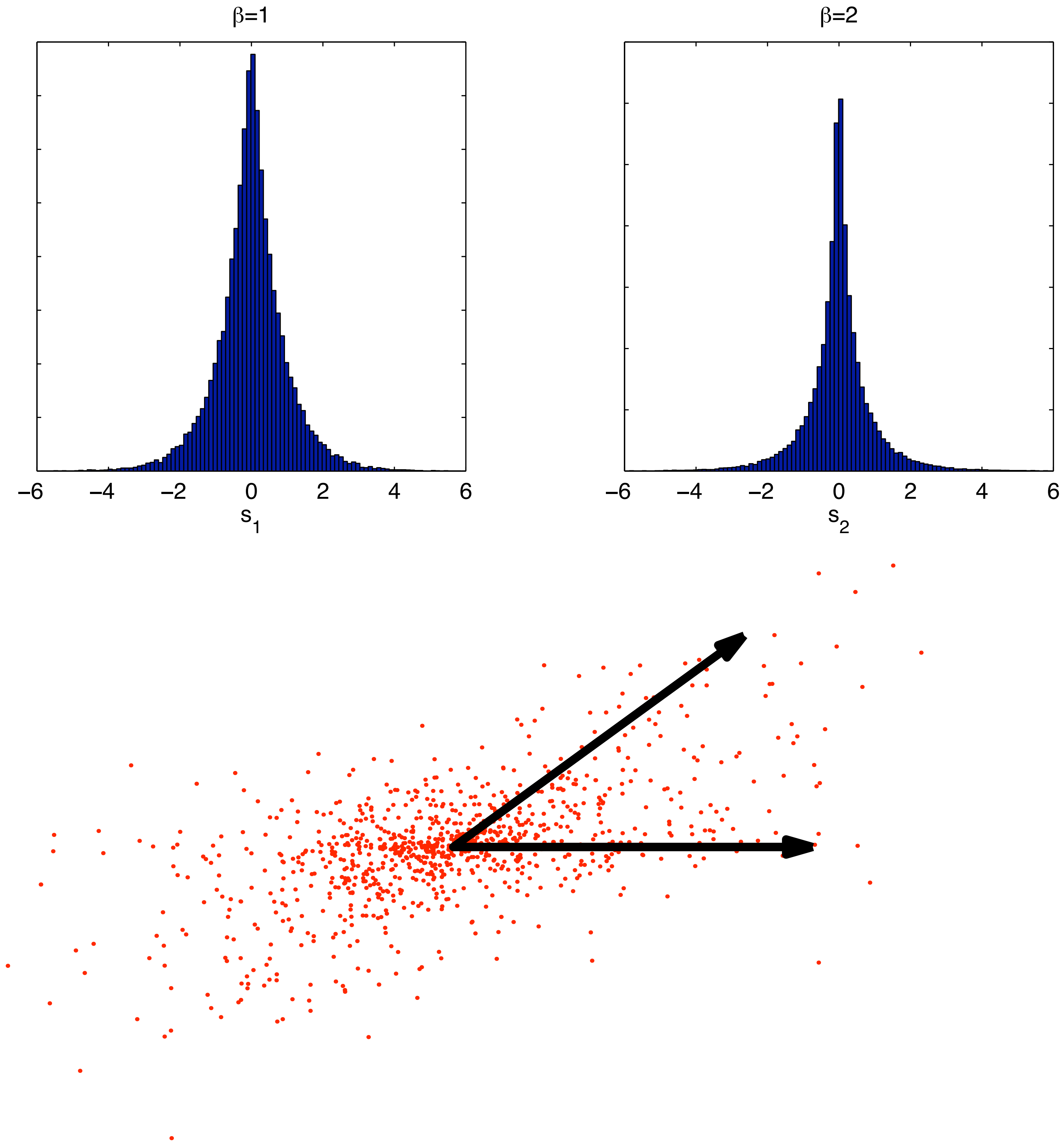


- The histograms show the amplitude distribution of each source.
- The scatter plot shows the 2D distribution of $x_1(t)$ vs $x_2(t)$.
- The parameter β characterizes the sharpness of the data using the distribution

$$P(s) \propto e^{-|s|^{2/(1+\beta)}/2}$$

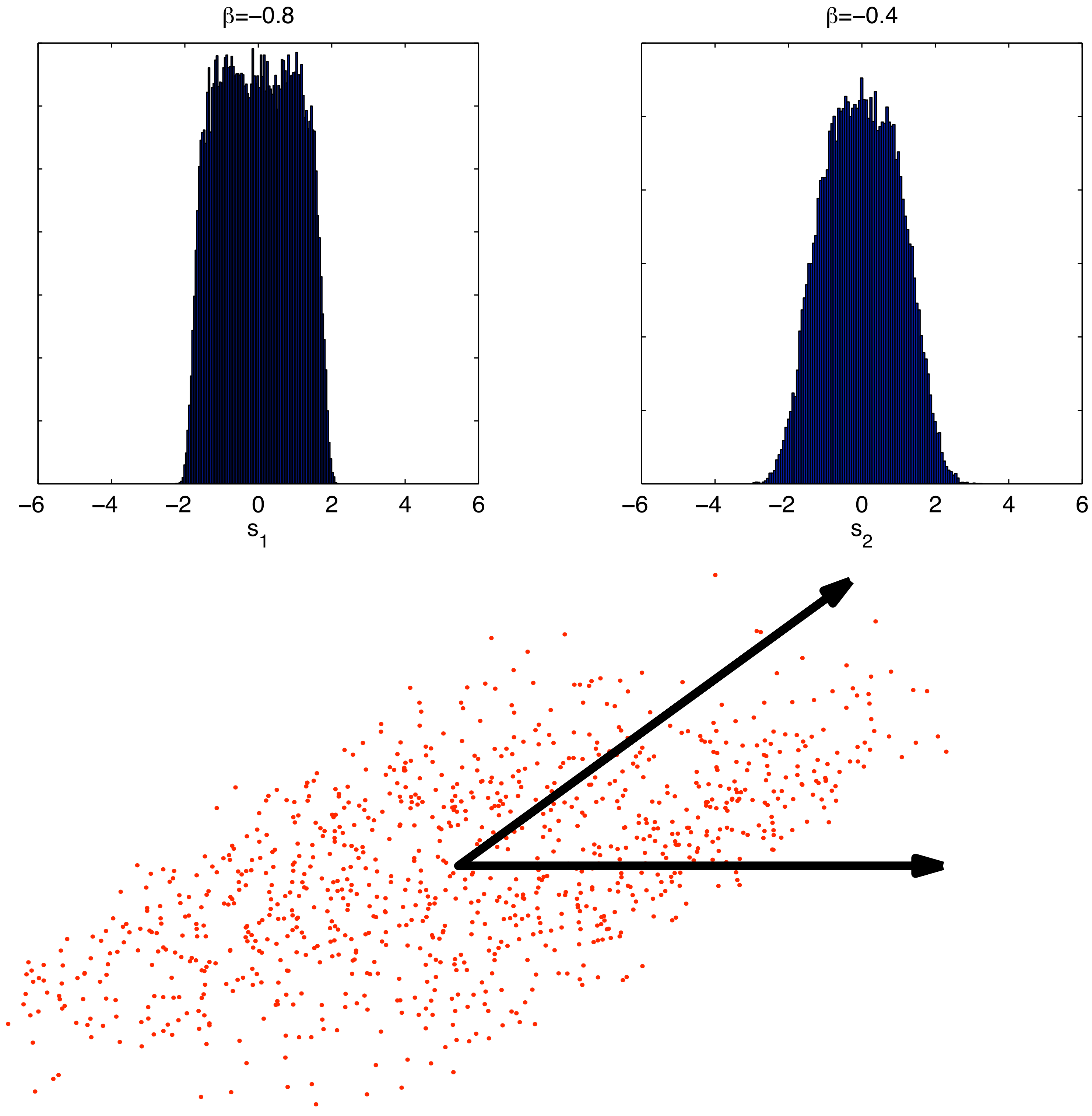
- $\beta = 0$ corresponds to a Gaussian distribution

The distribution of sample points



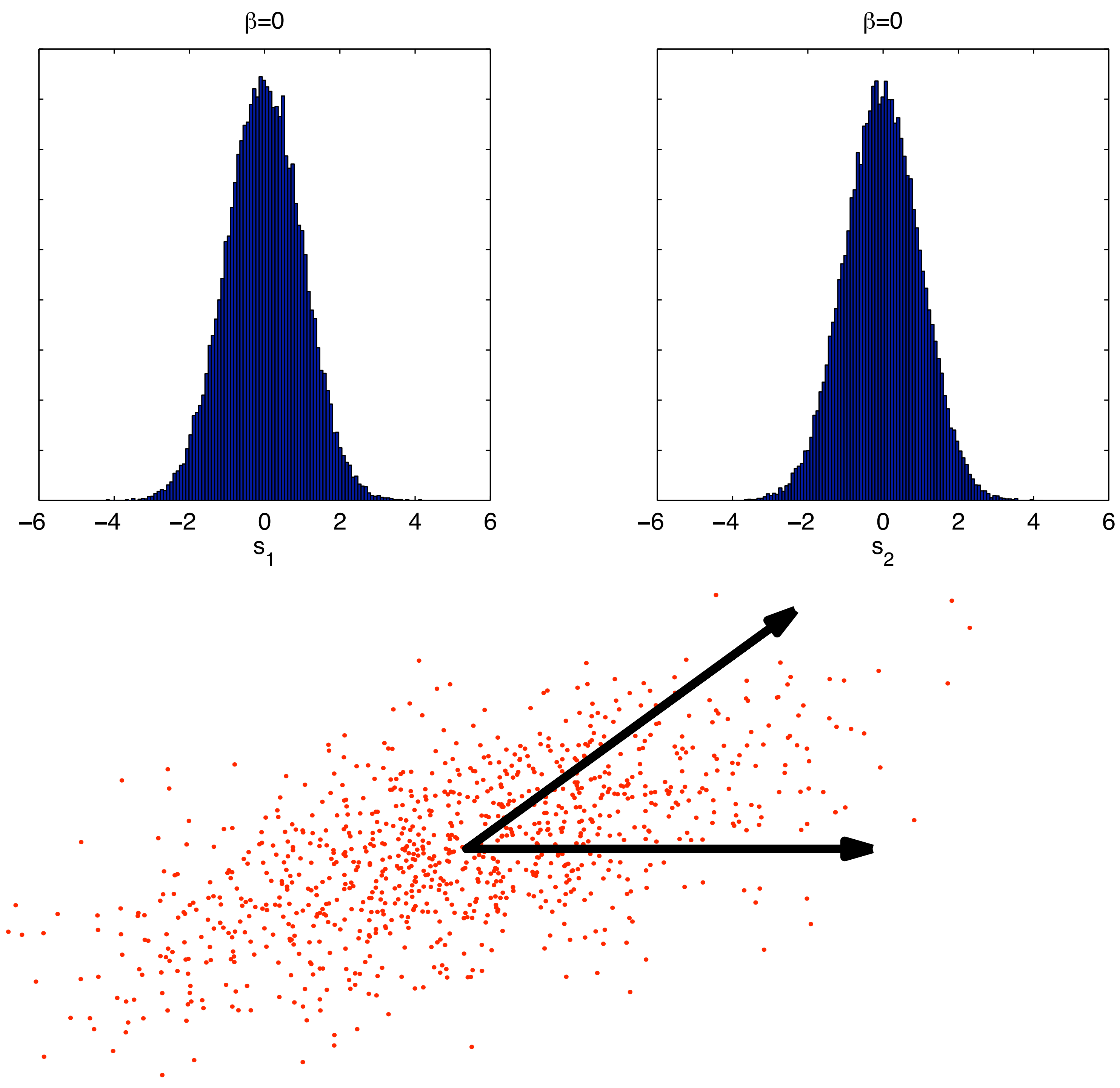
- Sources can have a wide range of amplitude distributions.
- Different directions correspond to different mixing matrices.
- A mixture of non-Gaussian sources create a distribution whose axes can be uniquely determined (within a sign change).

The distribution of sample points



The axes of Sub-Gaussian ources, i.e $\beta < 0$ or negative kurtosis, can still be determined.

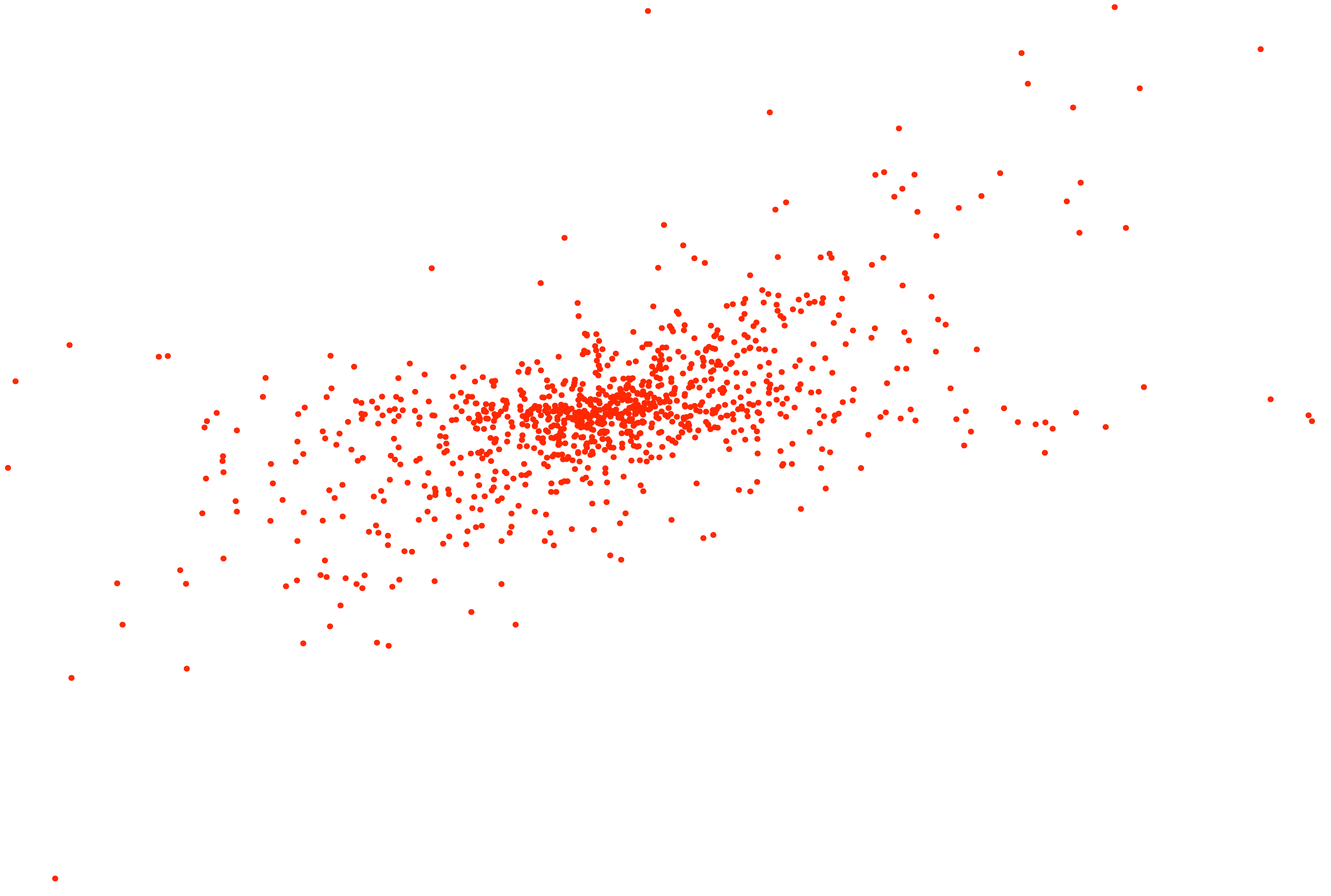
A Gaussian distribution of sample points



The principal axes of two Gaussian sources are ambiguous.

- Why?
- Because the product of two Gaussians is still a Gaussian, so there are an infinite number of directions that fit the 2D distribution.

Inferring the (un)mixing matrix



How do we determine the axes from just the data?

Modeling non-Gaussian distributions

Learning objective: model statistical density of sources:
$\Rightarrow$ maximize $P(\mathbf{x}|\mathbf{A})$ over $\mathbf{A}$.

Probability of pattern ensemble is:

$$P(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N|\mathbf{A}) = \prod_k P(\mathbf{x}_k|\mathbf{A})$$

To obtain $P(\mathbf{x}|\mathbf{A})$ marginalize over $\mathbf{s}$:

$$\begin{aligned} P(\mathbf{x}|\mathbf{A}) &= \int d\mathbf{s}\, P(\mathbf{x}|\mathbf{A}, \mathbf{s}) P(\mathbf{s}) \\ &= \frac{P(\mathbf{s})}{|\det \mathbf{A}|} \end{aligned}$$

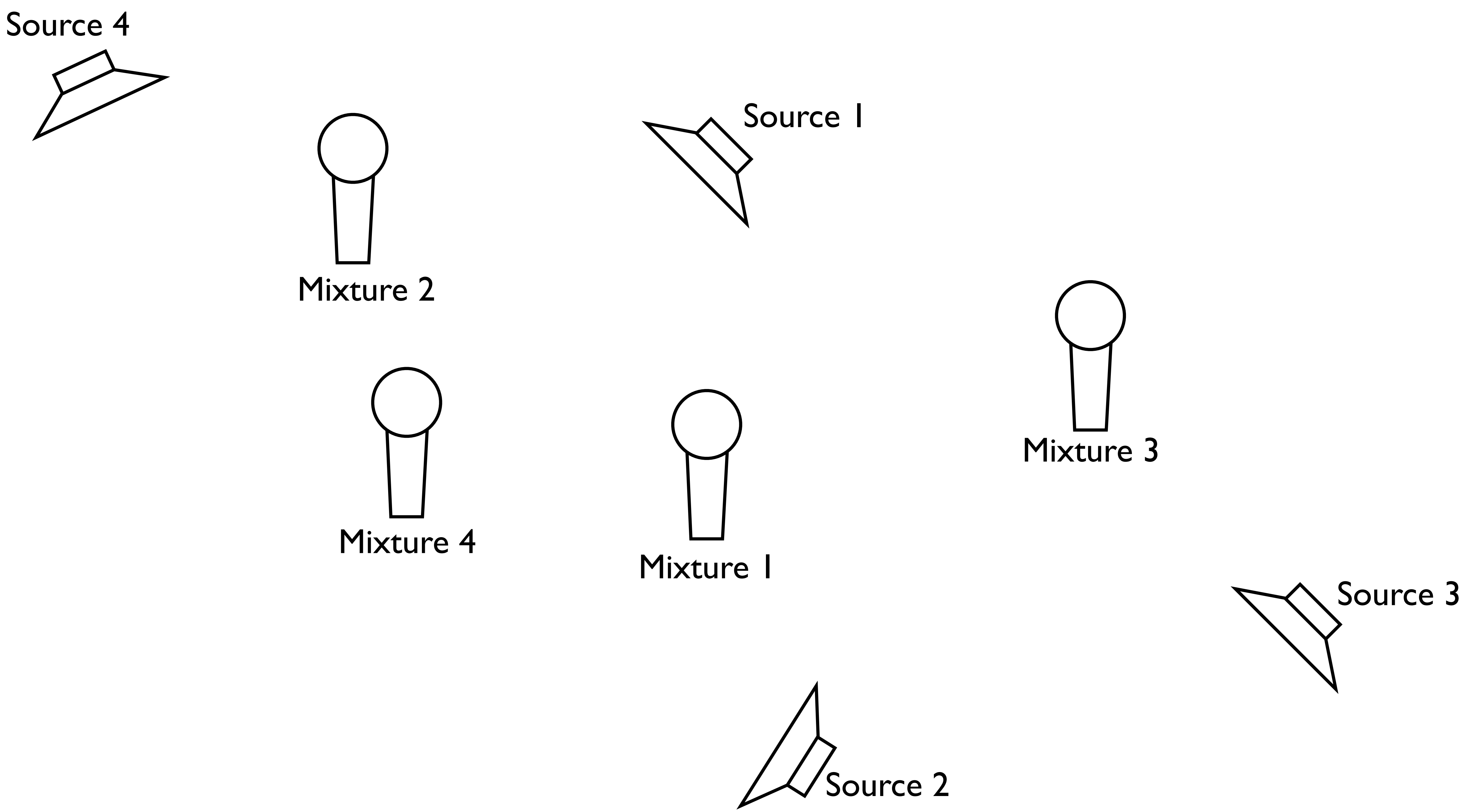
Learning rule (ICA):

$$\begin{aligned} \Delta\mathbf{A} &\propto \mathbf{A}\mathbf{A}^T \frac{\partial}{\partial \mathbf{A}} \log P(\mathbf{x}|\mathbf{A}) \\ &= -\mathbf{A}(\mathbf{z}\mathbf{s}^T - \mathbf{I}), \end{aligned}$$

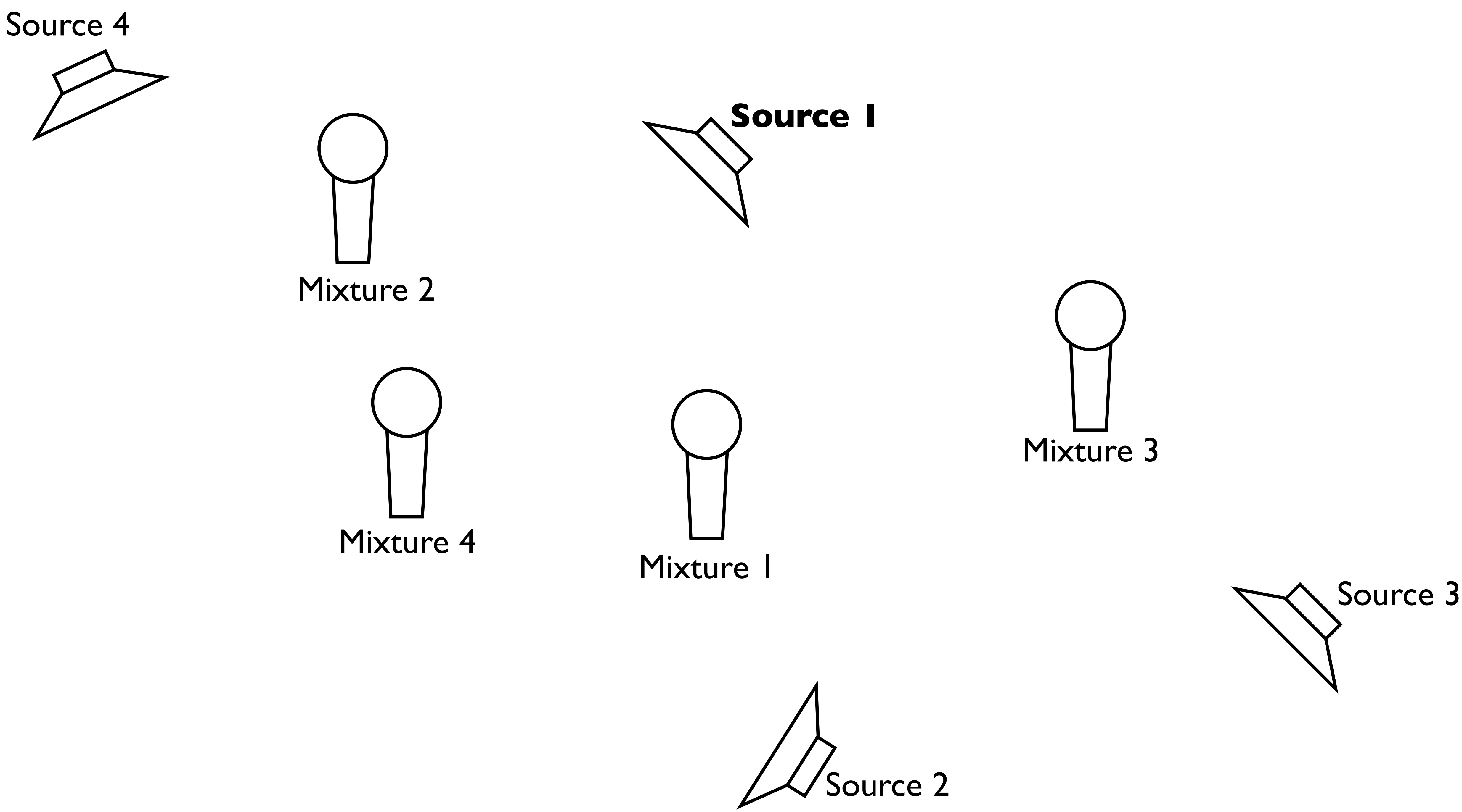
where $\mathbf{z} = (\log P(\mathbf{s}))'$. Use $P(s_i) \sim \mathsf{ExPwr}(s_i|\mu, \sigma, \beta_i)$.

This is identical to the procedure that was used to learn efficient codes, i.e *independent component analysis*.

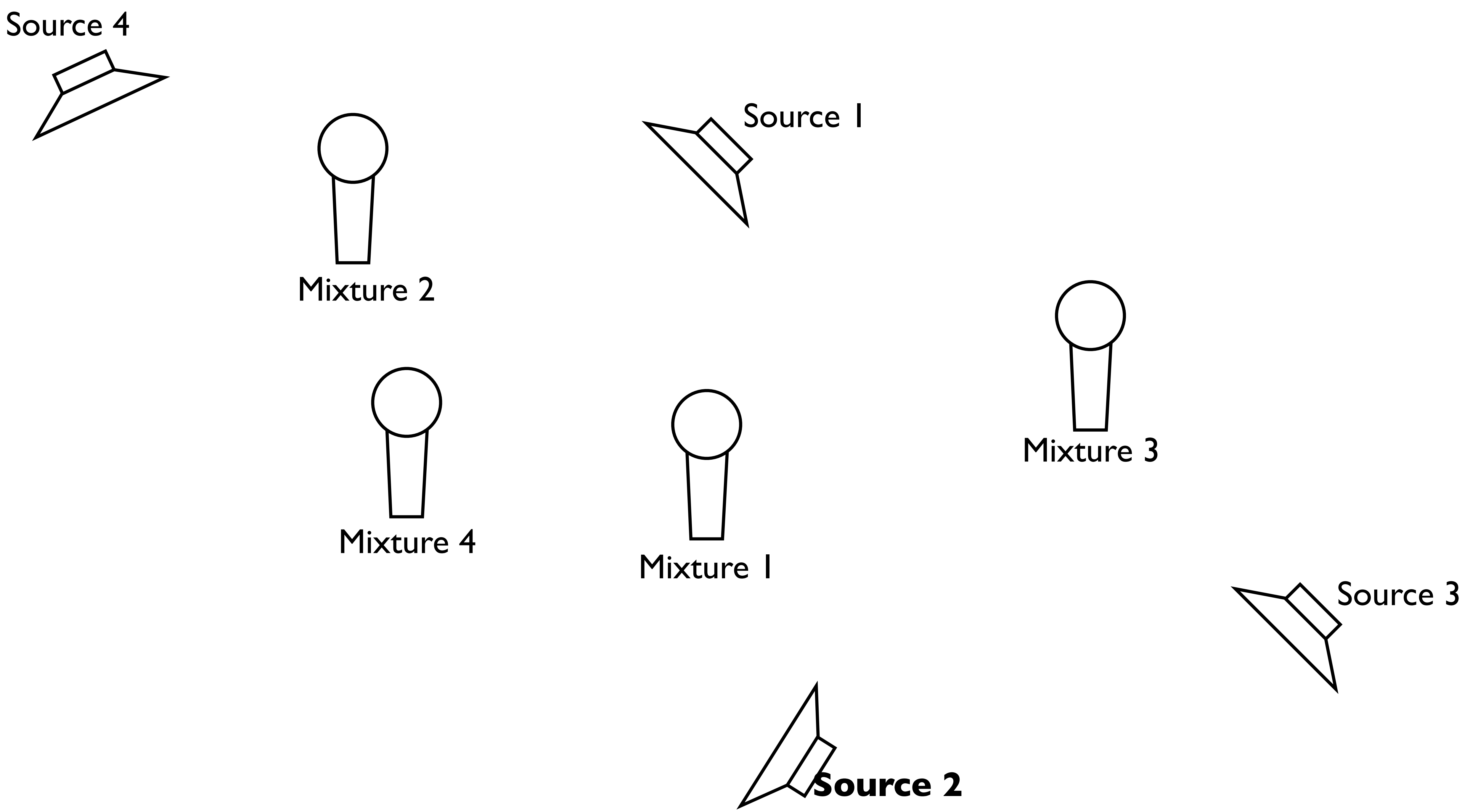
Separating mixtures of real sources



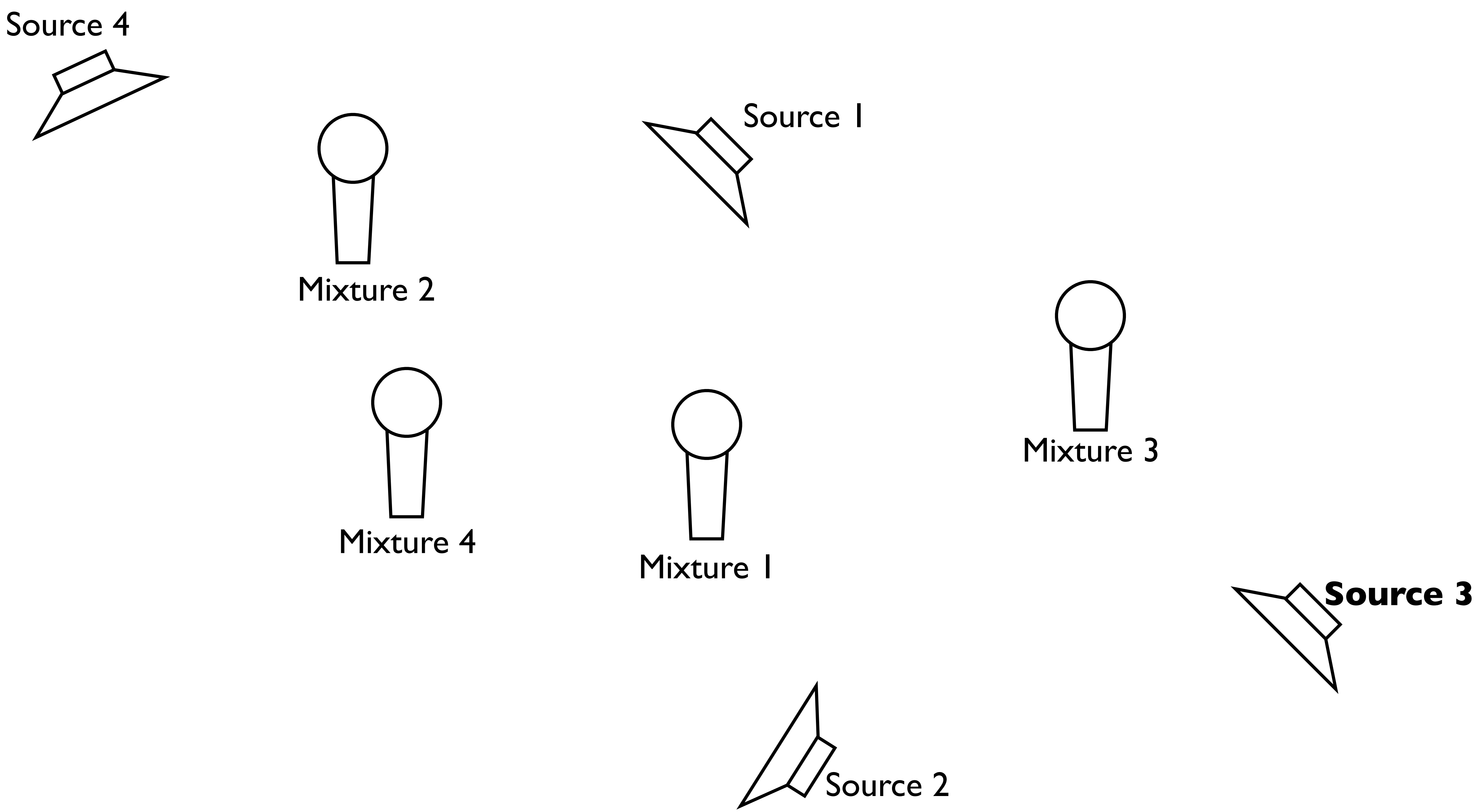
Separating mixtures of real sources



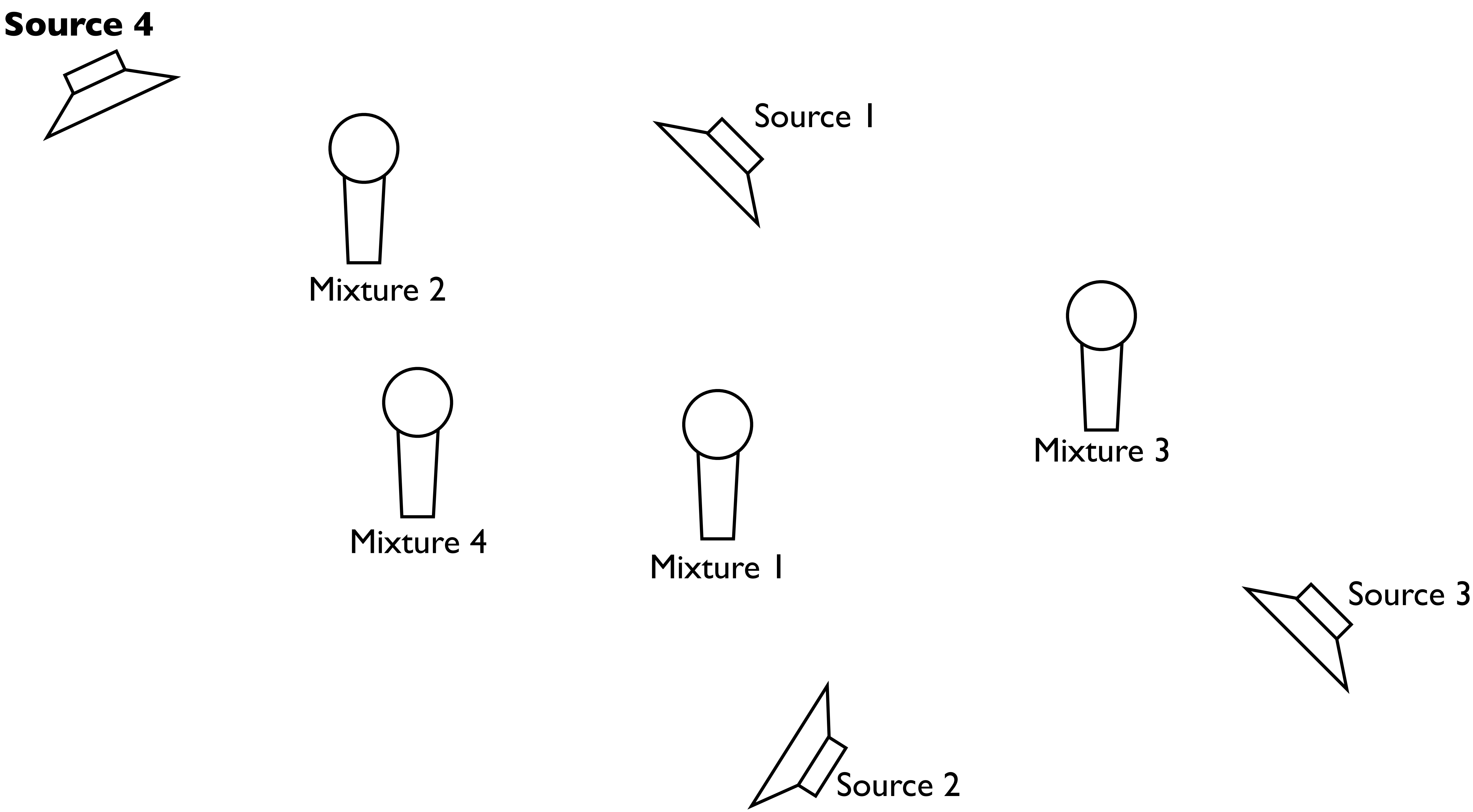
Separating mixtures of real sources



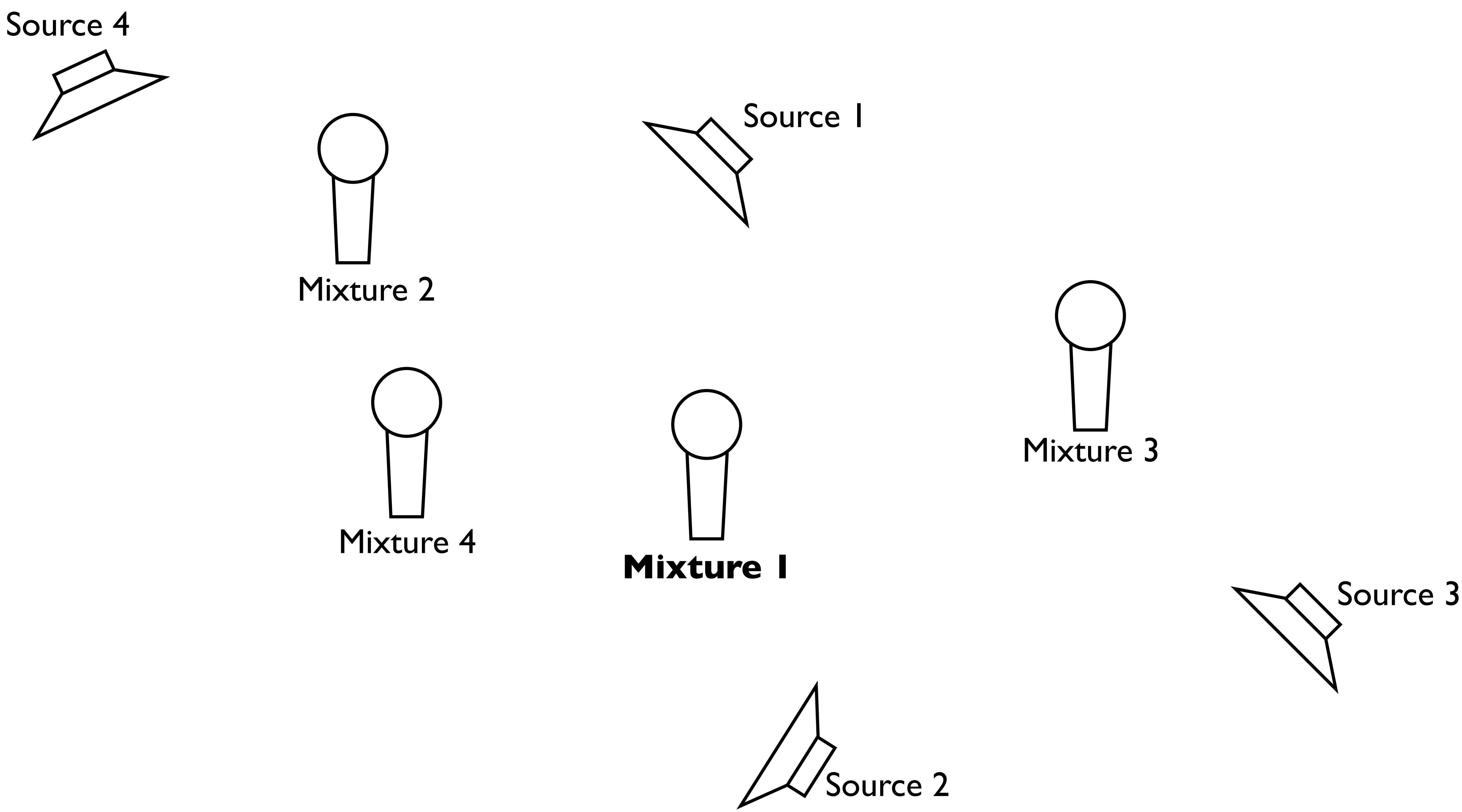
Separating mixtures of real sources



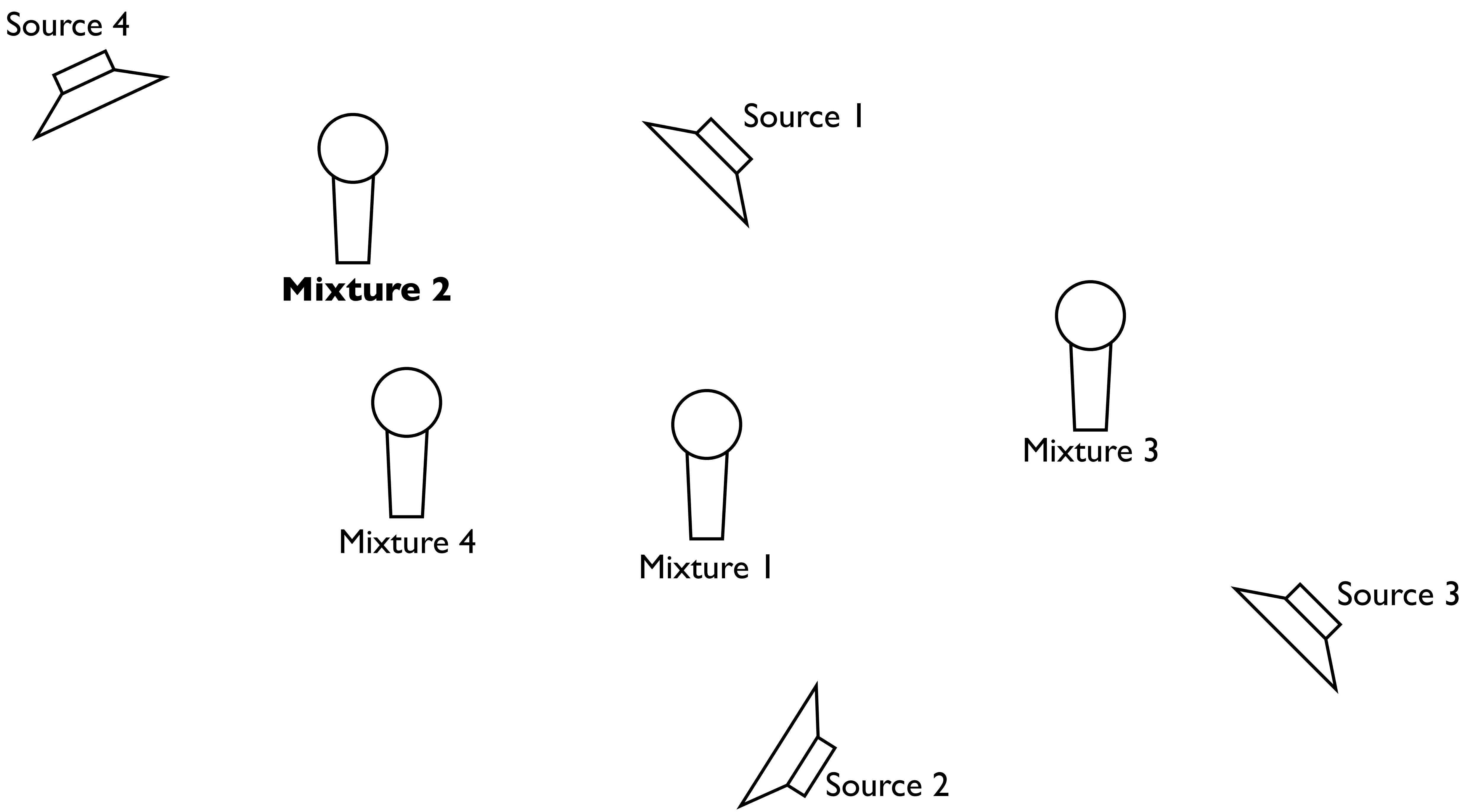
Separating mixtures of real sources



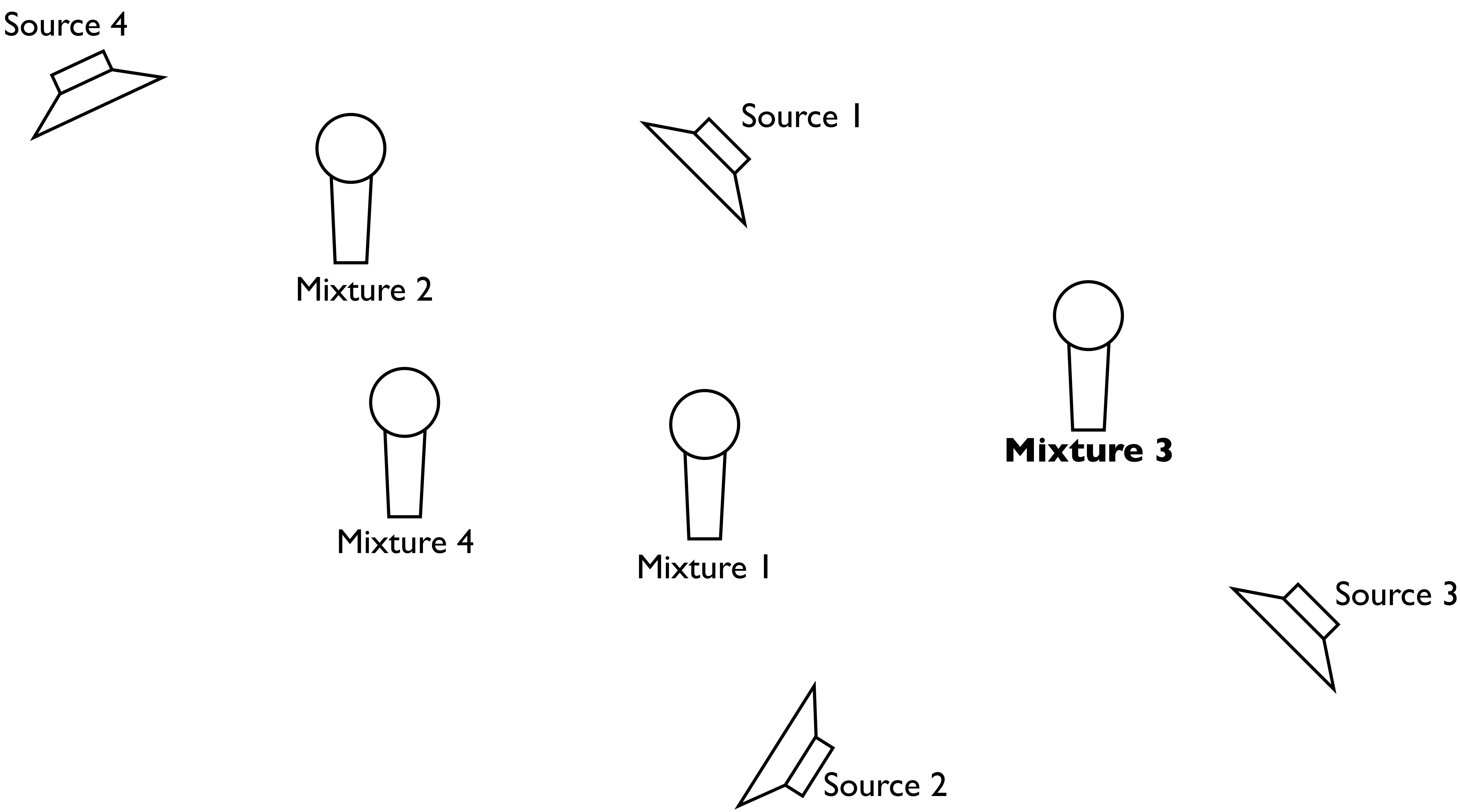
Separating mixtures of real sources



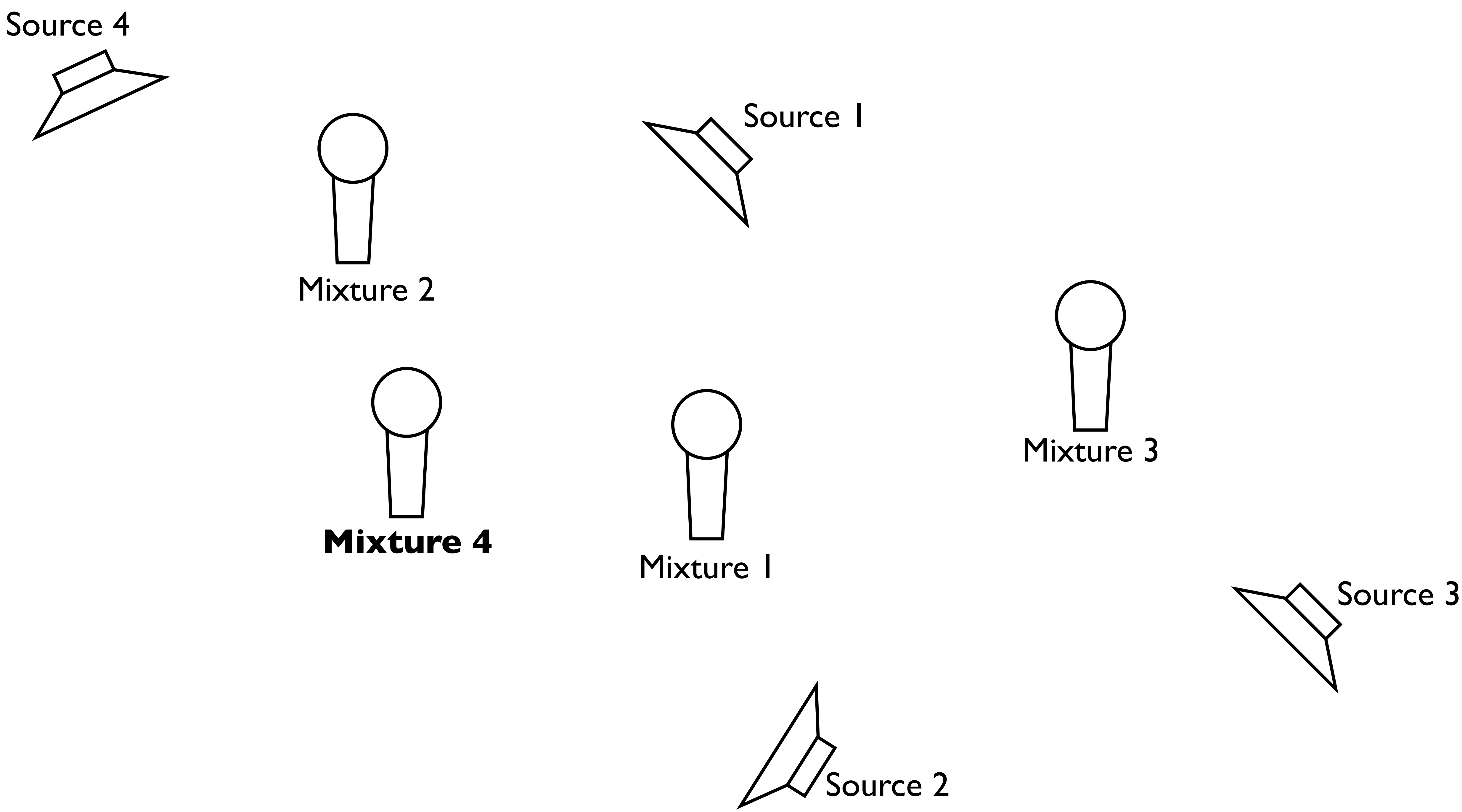
Separating mixtures of real sources



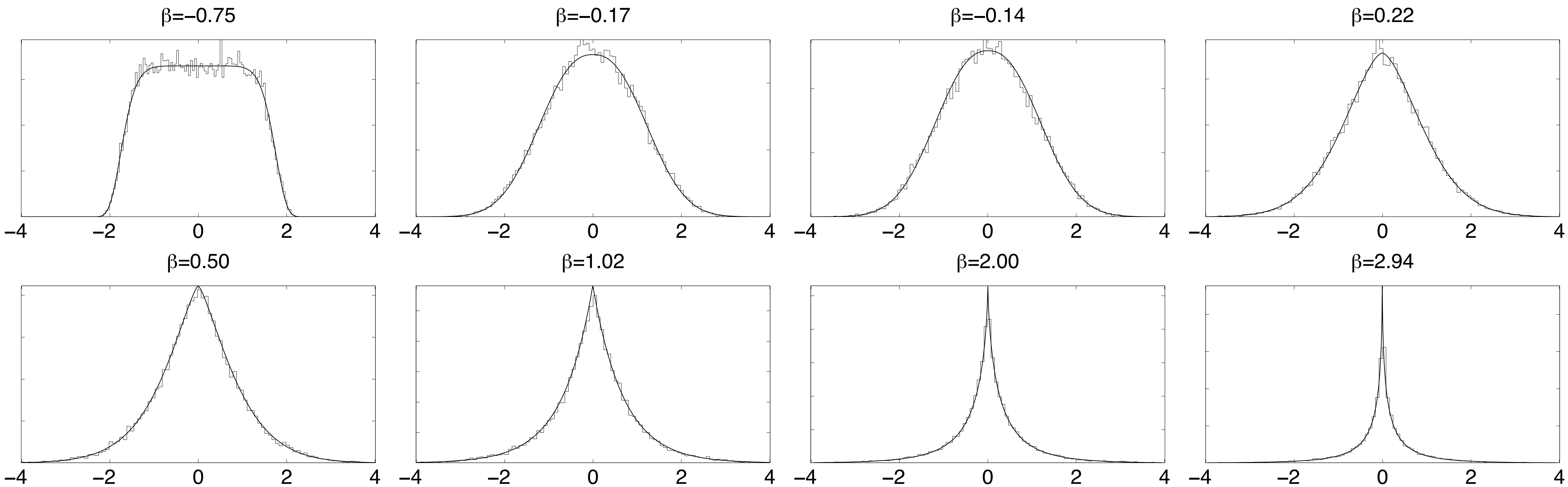
Separating mixtures of real sources



Separating mixtures of real sources



Separating mixtures of synthetic sources: 4 sources, 4 mixtures



source	1	2	3	4	5	6	7	8
avg. SNR (dB)	40.48	4.84	4.71	17.29	35.03	43.07	45.85	44.10
std. dev.	1.35	0.39	0.42	2.08	1.49	0.50	1.75	2.00

- Experiment: recover synthetic sources from a random mixing matrix. Repeat 5 times.
- SNR reflects the accuracy of in inferring the mixing matrix.
- The near-Gaussian sources cannot be accurately recovered.

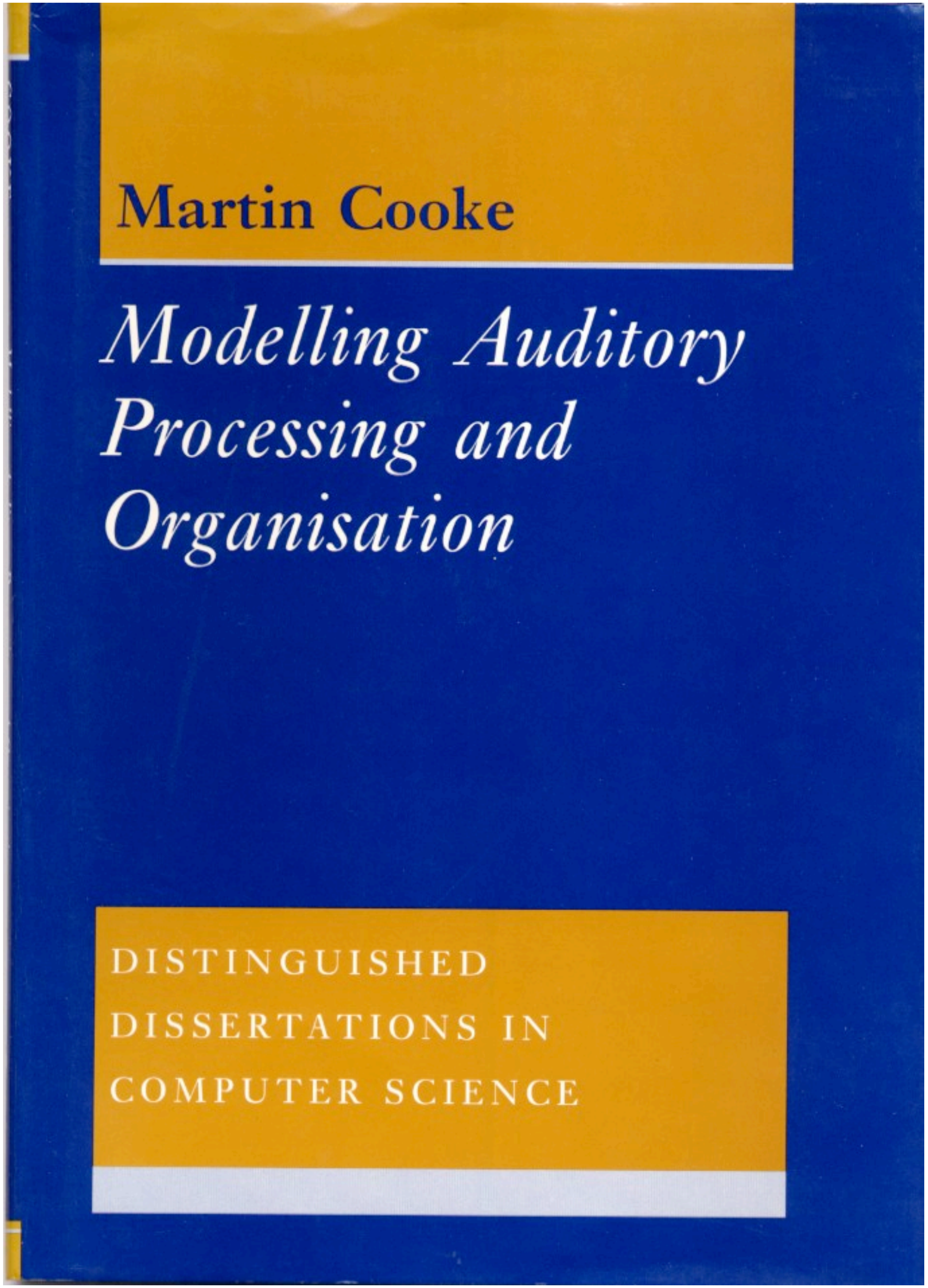
Computational auditory scene analysis

How do we incorporate these ideas in a computational algorithm?

Blind source separation (ICA) only solves special case

- non-Gaussian sources
- linear, stationary mixtures
- equal number of sources and mixtures
- need at least two mixtures

What about approaches that use auditory grouping cues?


Martin Cooke
Modelling Auditory Processing and Organisation
DISTINGUISHED DISSERTATIONS IN COMPUTER SCIENCE

Cooke (1993)

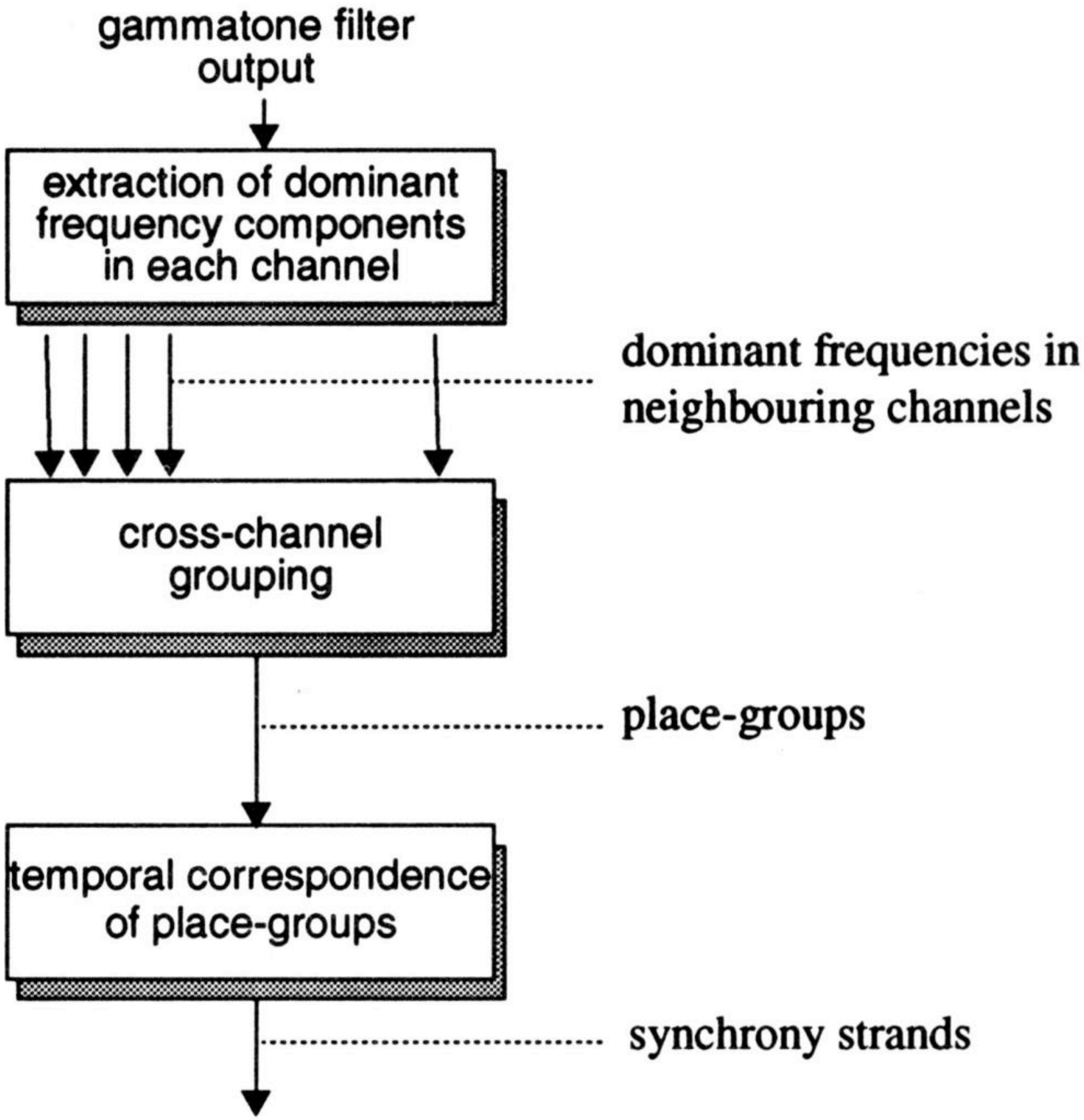
"Modeling Auditory Processing and Organisation"

- uses gammatone filter bank to model auditory periphery
- auditory "objects" are represented by "synchrony strands"
- synchrony strands are formed according to auditory grouping principles

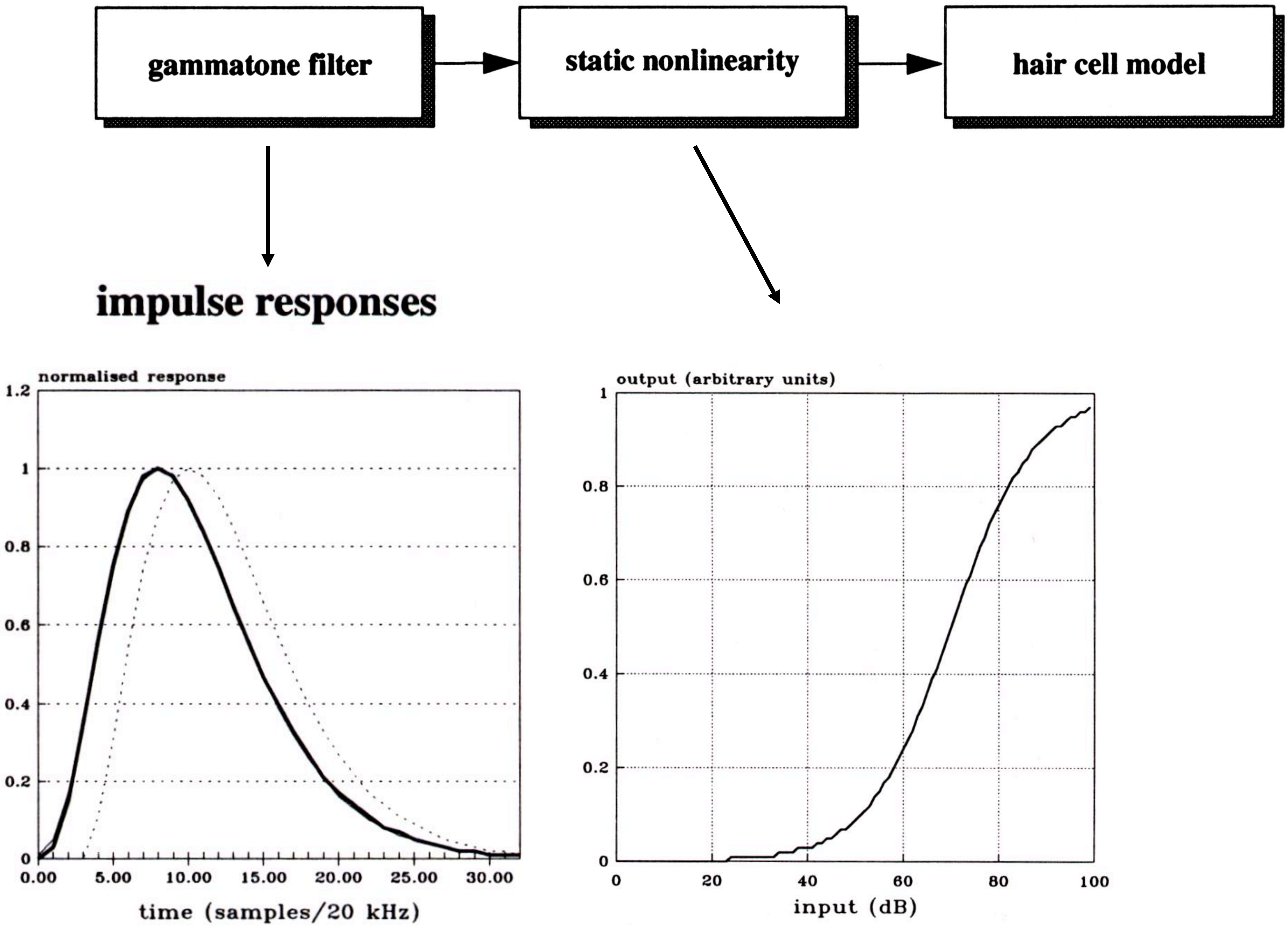
Want an auditory representation that

- describes the auditory scene in terms of time-frequency objects
- characterizes onsets, offsets, and movement of spectral components
- allows for further parameters to be calculated

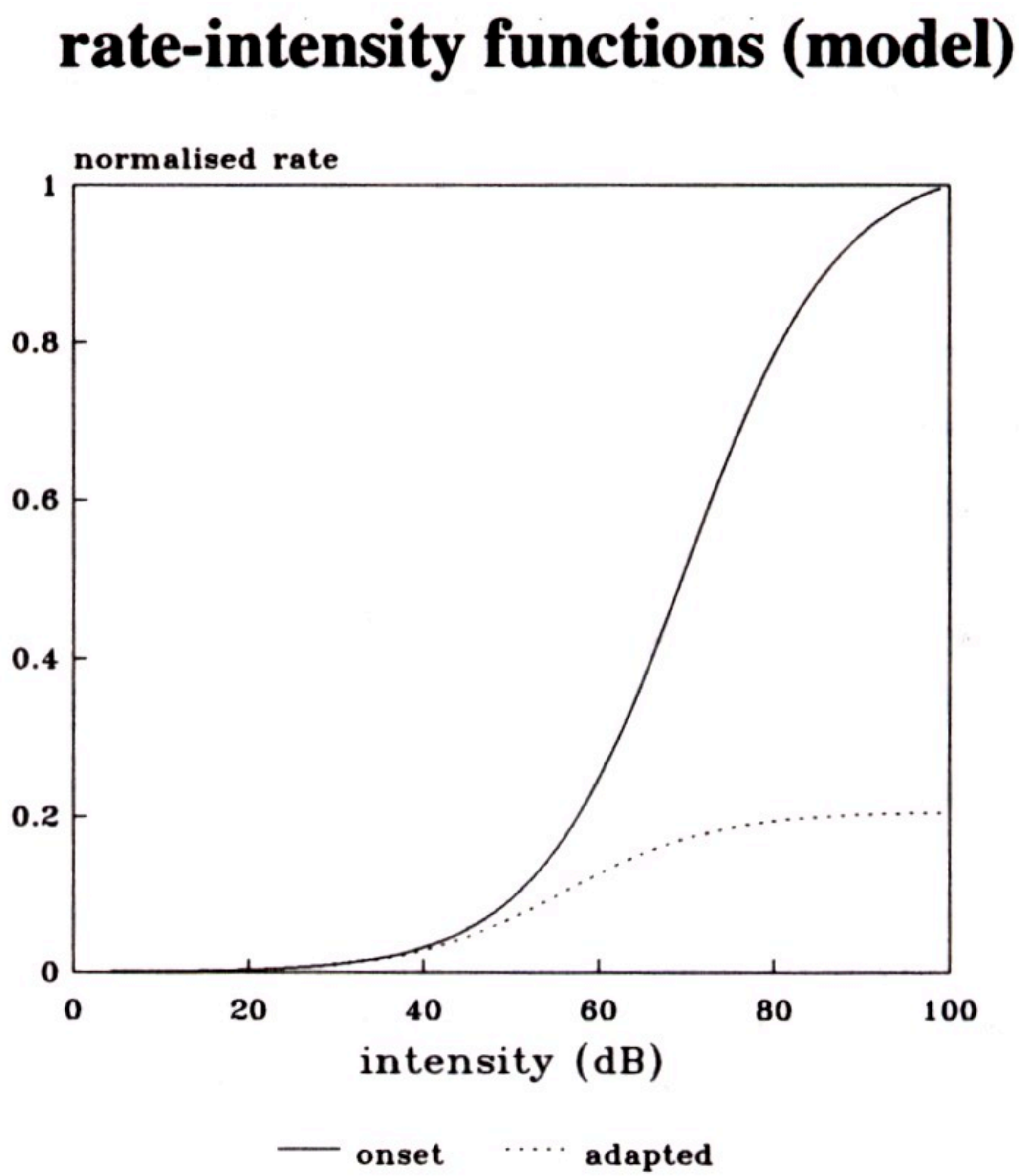
Overview of stages in synchrony strand formation

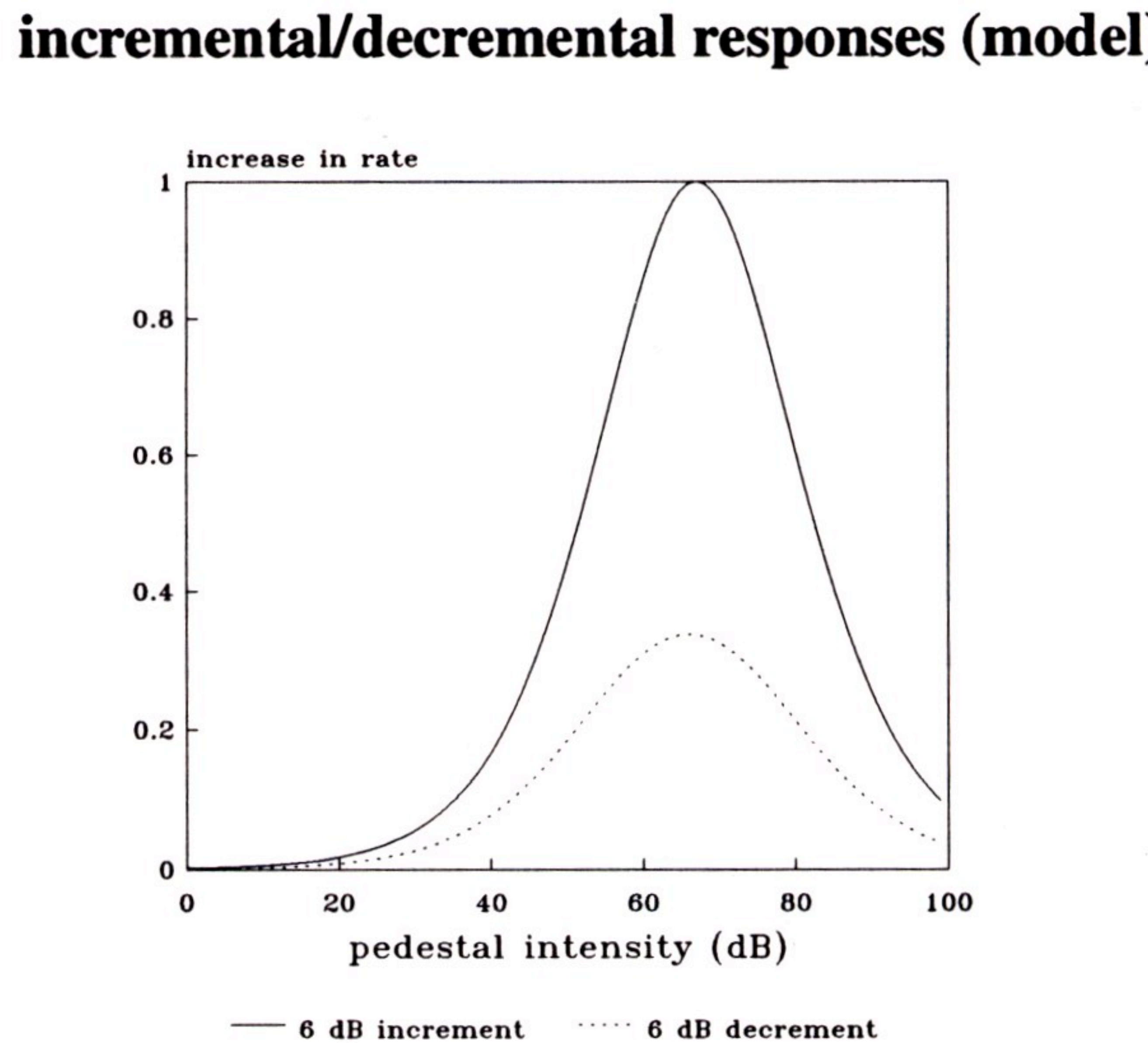


Auditory periphery model



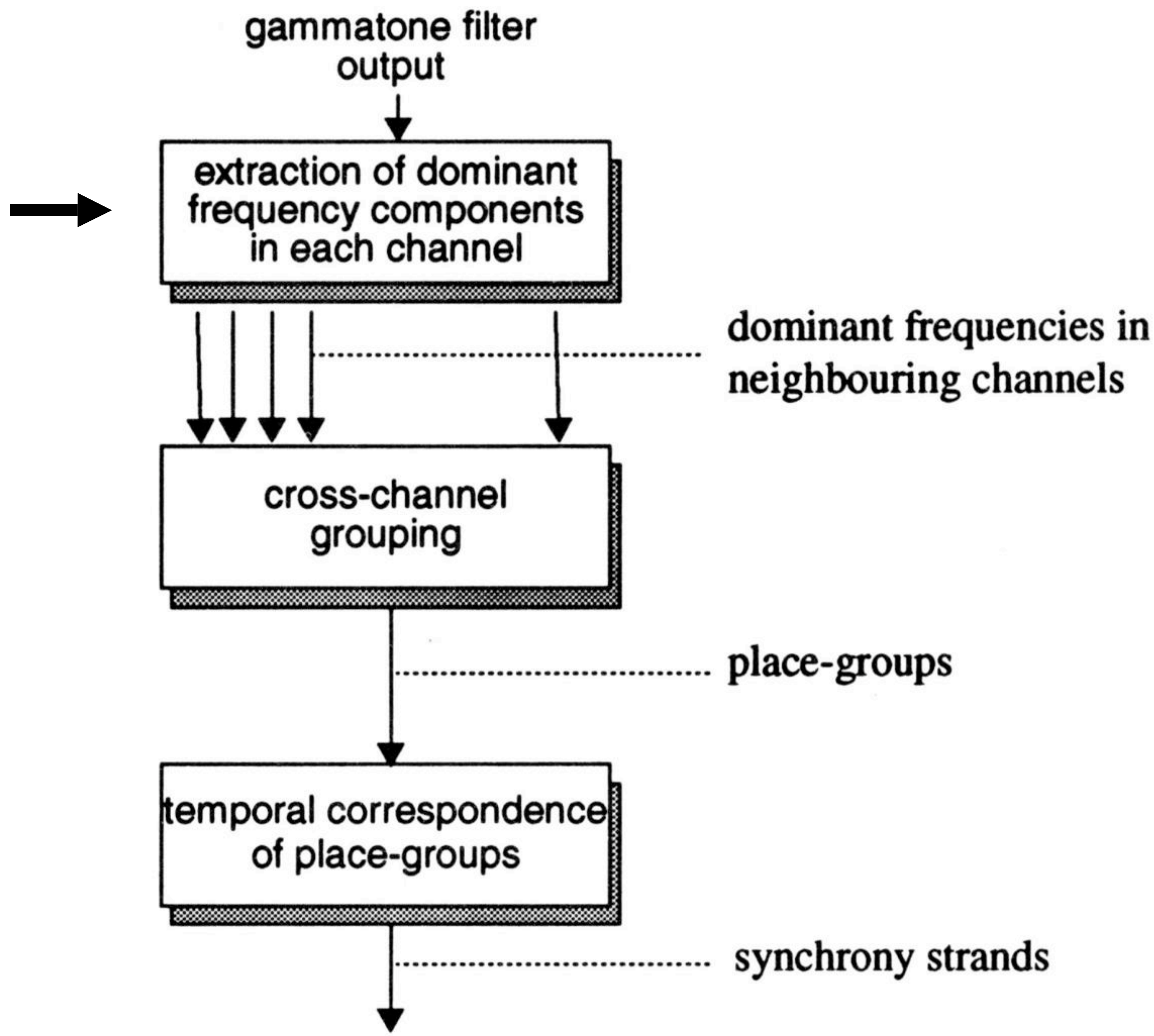
Hair cell adaptation





Hair cell model is based on adaptive non-linearities of cochlear responses

Next stage: extraction of dominant frequency



Dominant frequency estimation

Want a pure frequency representation in order to apply grouping rules.

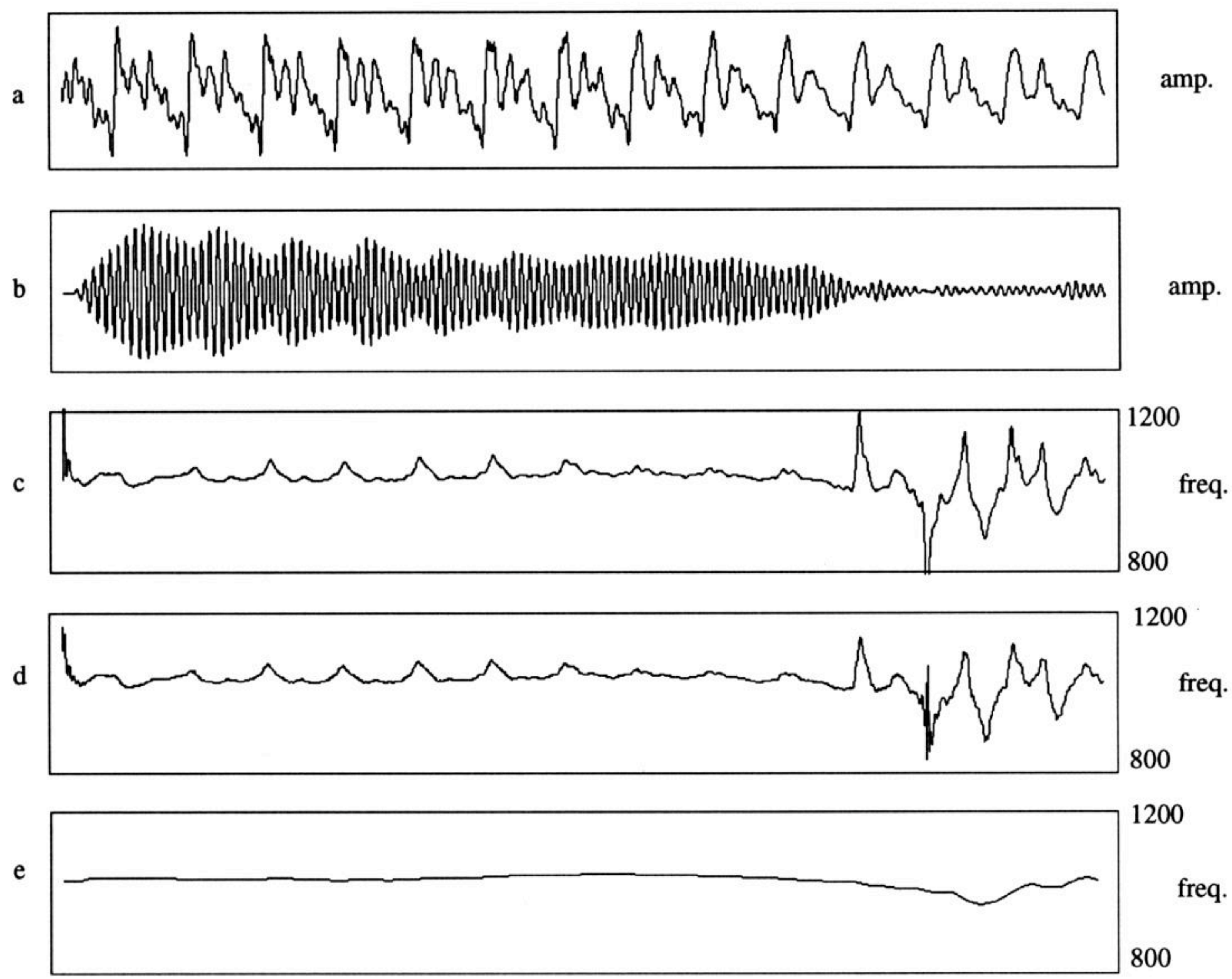


Figure 3.2 Dominant frequency estimation: *a*: waveform for the vowel /ae/; *b*: auditory filter output (CF: 1 kHz); *c*: $\nu(t)$ by analytic signal method; *d*: instantaneous frequency by linear prediction analysis; *e*: $\hat{\nu}(t)$.

a) waveform for the vowel /ae/

b) auditory filter output for 1kHz center frequency

c) instantaneous freqency

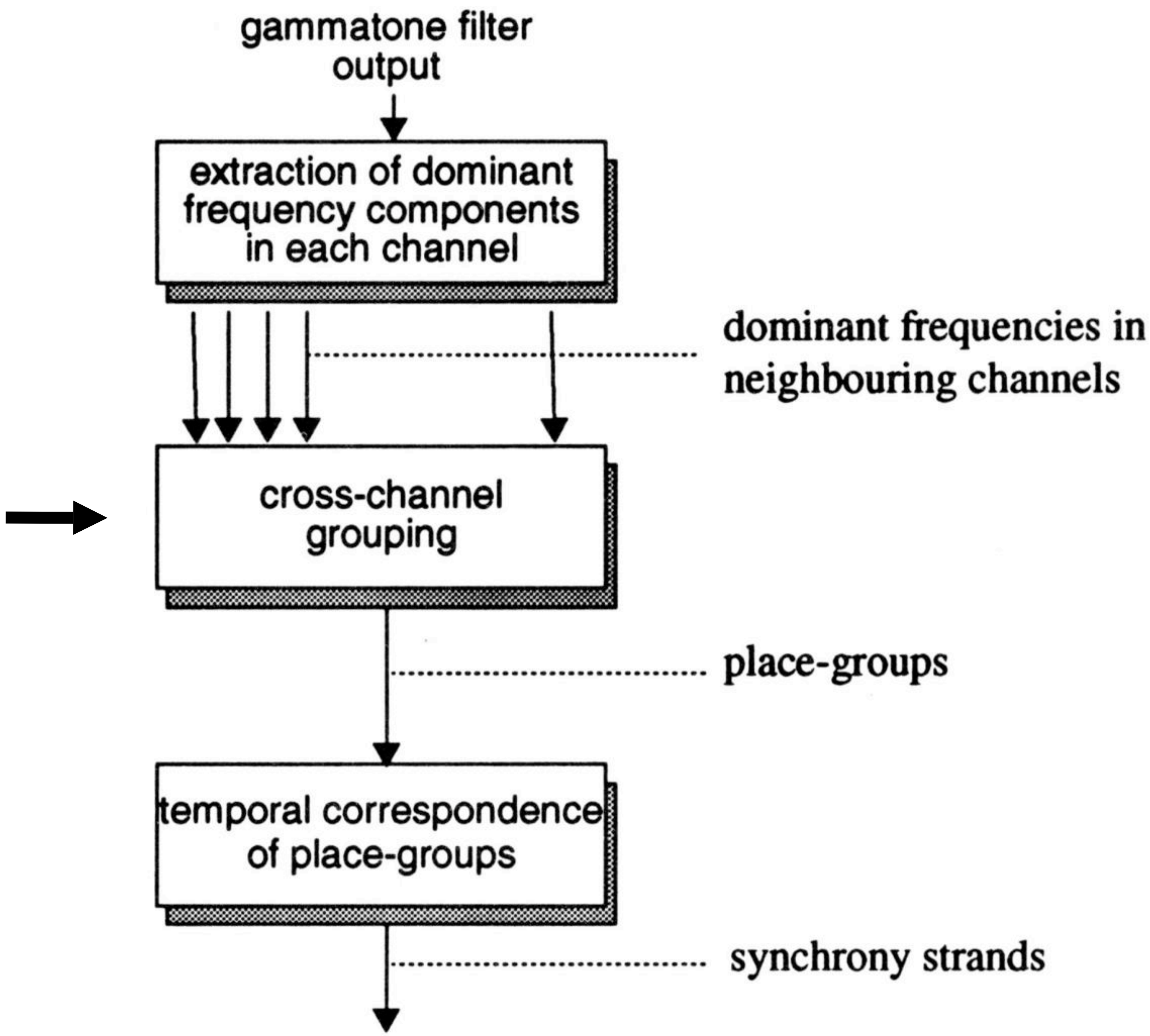
- from Hilbert transform of gammatone filter

d) an alternative method of estimating instantaneous frequency

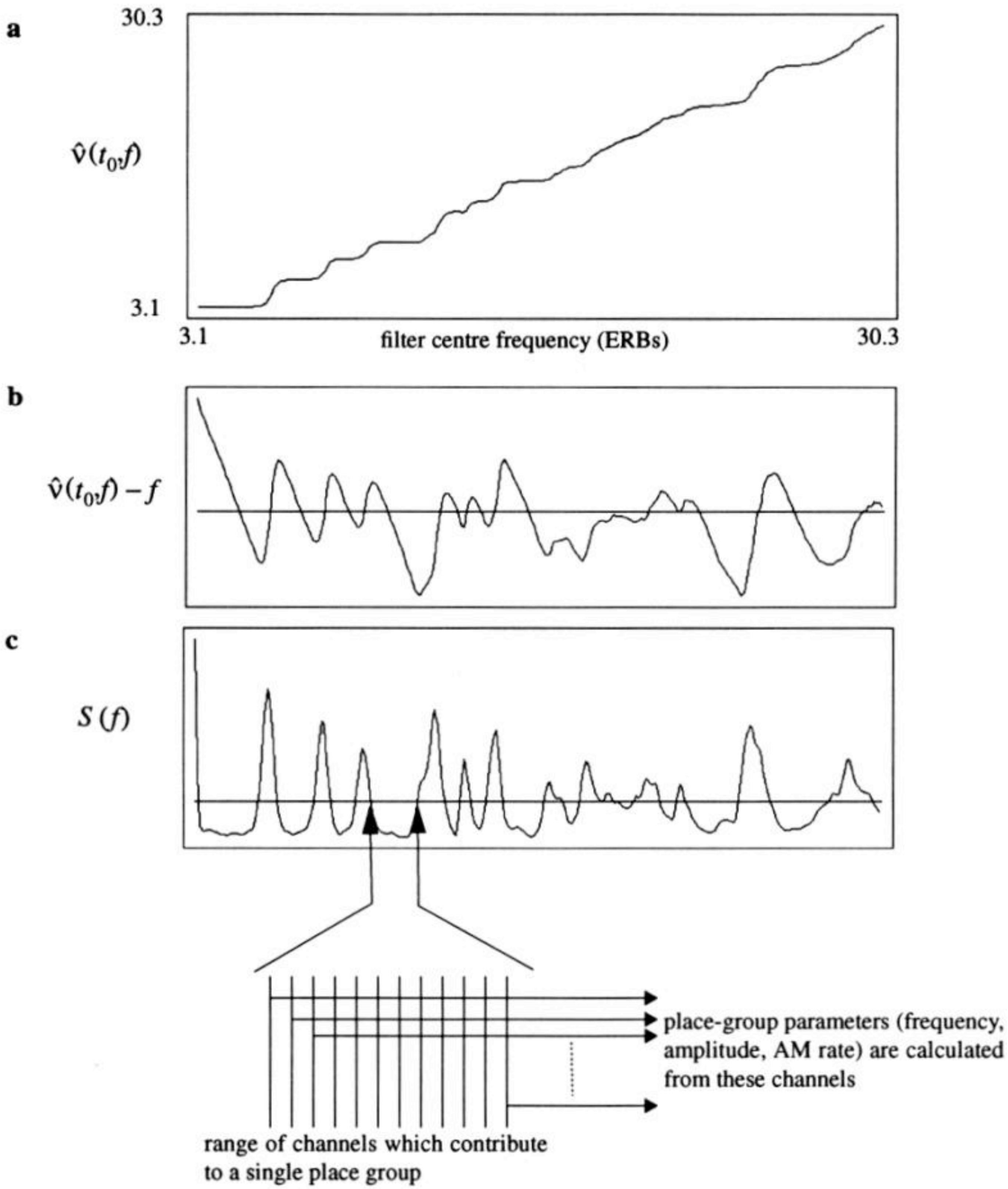
e) smoothed instantaneous frequency

- median filtering over 10ms, plus linear smoothing
- also allows for data reduction: samples are collapsed over 1ms window

Next stage: grouping frequency channels



Calculation of place groups



- What frequencies belong to the same sound?
- Goal of this stage is to locate and characterize intervals along filterbank with synchronous activity.
- (a) Notice that center frequency is not distributed evenly due the median filtering in the frequency estimation stage.
- (b) The instantaneous frequency varies around the estimated frequency.
- Idea is to group all channels centered around the "dominant frequency", i.e. grouping by spectral location.
- (c) $S(f)$ is a smoothed frequency derivative estimate. Channels at the minimum are grouped together.

Plot of place groups for a speech waveform

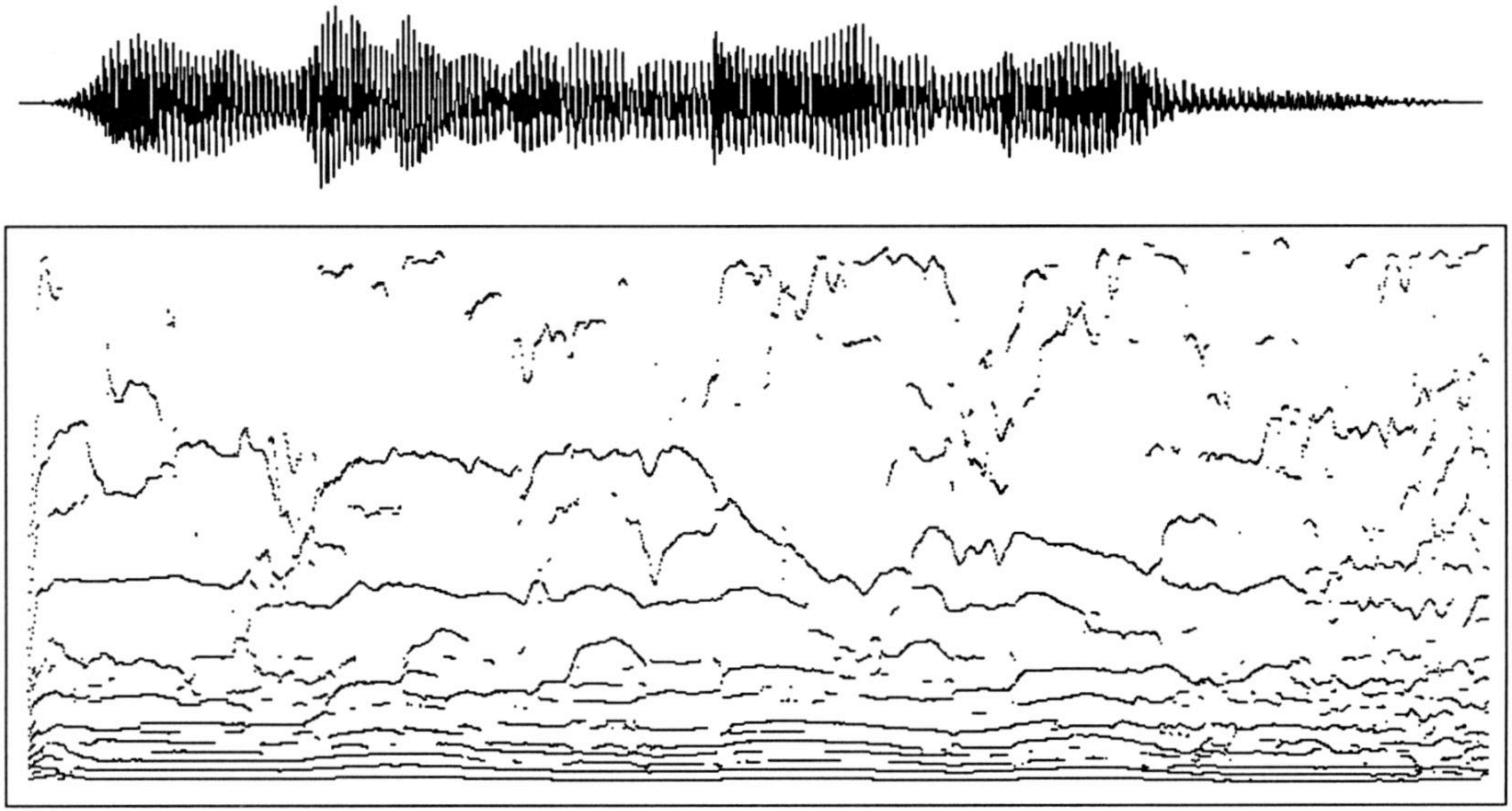


Figure 3.4 Place-groups for the utterance whose waveform is shown. Frequency axis is linear in Hz.

Waveform resynthesized from place groups

Resynthesize waveform by summing sines in place group

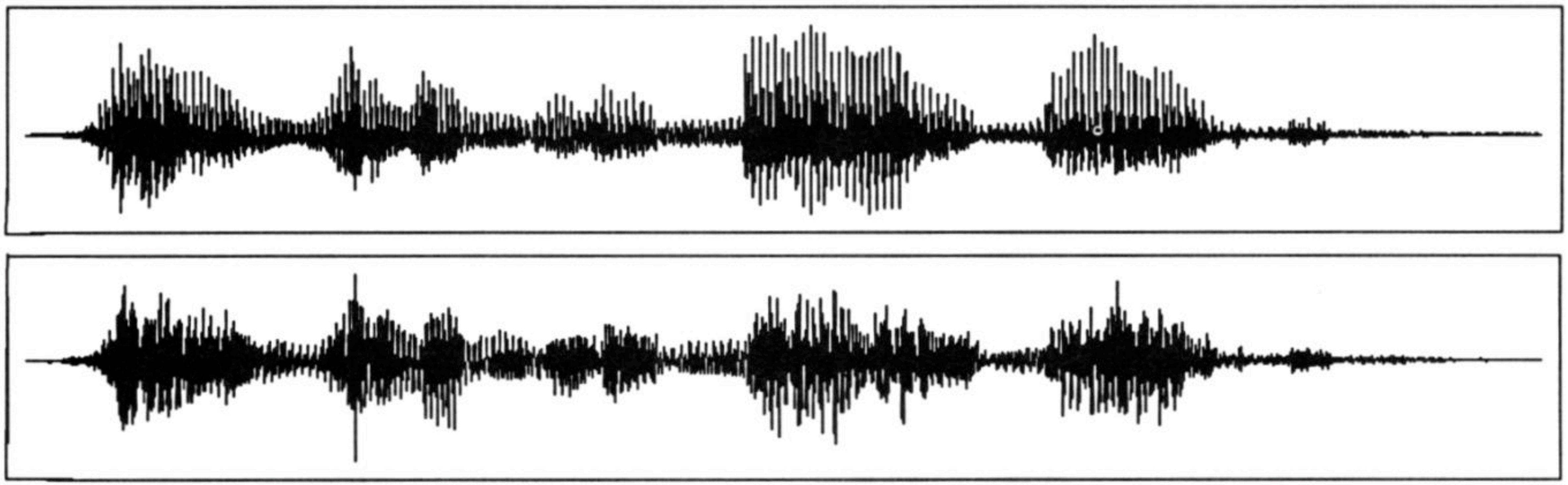


Figure 3.5 *Top*: Original waveform for the utterance "I'll willingly marry Marilyn". *Bottom*: Resynthesised waveform.

Reconstruction quality is good

- best is indistinguishable from original
- worst is still clearly intelligible
- best for harmonic sounds

More synchrony strand displays: female speech

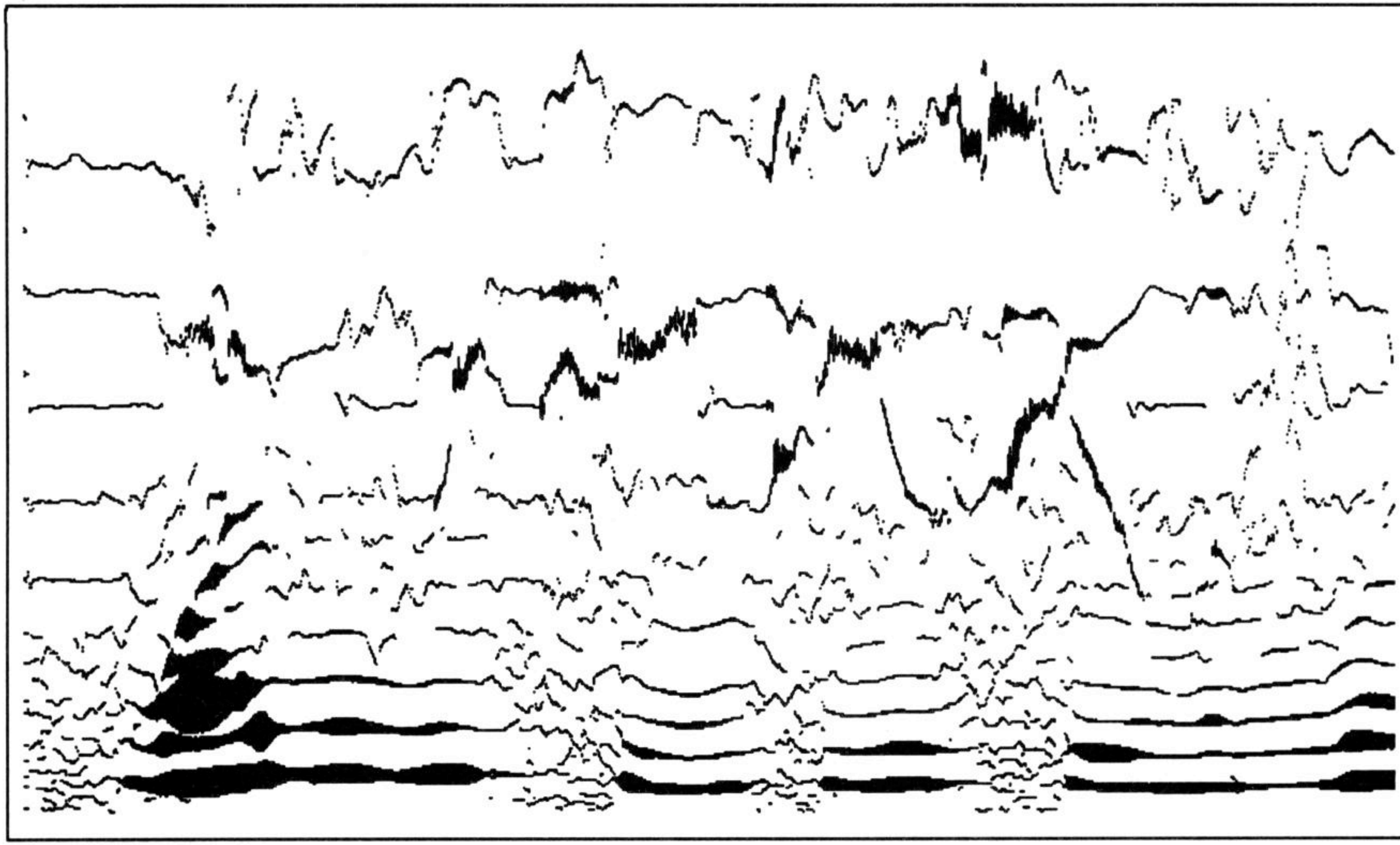


Figure 3.6 Female speech (1.83 s duration).

Synchrony strand display of male speech

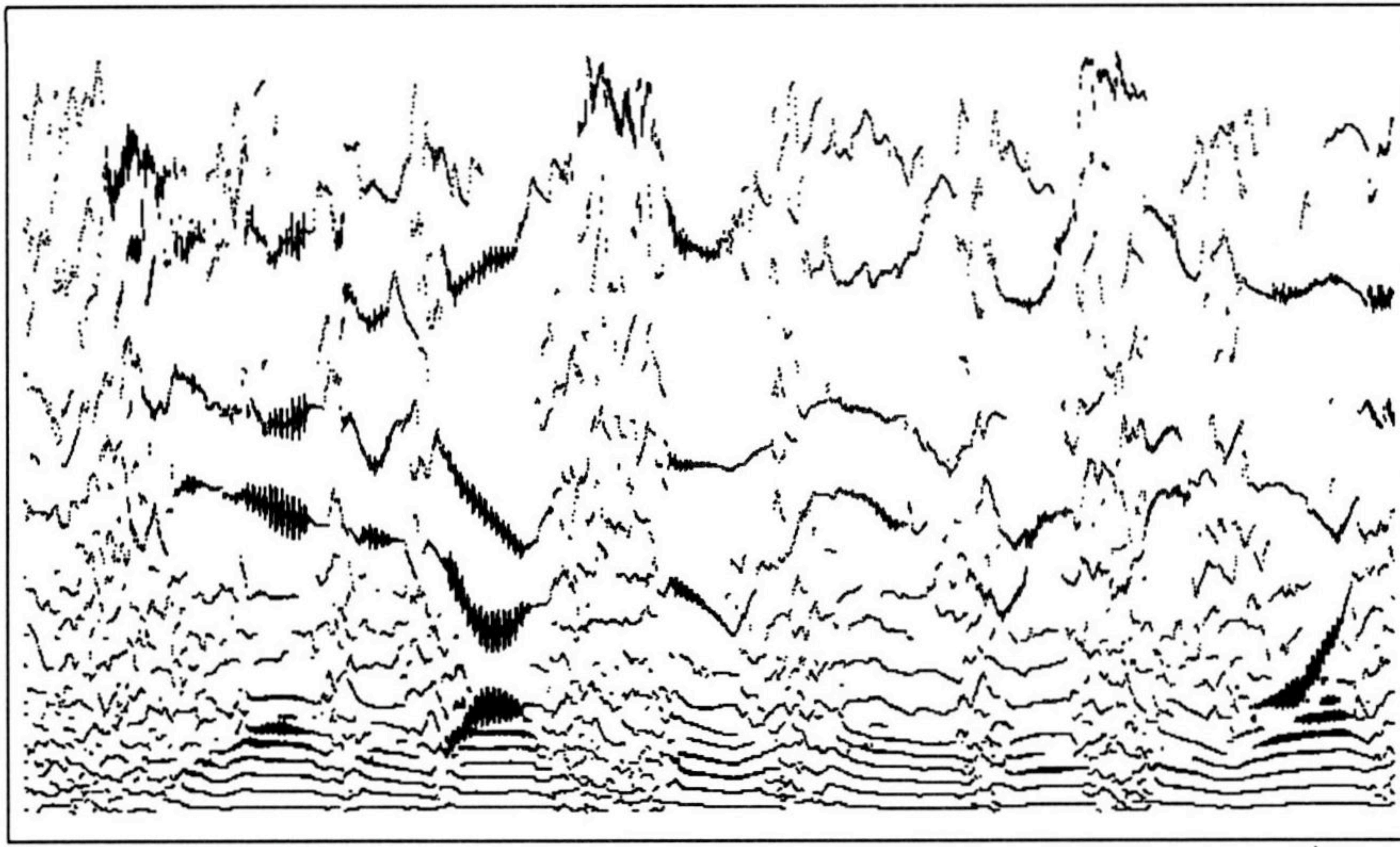


Figure 3.7 Male speech (2.5 s duration).

Synchrony strand display of noise burst sequence

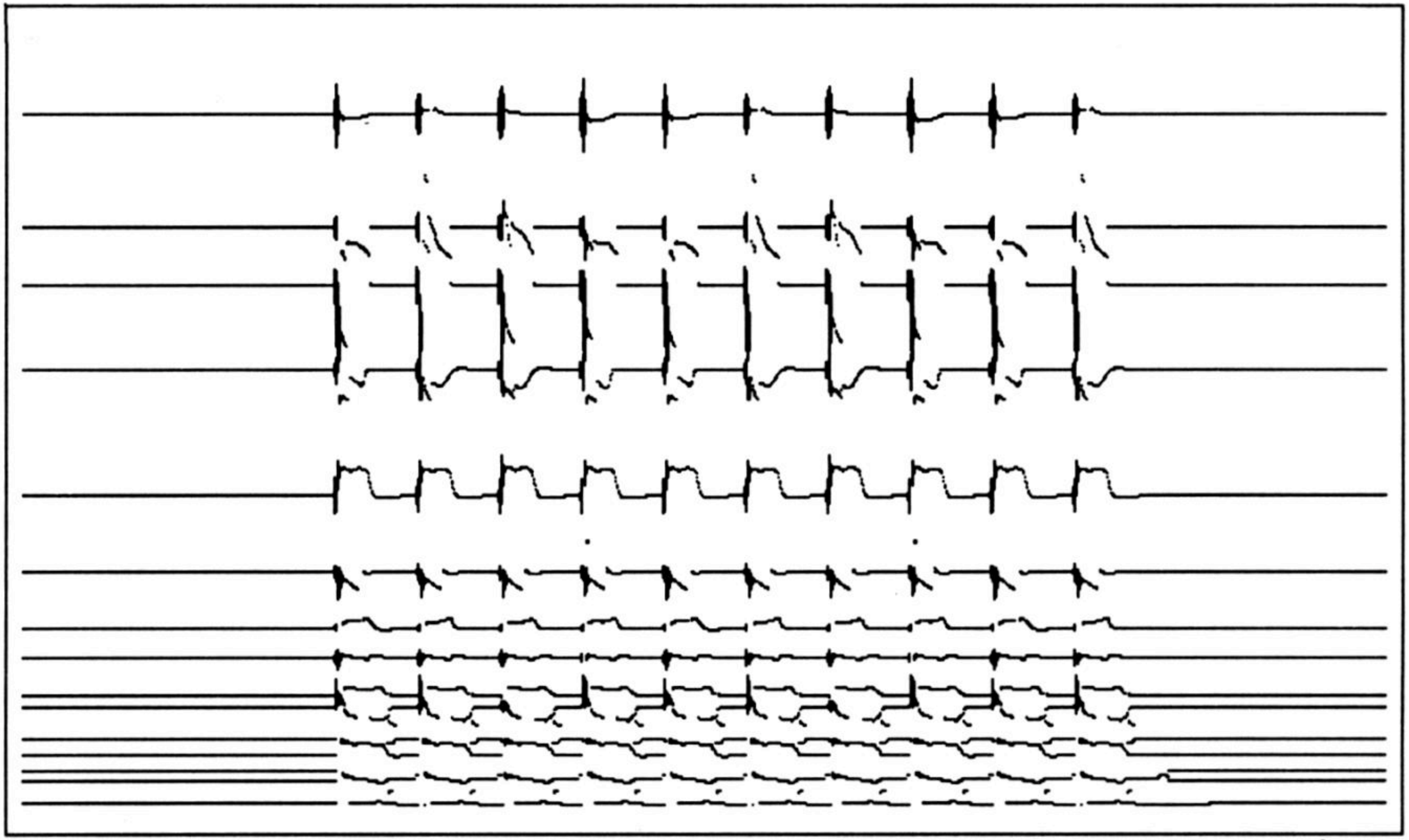


Figure 3.8 Noise burst sequence (1.76 s duration).

Synchrony strand display of new wave music

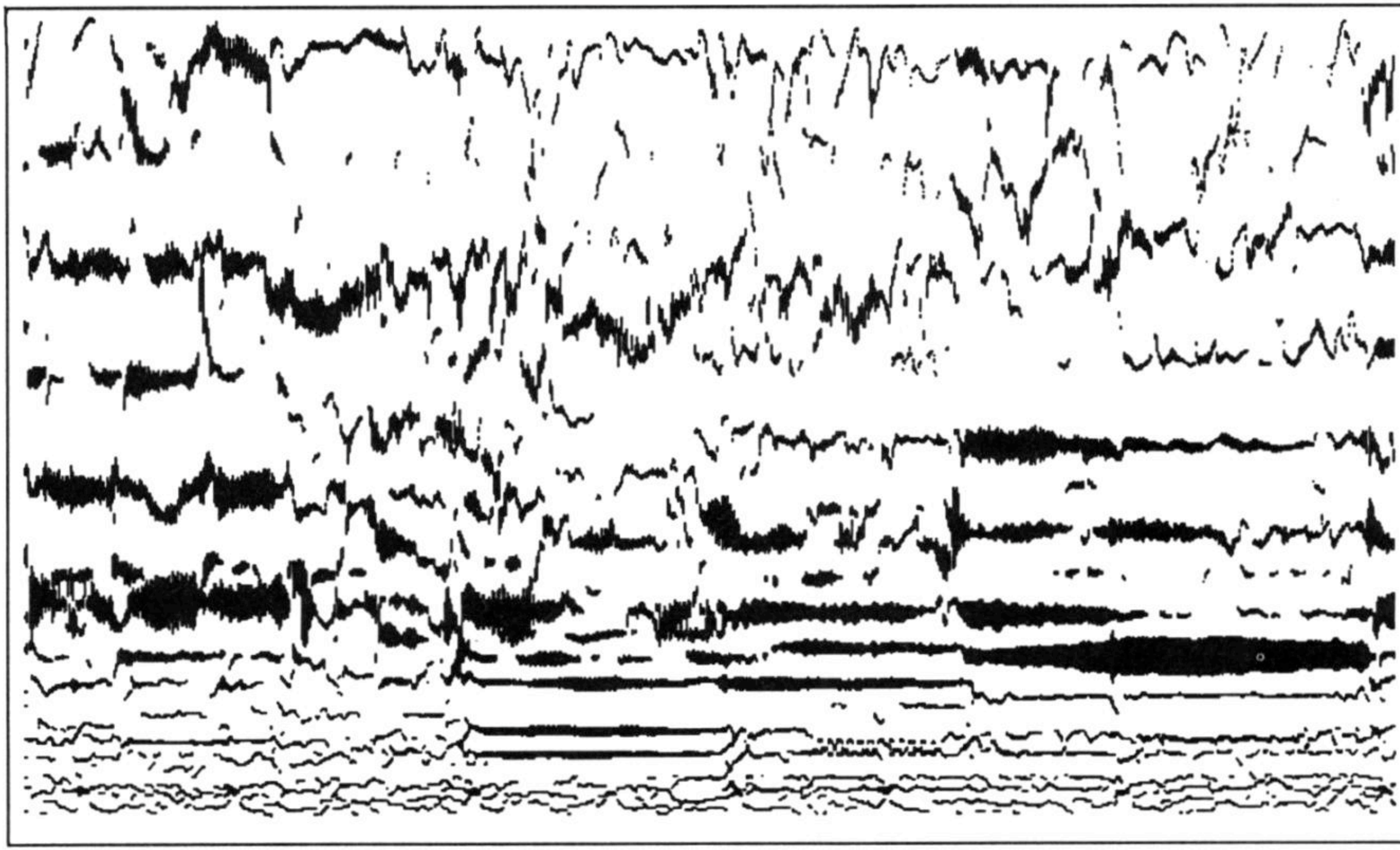


Figure 3.9 New wave music (2 s duration).

Modeling auditory scene exploration

- The synchrony strand display has done local grouping of frequencies, but this alone does not provide a way to group sounds from the same source.
- The next part of the model develops ways to construct higher order groups

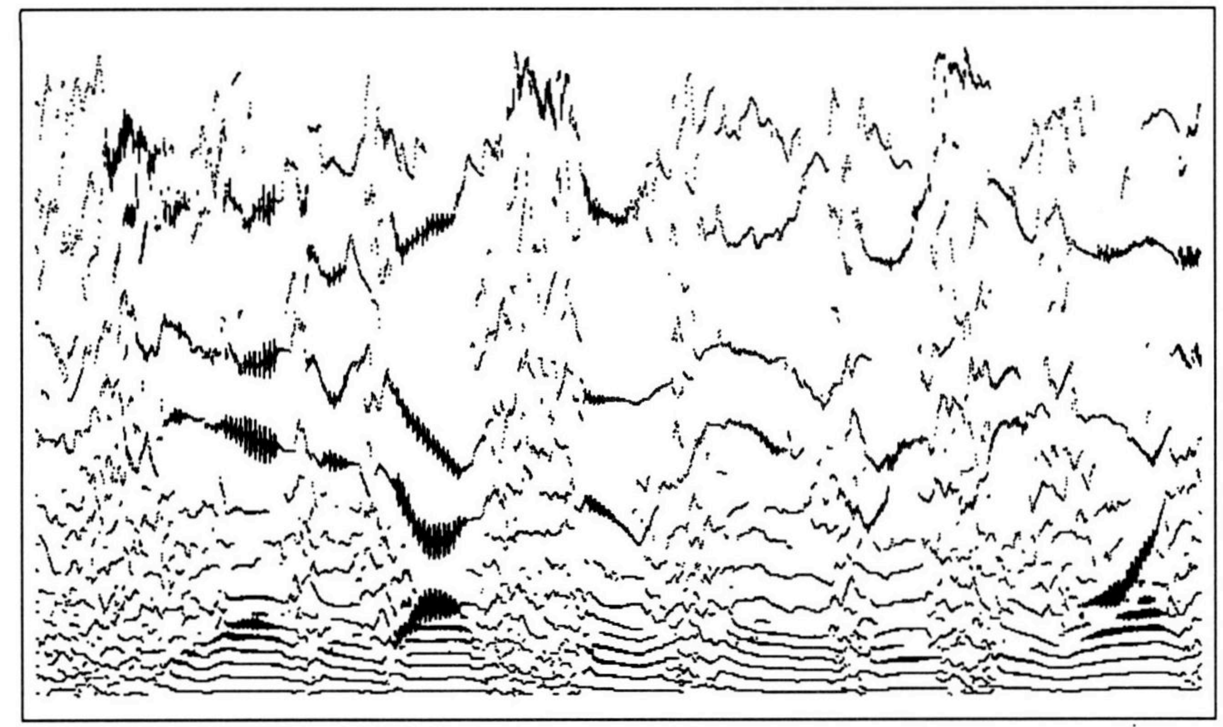


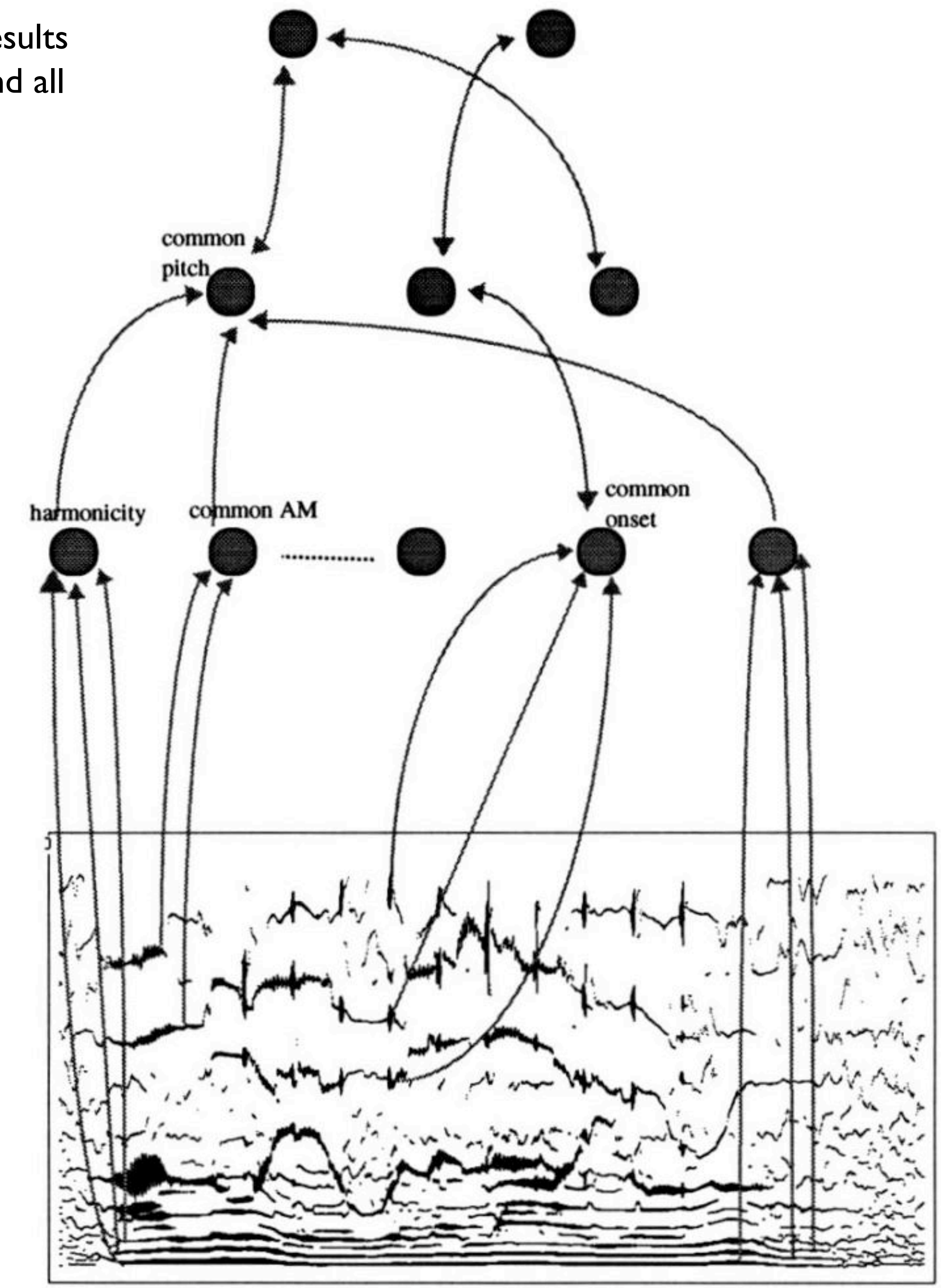
Figure 3.7 Male speech (2.5 s duration).

A hierarchical view of auditory scene analysis

Top level: the stable interpretation which results when all organization has been discovered and all competitions resolved.

Intermediate level: results from conbining groups with similar derived properties such as pitch and contour

First level: Results from the independent application of organizing principles such as harmonicity, common onset etc.



But it's not as simple as assigning strands to different sounds

Duplex perception (Lieberman, 1982)

- Stimulus: a synthetic syllable with 3 formants
- Syllable minus the third formant transition is played to one ear
- Missing formant transition is played to the other ear

What do subjects hear?

- There hear the syllable as if all formants had been presented to one ear.
- But, they also hear an isolated chirp

This suggests that components are shared across groups

Cooke's framework for auditory scene exploration

1. Seed selection (thick line)

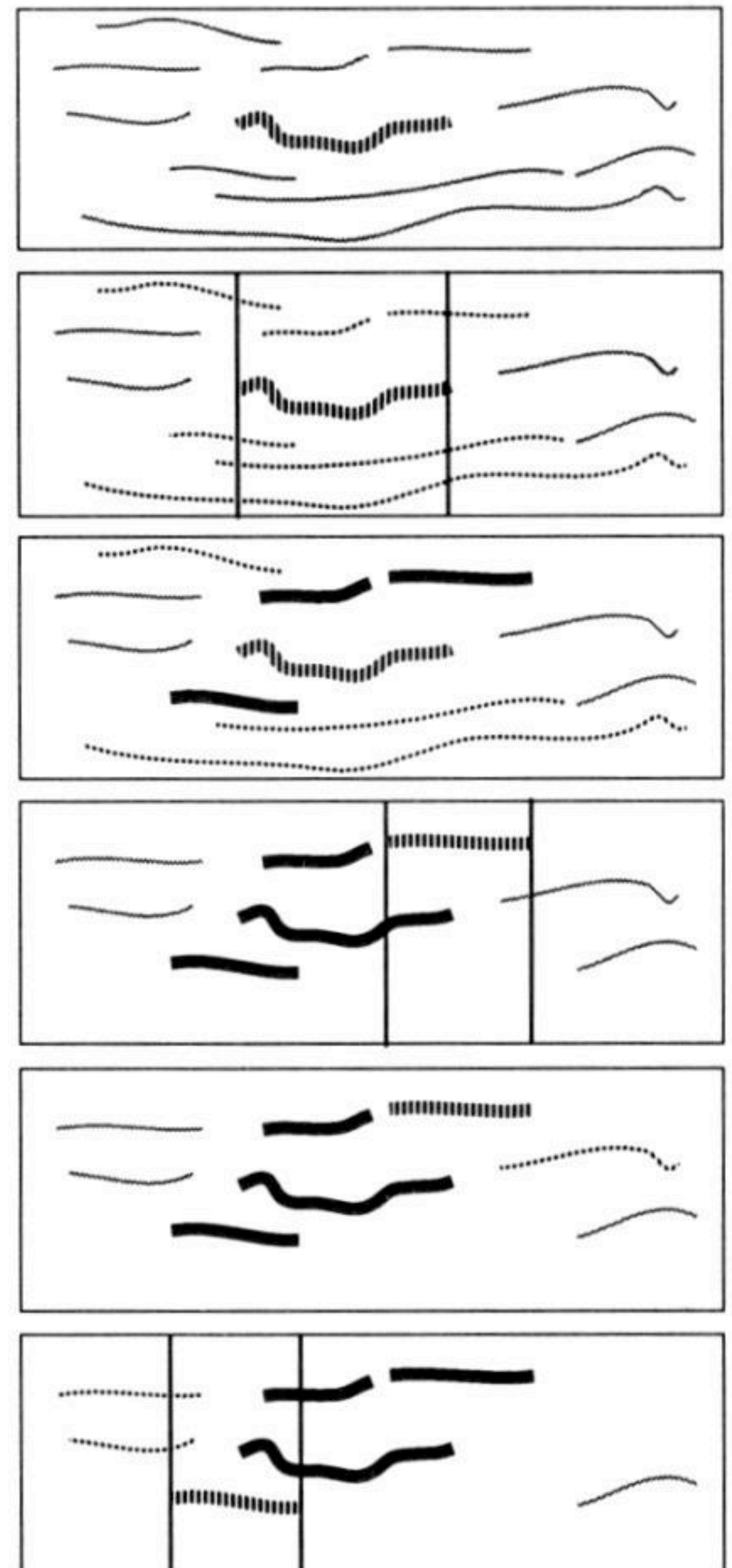
2. Choose 'simultaneous set' i.e. those strands which overlap in time with the seed

3. Apply grouping constraint, e.g. suppose the black strands share a common rate of amplitude modulation with the seed

4. Sequential phase: choose strand to extend the temporal basis of the group

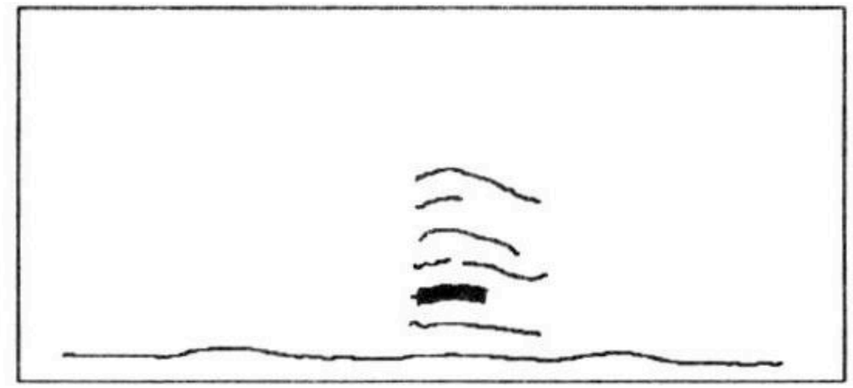
5. Back to simultaneous phase: consider for similarity strands that overlap with new seed

6. Seeds may be selected from any in the group which extended in time

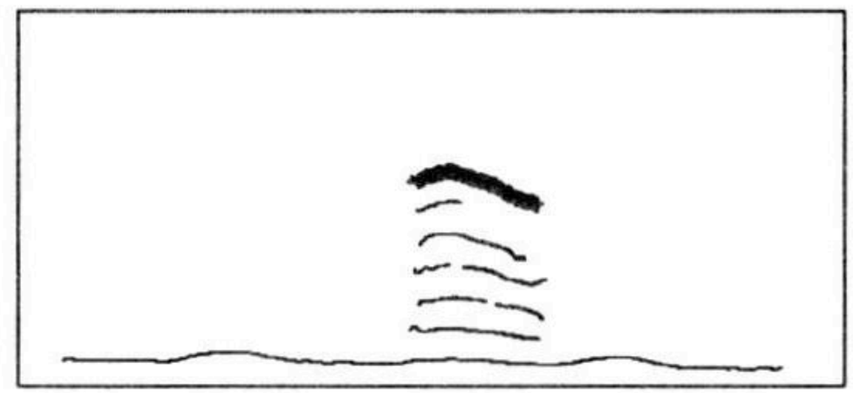


Harmonic constraint propagation for a voiced utterance

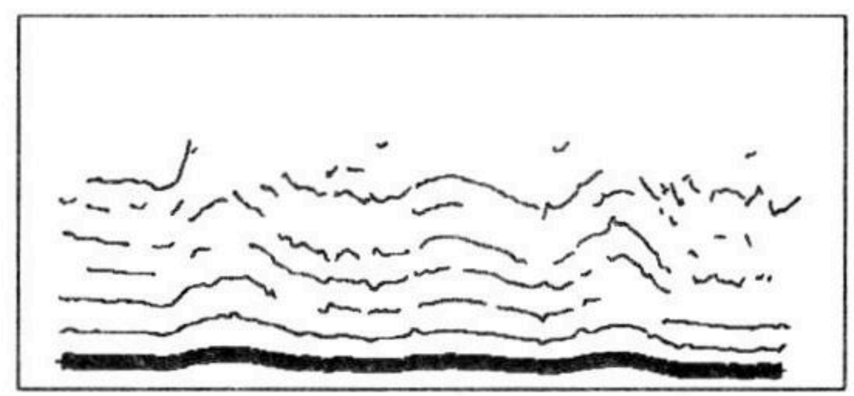
Seed (highlighted) attracts several supporters to form a harmonic group.



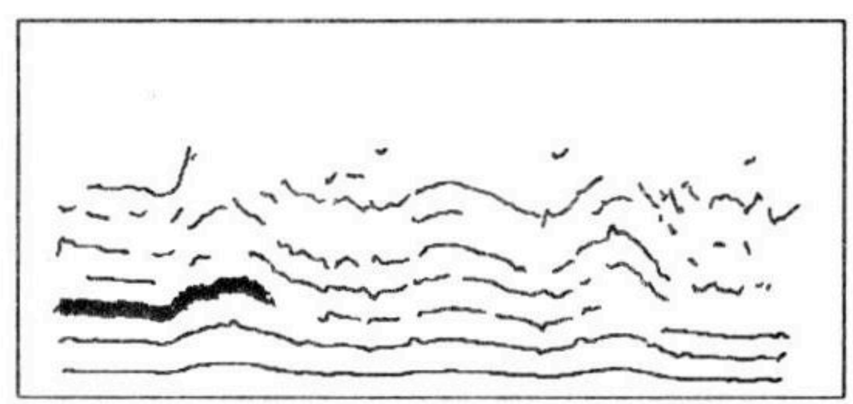
A new focus is chosen, but recruits few new supporters to group.



New focus (f0 itself) successfully attracts virtually all the harmonically related strands in the utterance.



Process halts when no temporal extension to the group is possible.



Implementation of auditory grouping principles

Cooke’s algorithm implements the following grouping principles:

- harmonicity
- common amplitude modulation
- common frequency movement

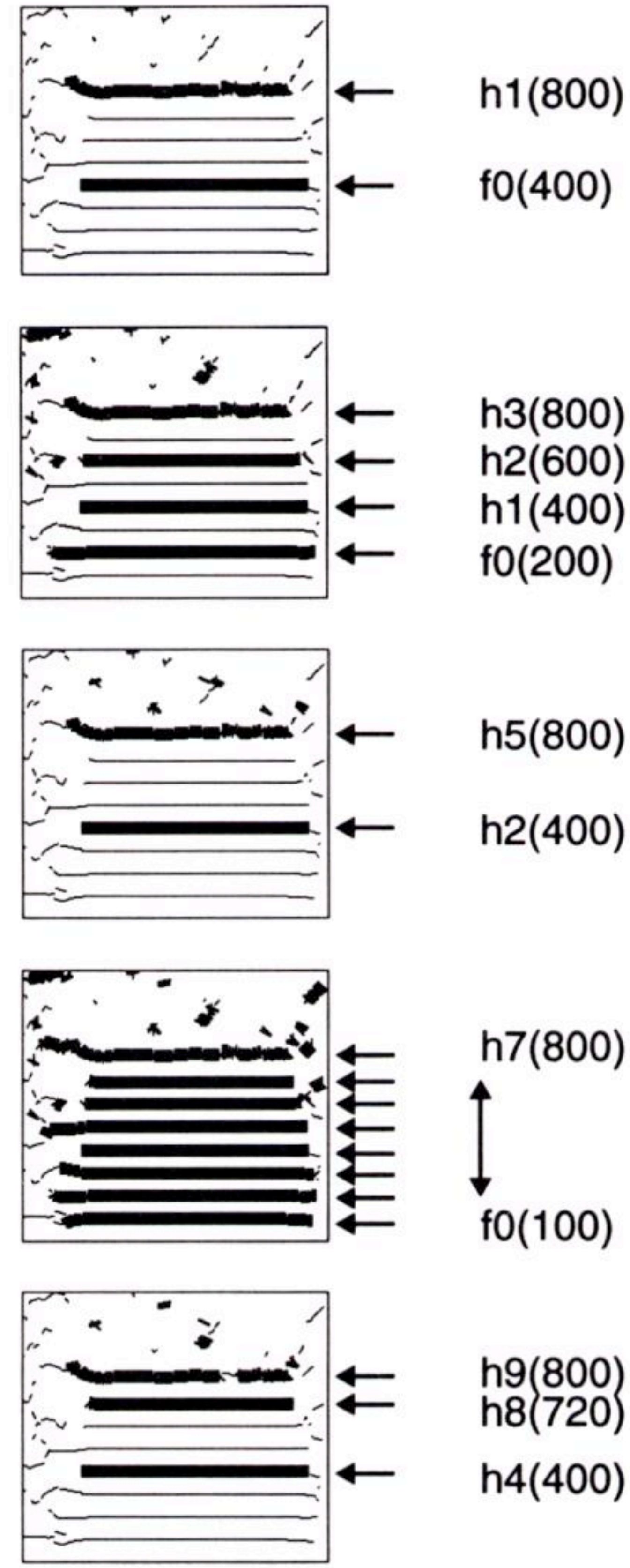
Also need “subsumption” to form higher level groups.

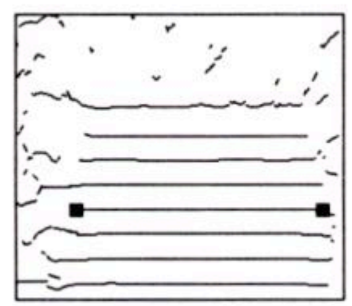
- Idea is to remove groups that are contained within a larger group
- Groups expand until at some point they are assigned to a source

Also note that the algorithm involves many heuristics which are not covered here.

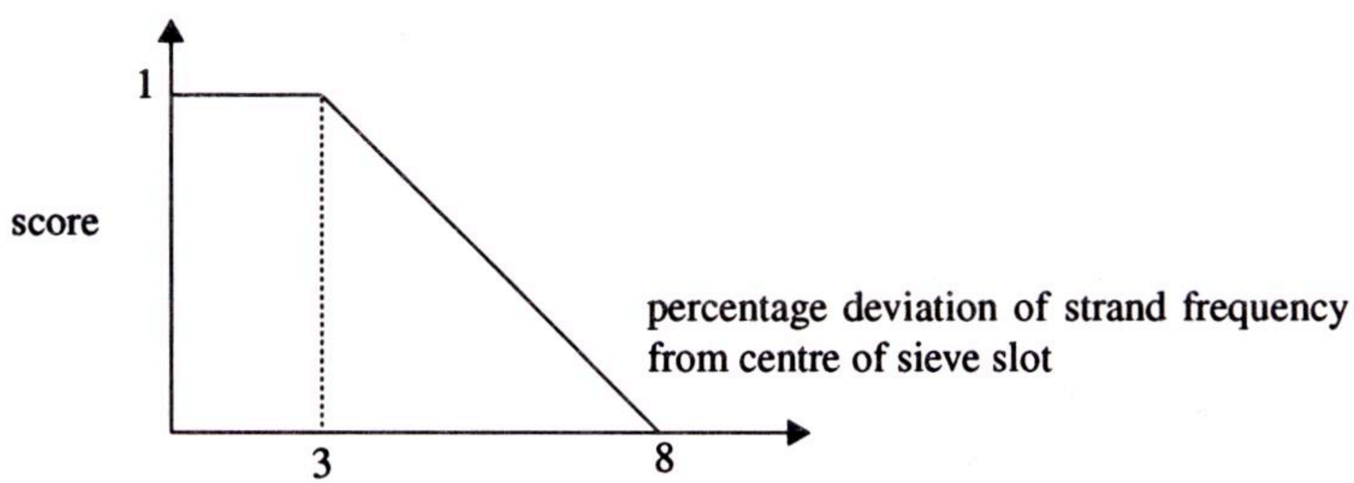
Harmonic grouping

Harmonic groups discovered by various synchrony strands





Scoring function used in assessing how well strands fit sieve slots



Harmonic grouping of frequency modulated sounds

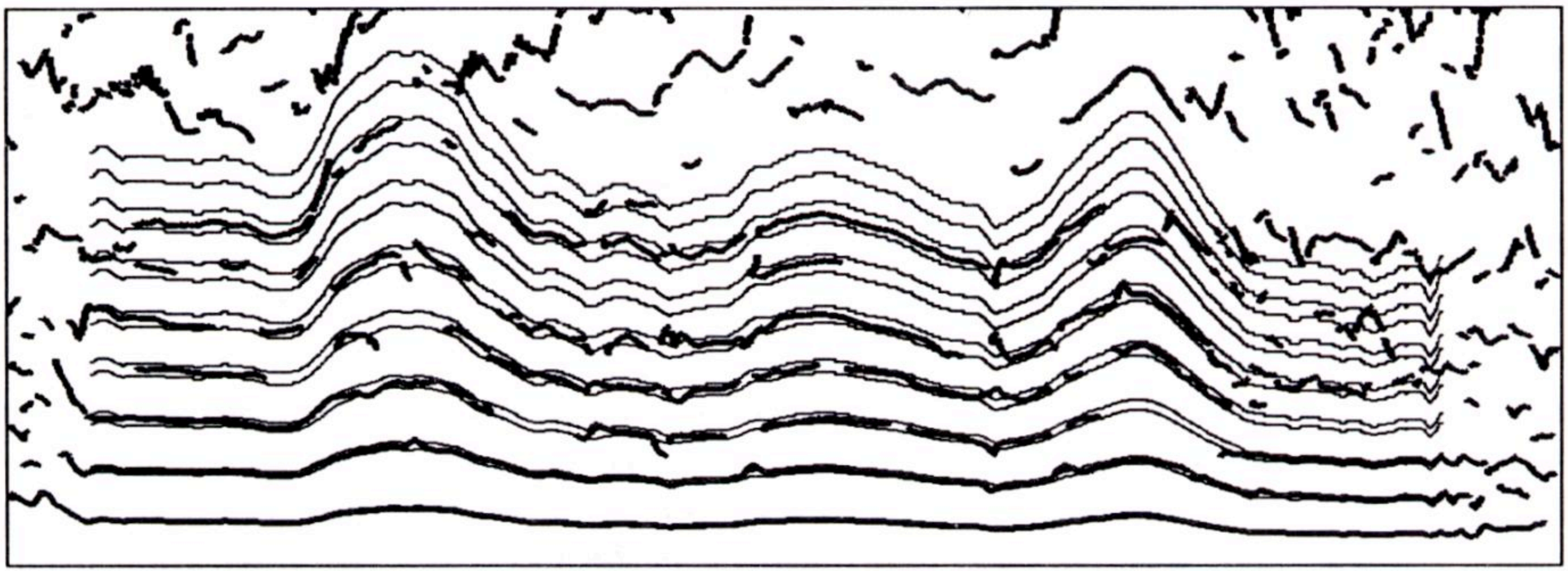
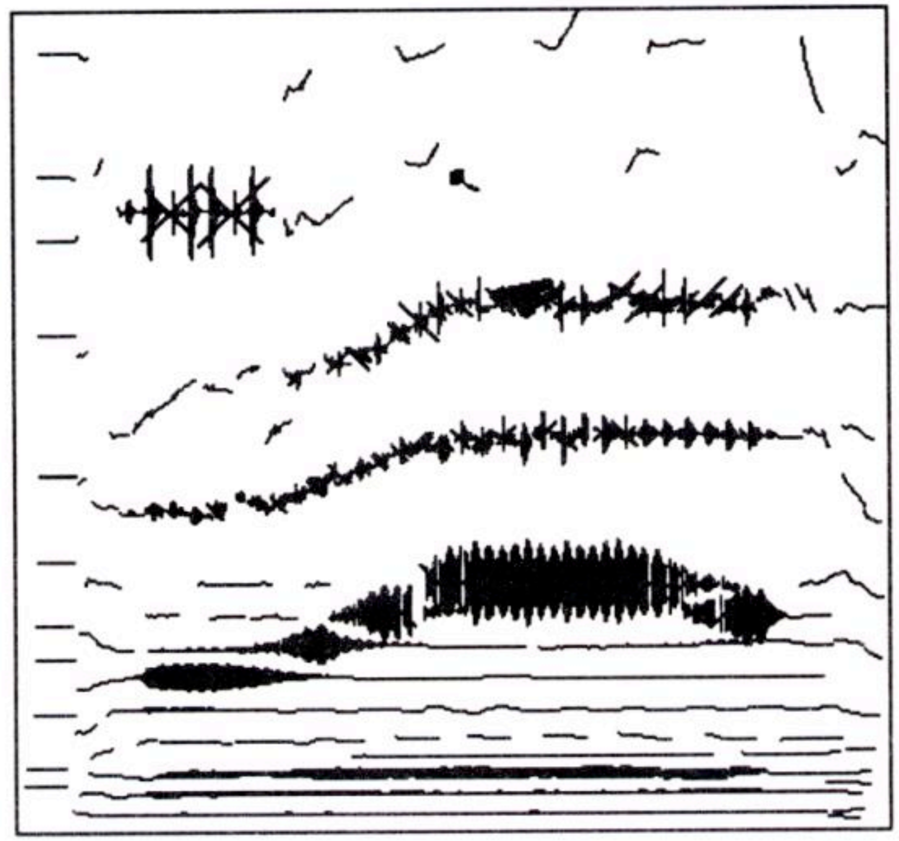
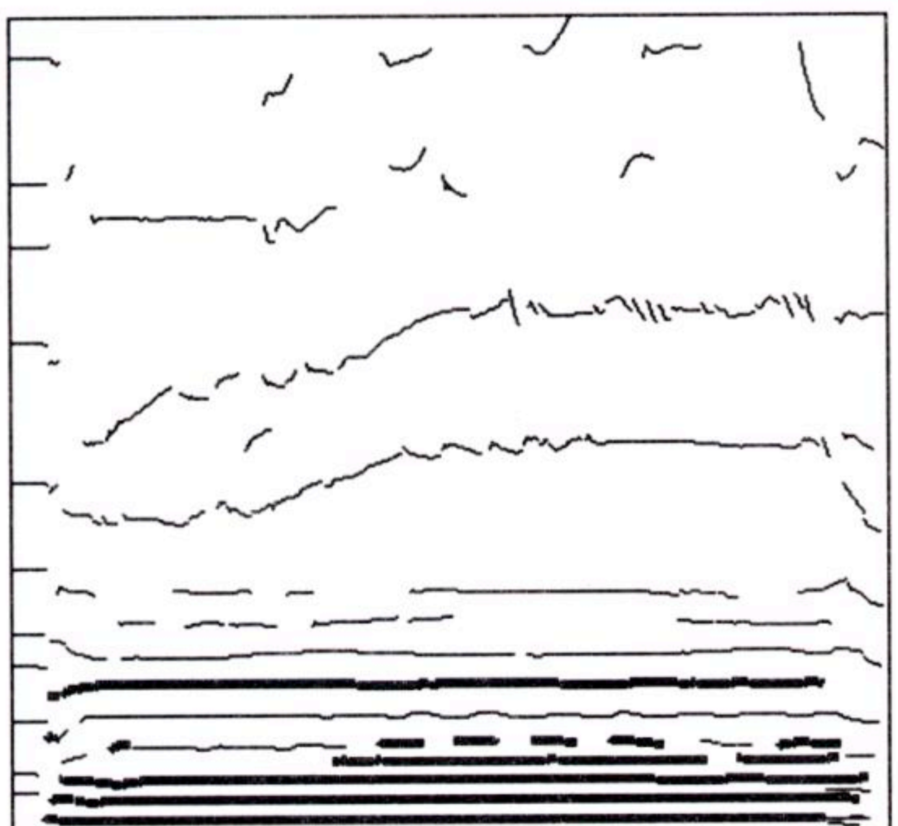
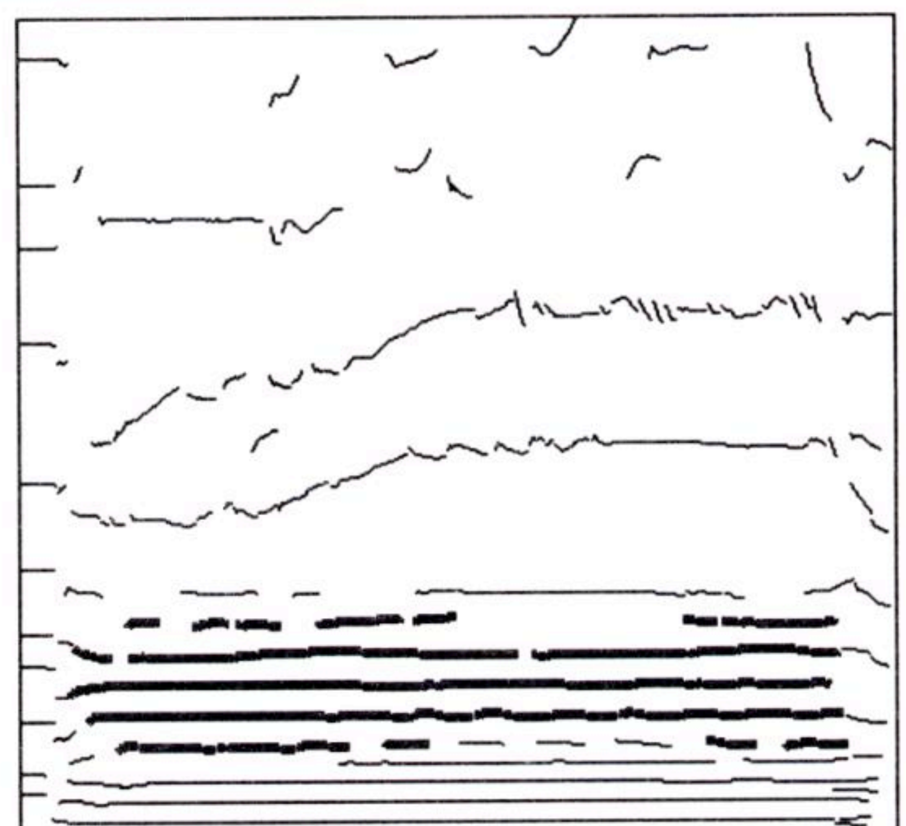


Figure 5.2 One of a series of temporally-extended harmonic sieves generated from a seed strand. Thin lines represent the +/-3% boundaries of sieve channels. Some strands fall wholly or partly into the sieve, whilst others do not. Strands may contribute to more than one sieve channel.

Harmonic grouping for synthetic 4-formant syllable





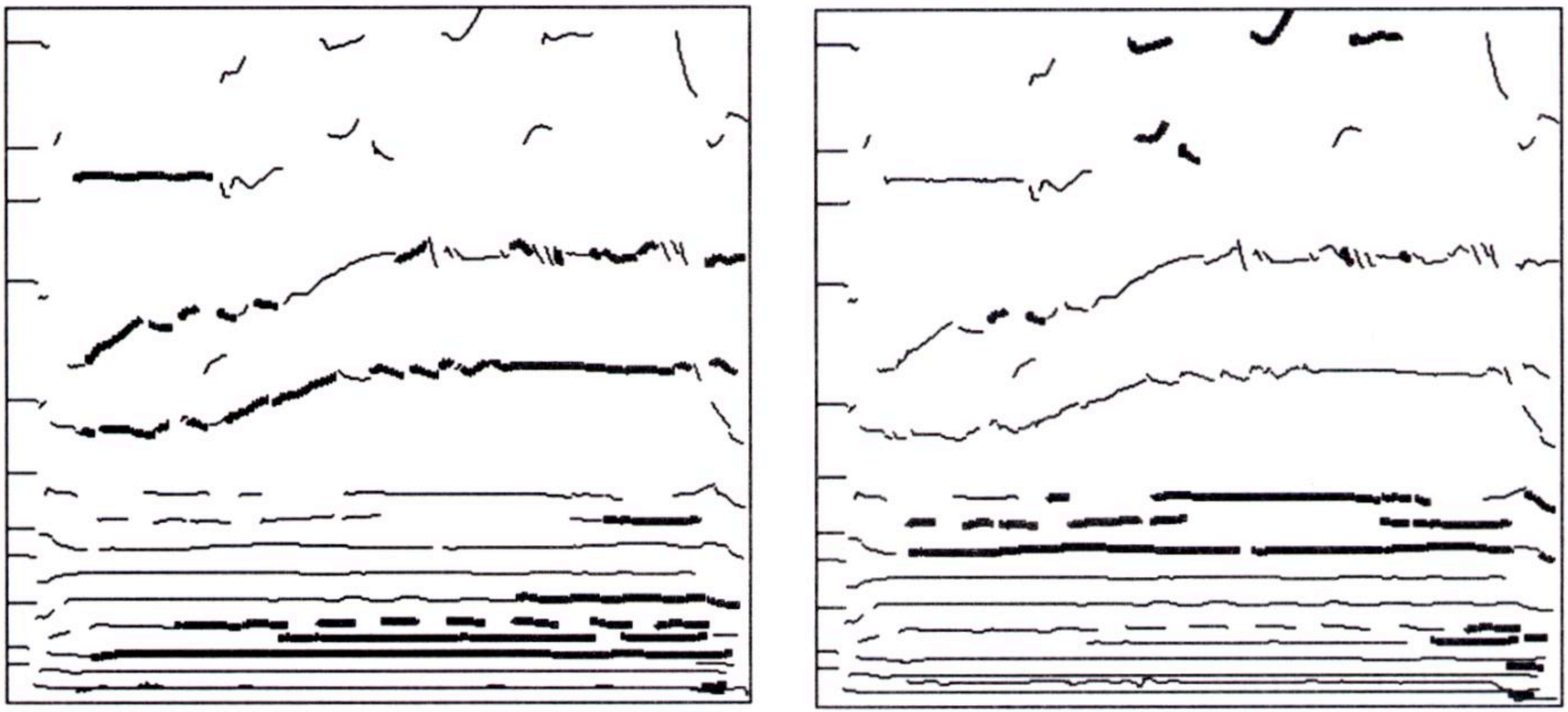


1^{st}, 3^{rd}, and 4^{th} formant are synthesized on a fundamental of 110Hz

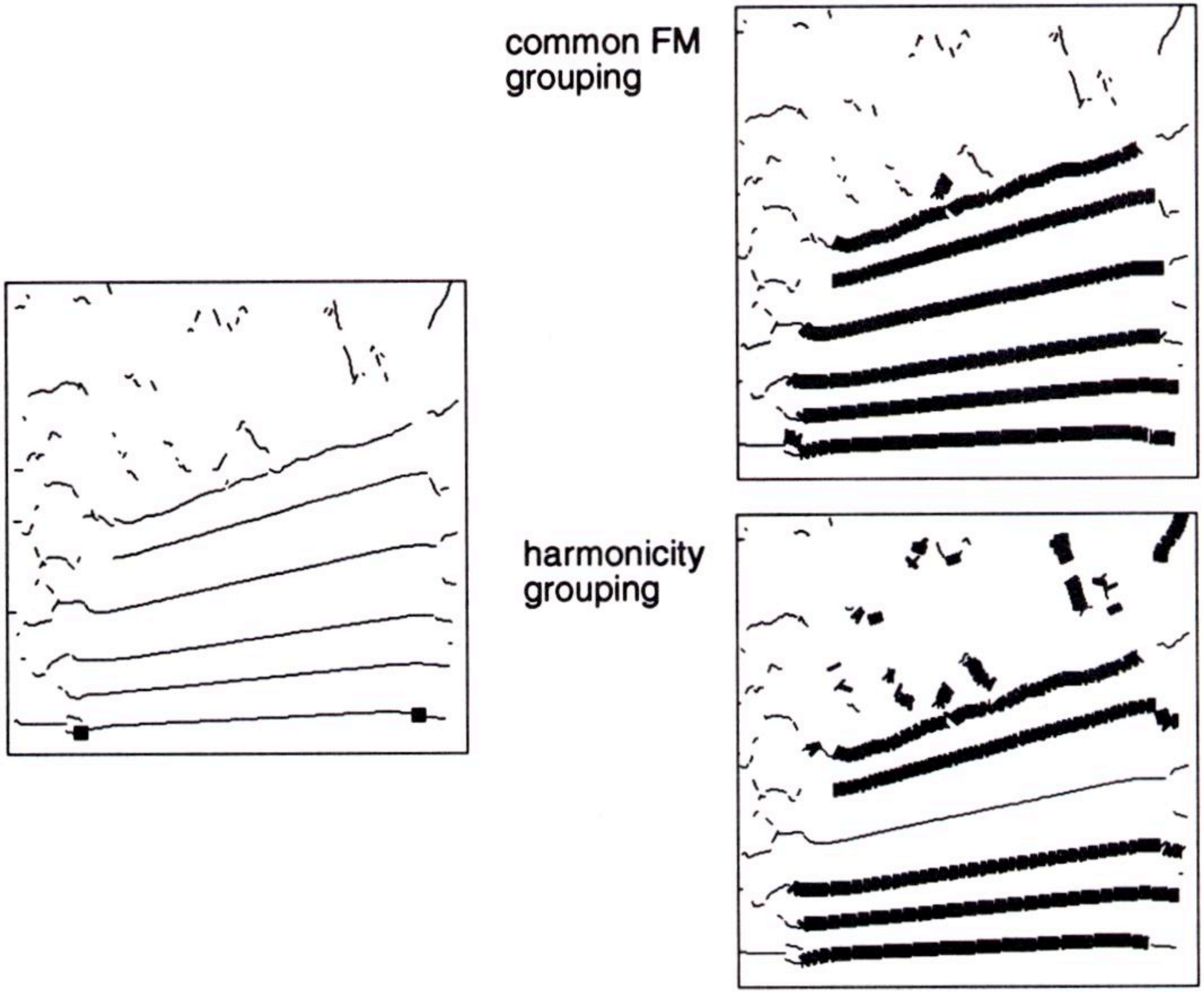
2^{nd} formant synthesized on fundamental of 174Hz

Algorithm finds two groups

Grouping by common amplitude modulation

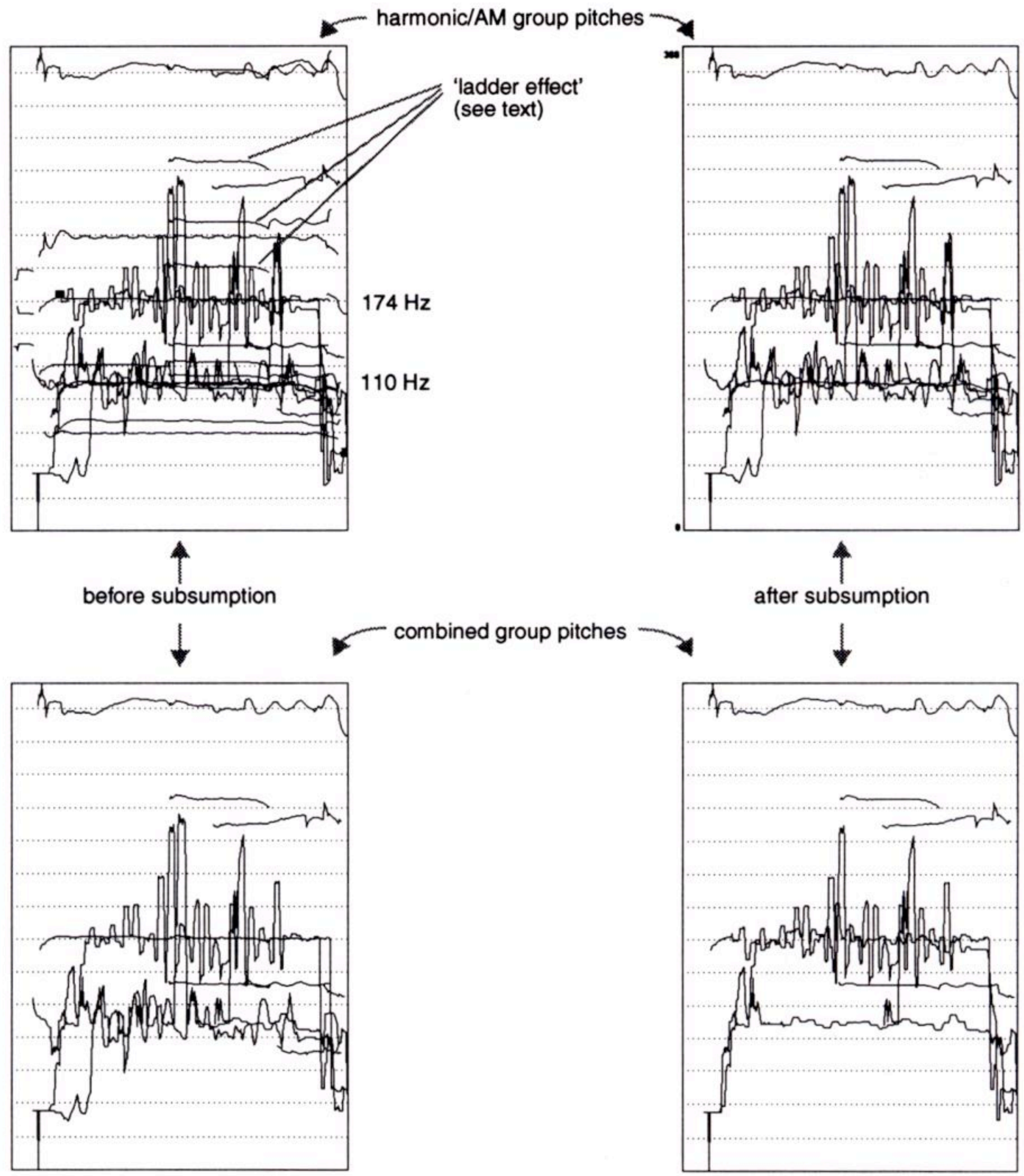


Grouping by common frequency movement

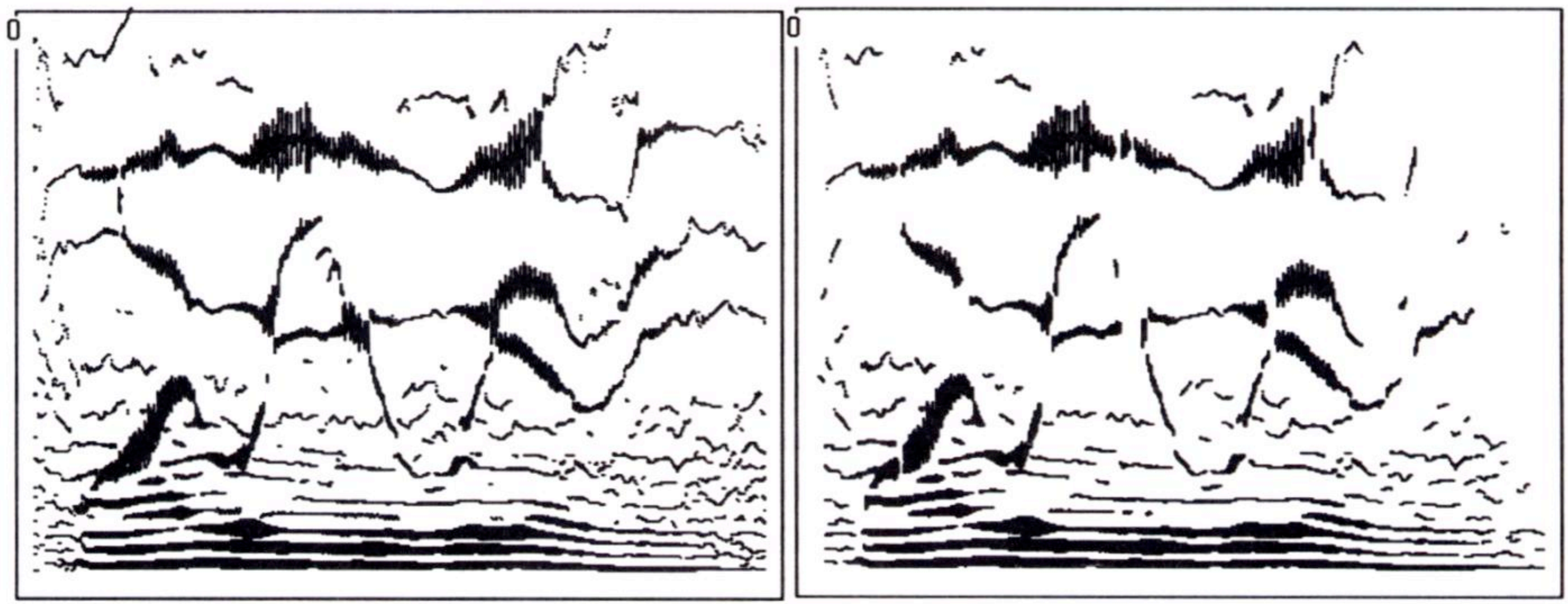


Subsumption: forming higher level groups

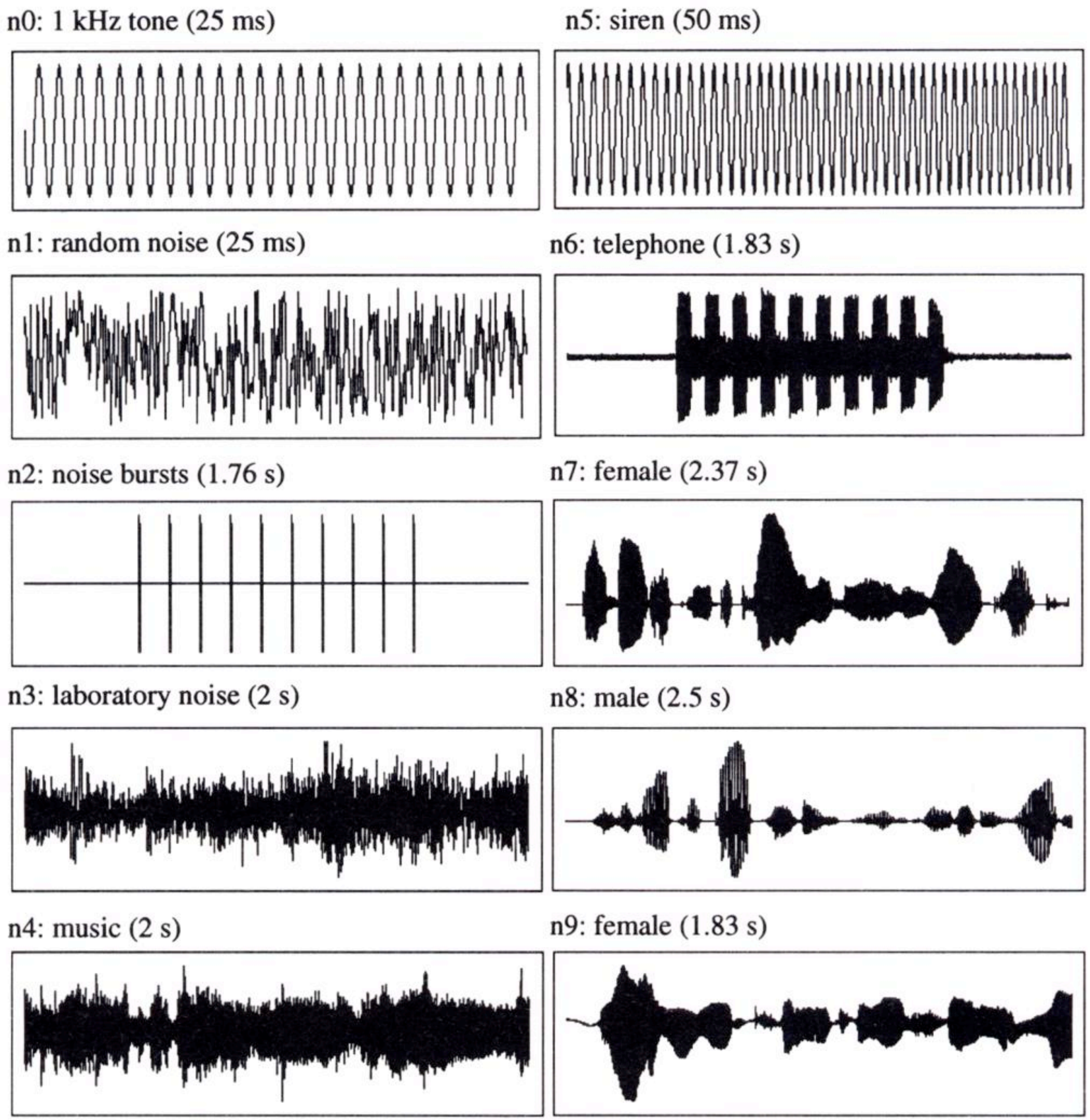
Remove groups that are contained within a larger group



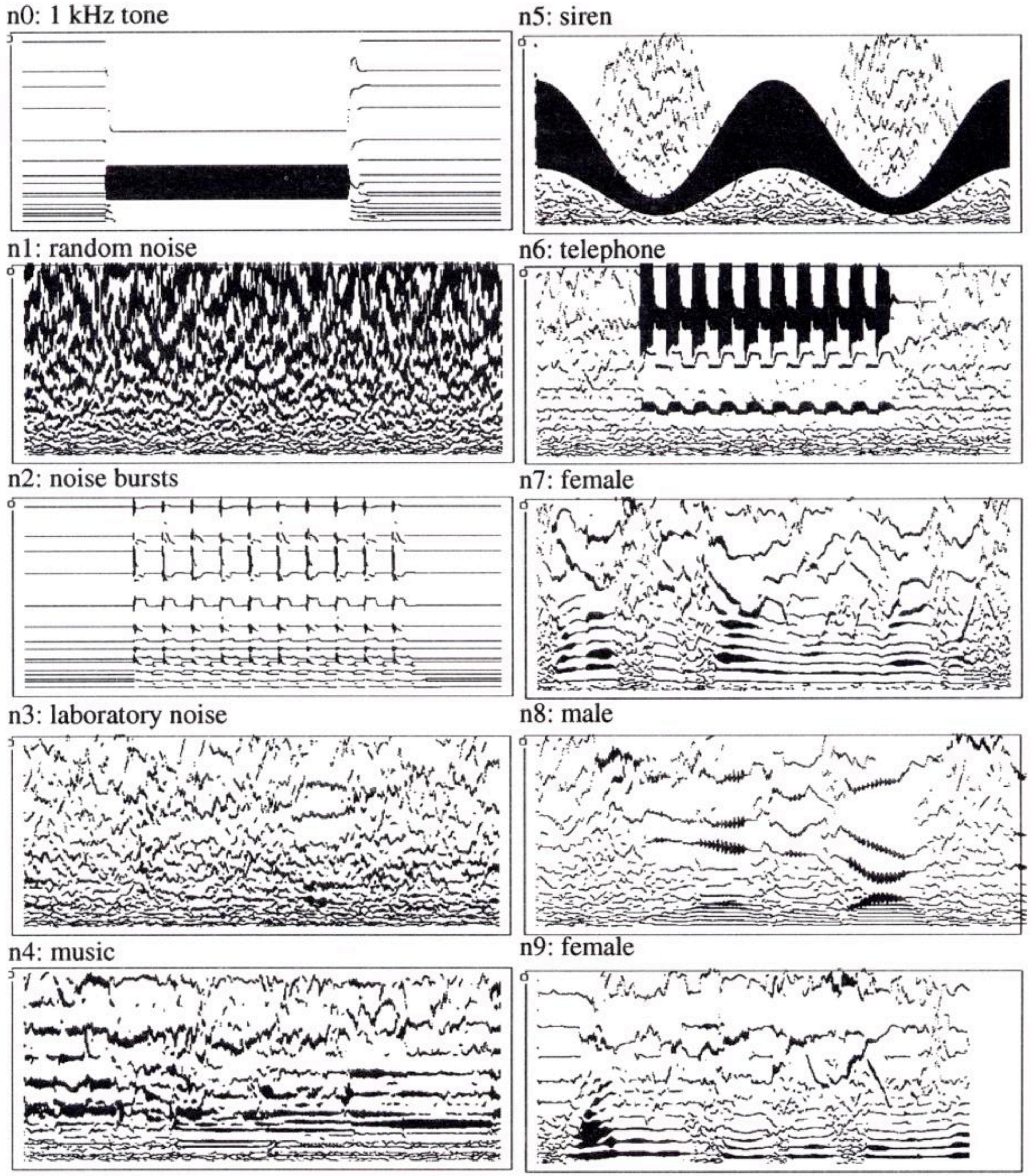
Grouping natural speech



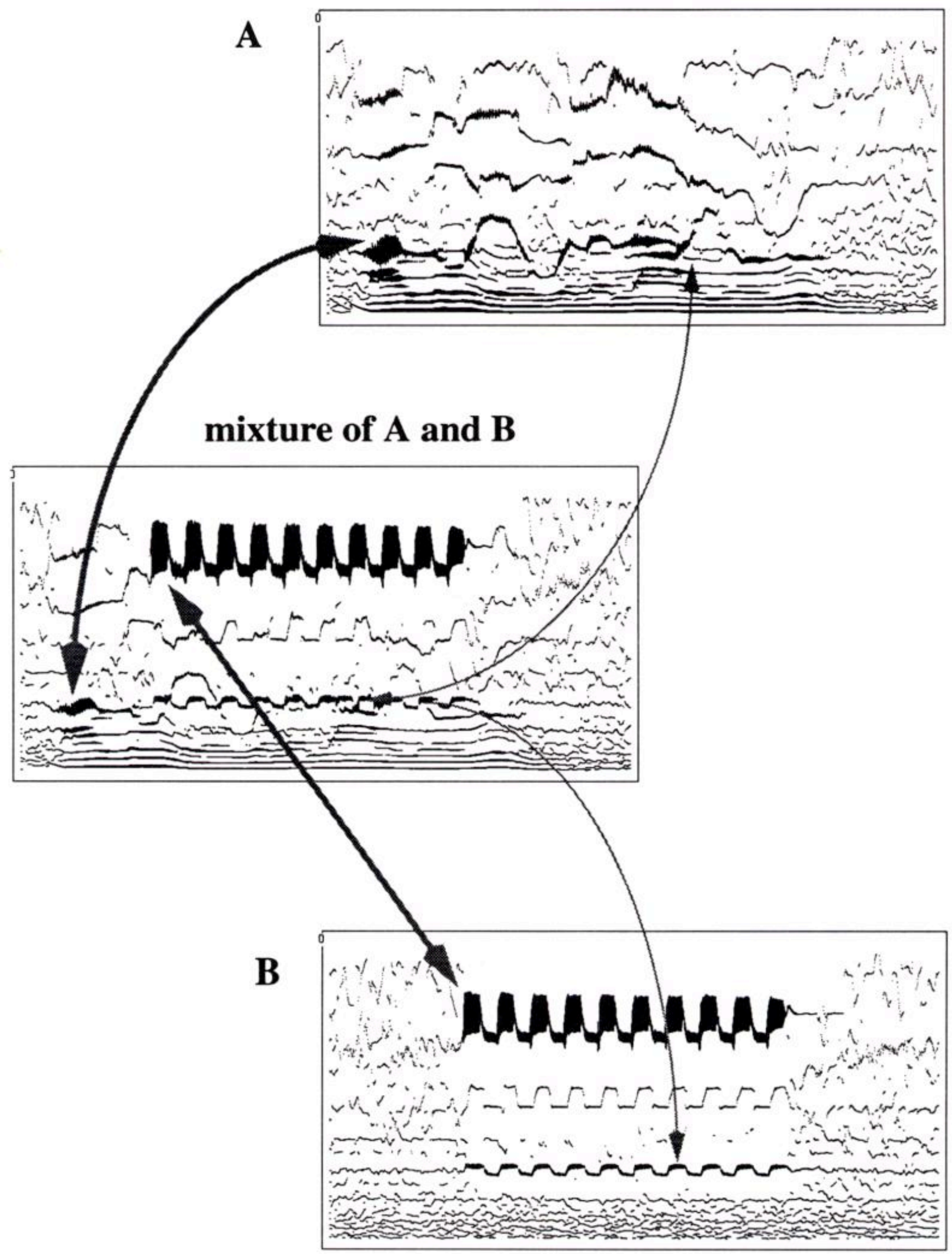
Test sounds and noise sources



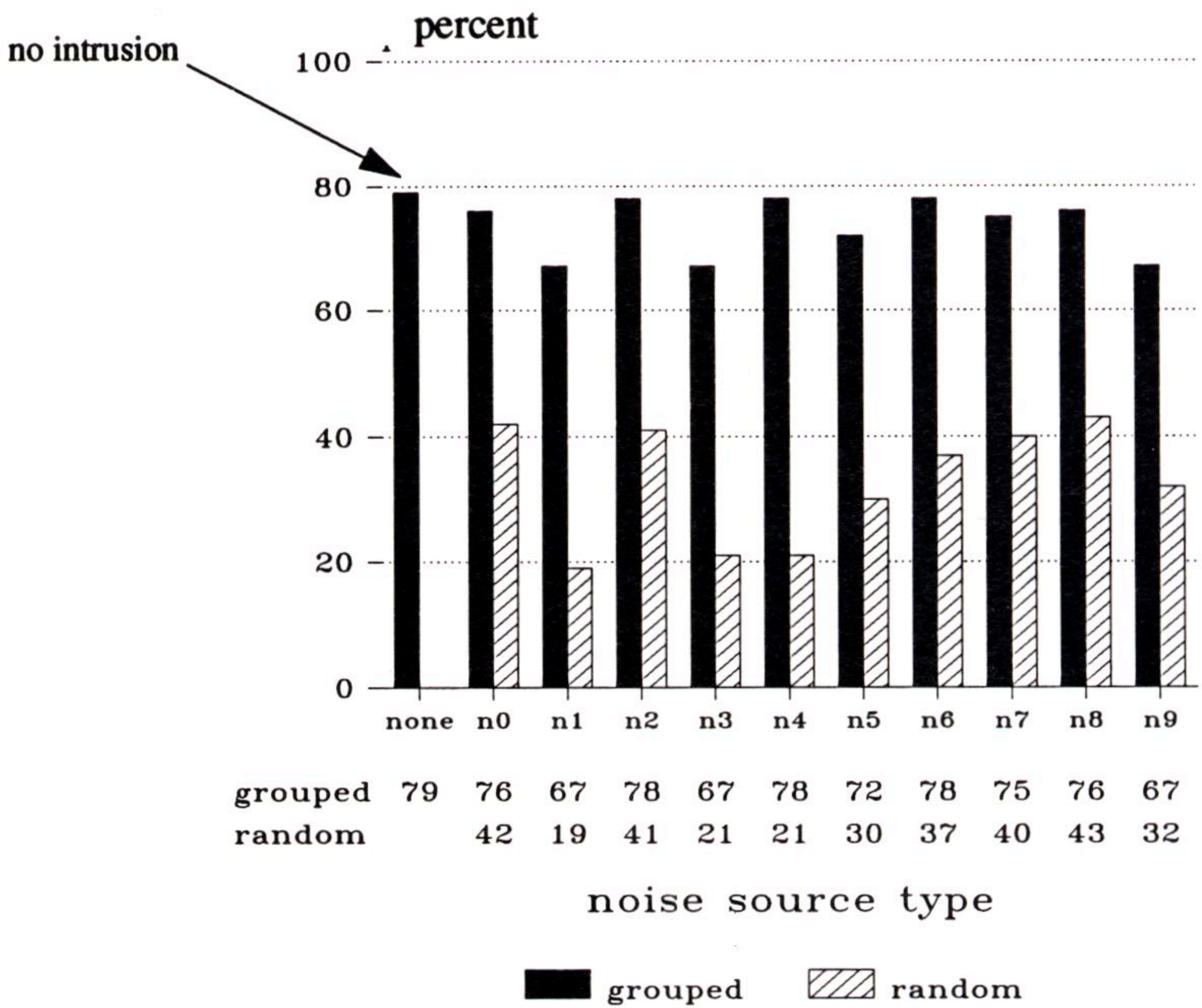
Synchrony strand representations of test sounds



The mixture correspondence problem



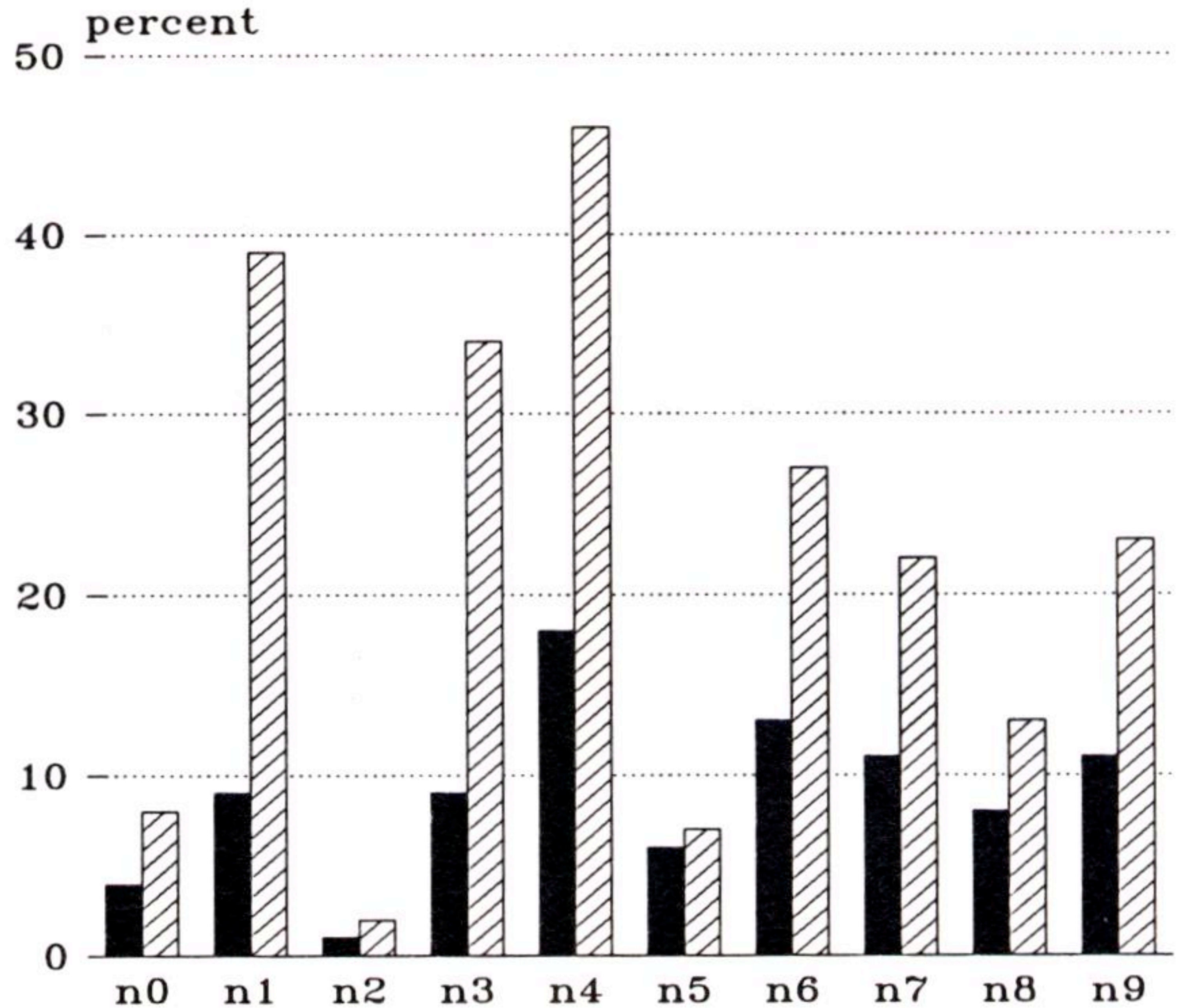
Performance: Utterance characterization



Performance: crossover from intrusive source

How much is the intrusive source incorrectly grouped?

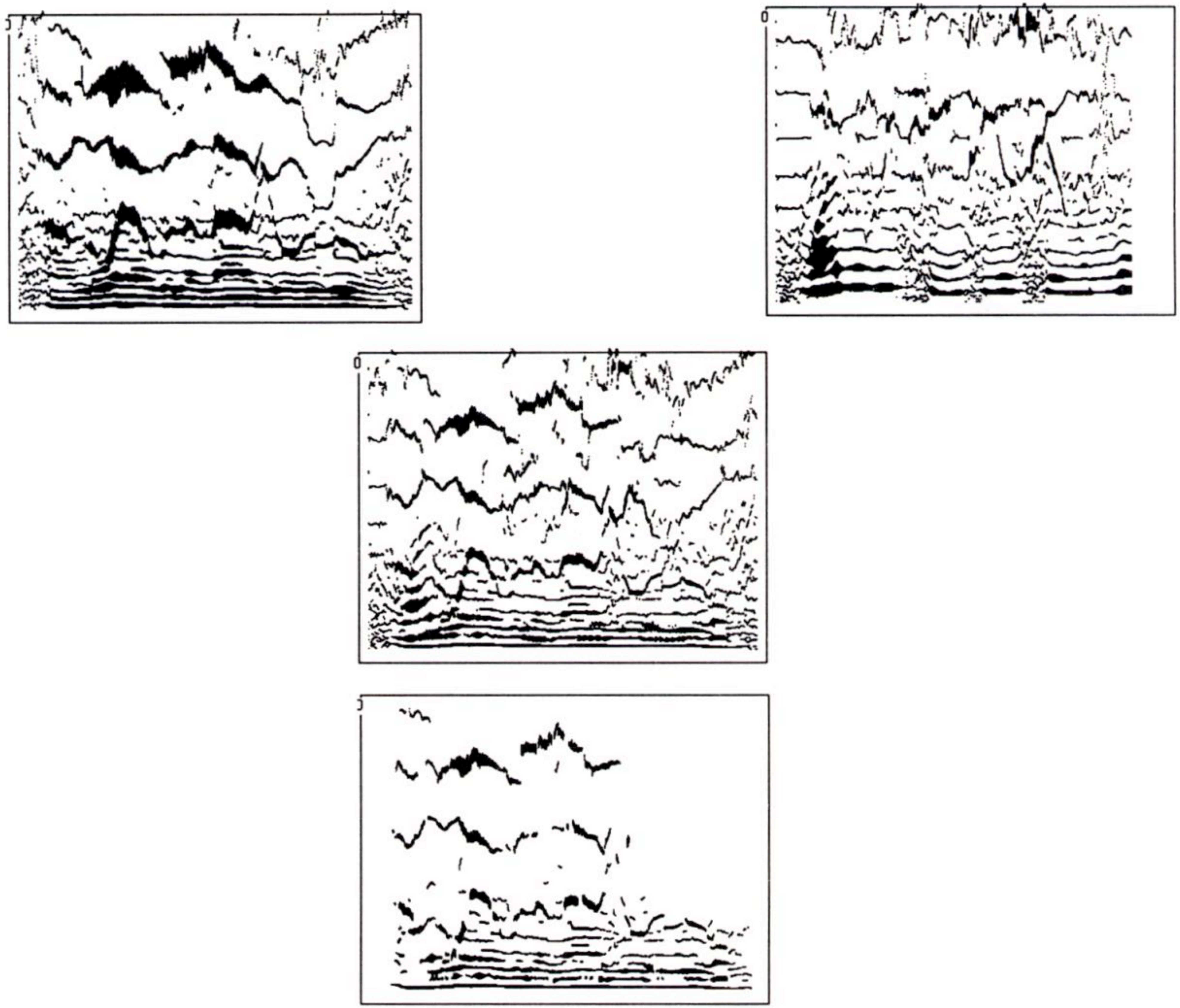
- Least intrusion when intrusive source is noise bursts (n2)
- Worst is when intrusive source is music (n4)



	n0	n1	n2	n3	n4	n5	n6	n7	n8	n9
grouped	4	9	1	9	18	6	13	11	8	11
random	8	39	2	34	46	7	27	22	13	23

grouped random

Grouping of speech from a mixture (cocktail party)



Grouping speech in presence of laboratory noise