I. Real-world audition

- **The hearing problem facing a listener**
- **Listener's performance**

Real-world audition

What?

- Speech
  - message
  - speaker
    - age, gender, linguistic origin, mood, …
- Music
- Car passing by

Where?

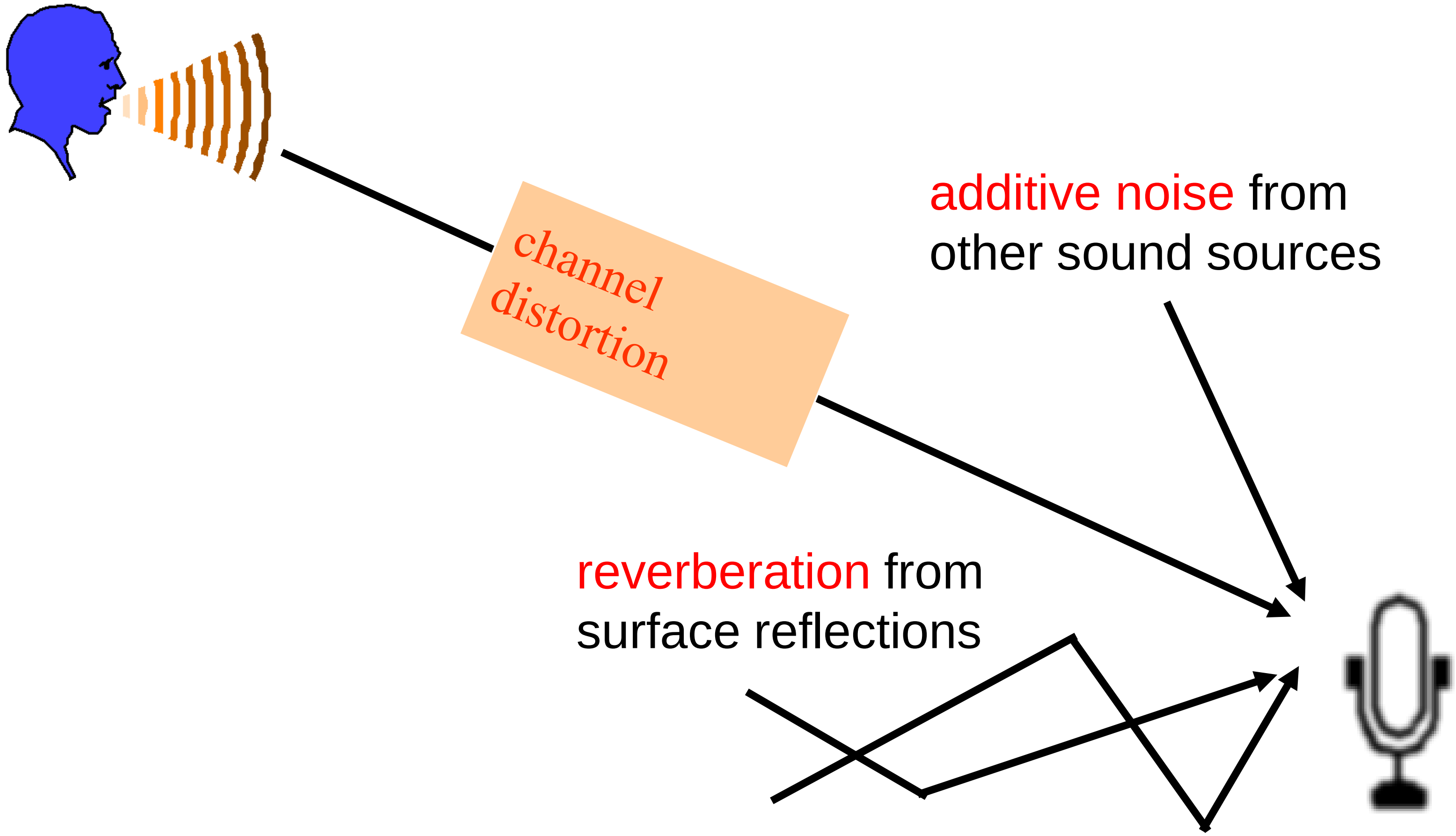
- Left, right, up, down
- How close?

Channel characteristics

Environment characteristics

- Room reverberation
- Ambient noise

Sources of intrusion and distortion



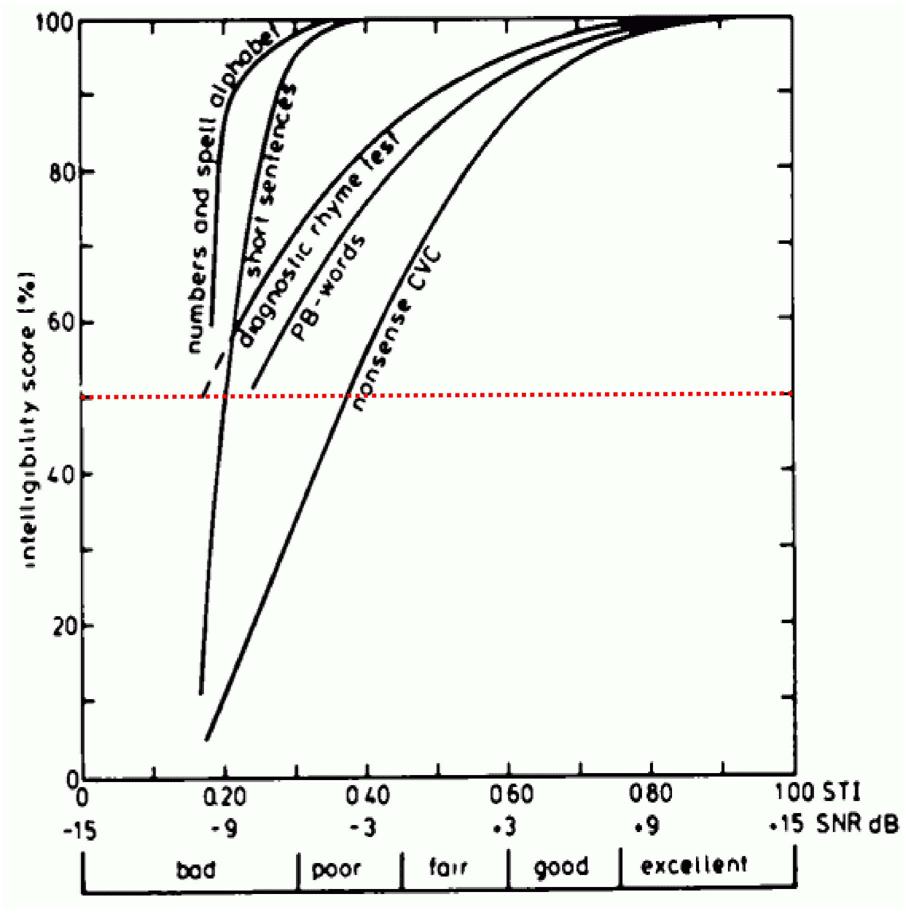
Cocktail party problem

- **Term coined by Cherry**
  - "One of our most important faculties is our ability to listen to, and follow, one speaker in the presence of others. This is such a common experience that we may take it for granted; we may call it 'the cocktail party problem'…" (Cherry, 1957)
  - "For 'cocktail party'-like situations… when all voices are equally loud, speech remains intelligible for normal-hearing listeners even when there are as many as *six* interfering talkers" (Bronkhorst & Plomp, 1992)
- **Ball-room problem by Helmholtz**
  - "Complicated beyond conception" (Helmholtz, 1863)
- **Speech segregation problem**

Listener performance

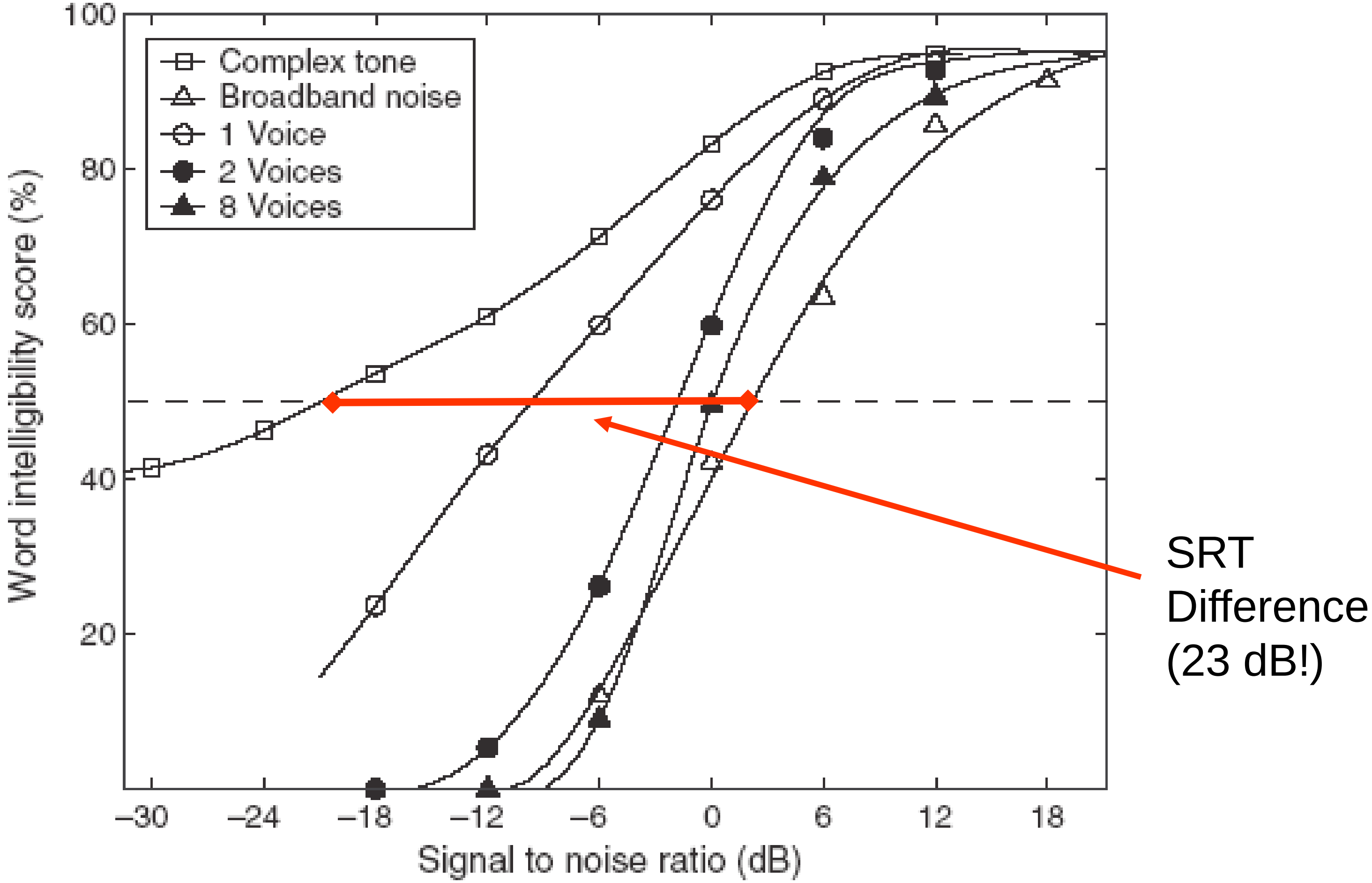
Speech reception threshold (SRT)

- The speech-to-noise ratio needed for 50% intelligibility
- Each 1 dB gain in SRT corresponds to 5-10% increase in intelligibility (Miller et al., 1951) dependent upon materials



Source: Steeneken (1992)

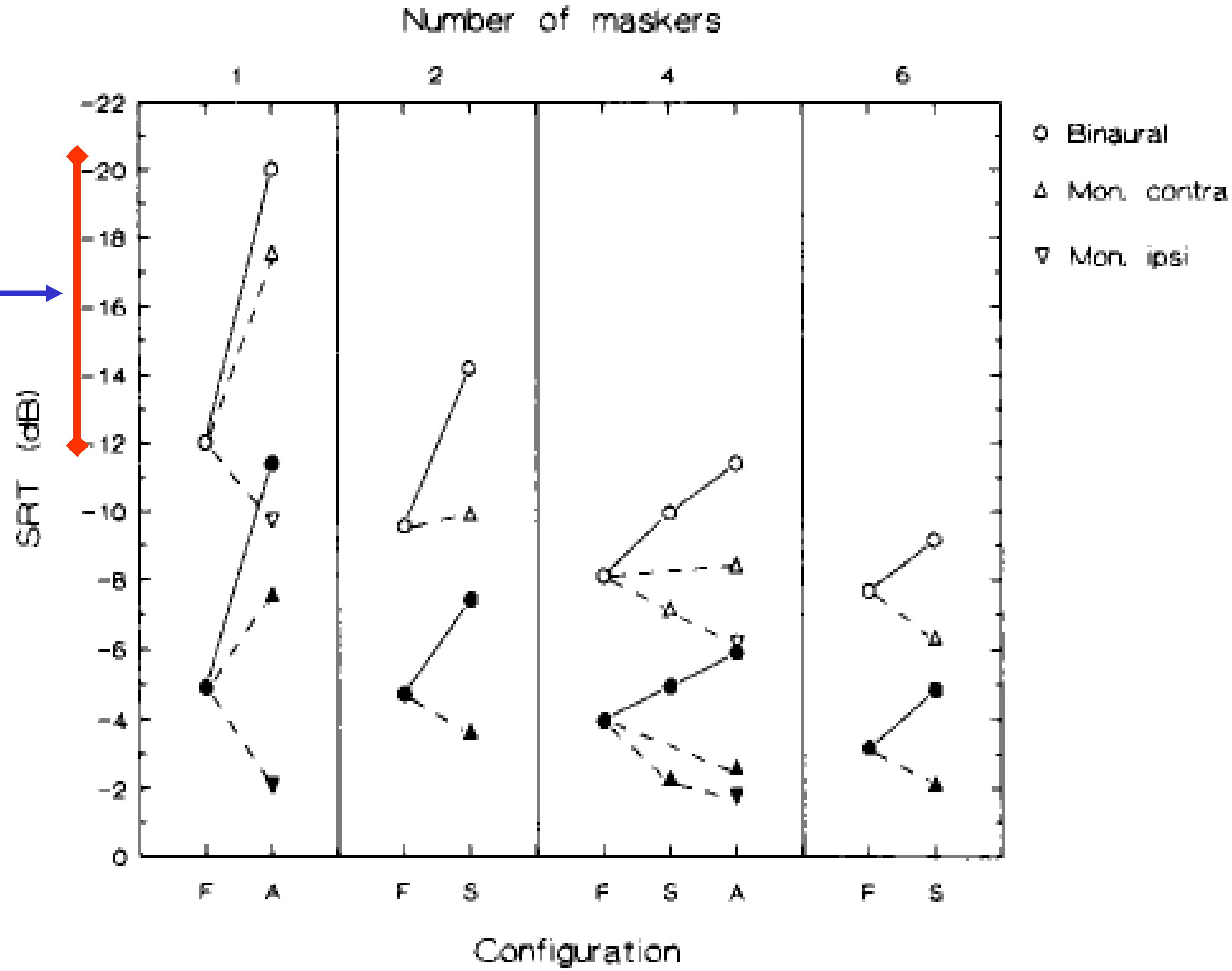
Effects of competing source



Source: Wang and Brown (2006)

Location

SRT gain



Source: Bronkhorst & Plomp (1992)

Part II. Fundamental auditory representations

- **Modeling of the auditory periphery**
- **Organization in speech**
- **Auditory representations**

Cochlear filtering model

The *gammatone* function approximates physiologically-recorded impulse responses

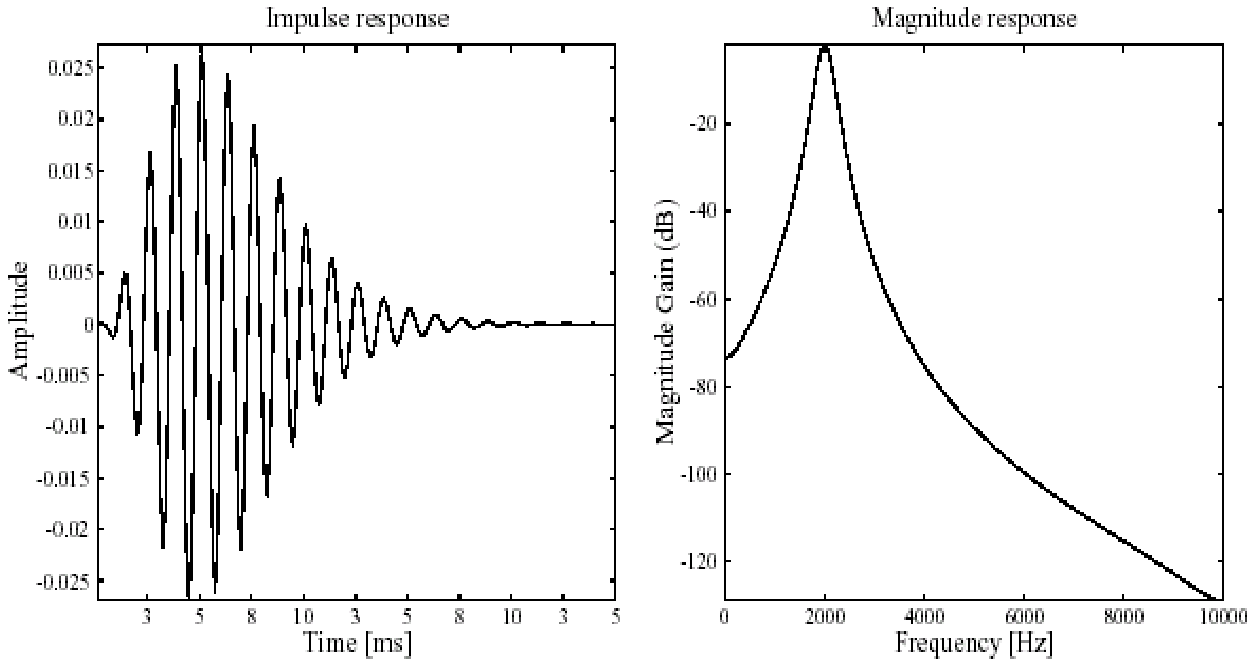
$$g(t) = t^{n-1} \exp(-2\pi b t) \cos(2\pi f_0 t + \phi)$$

n = filter order (typically 4)

b = bandwidth

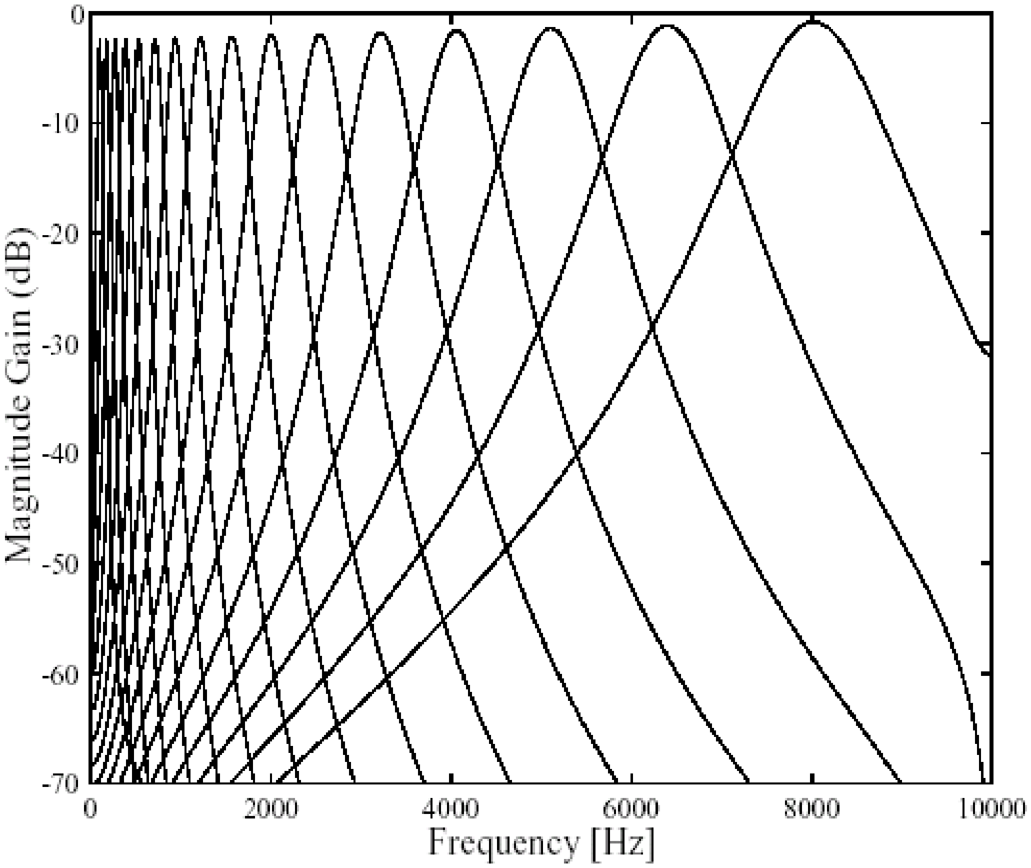
f_0 = centre frequency

ϕ = phase



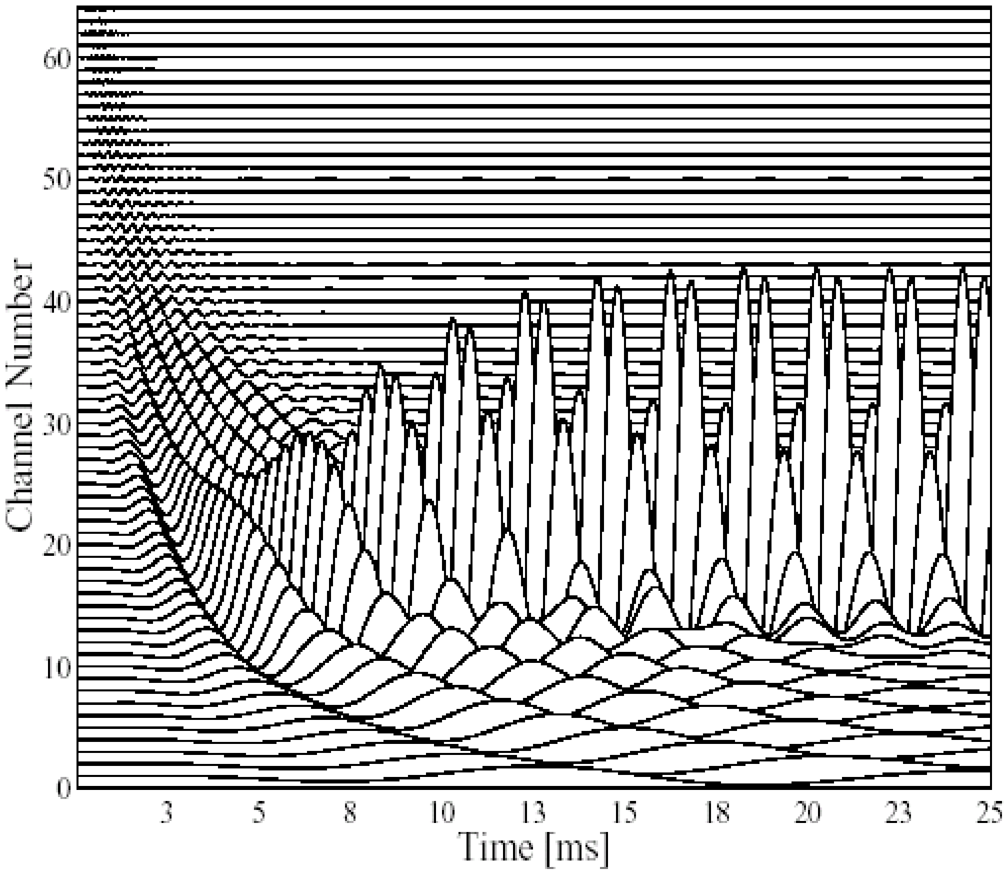
Gammatone filterbank

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



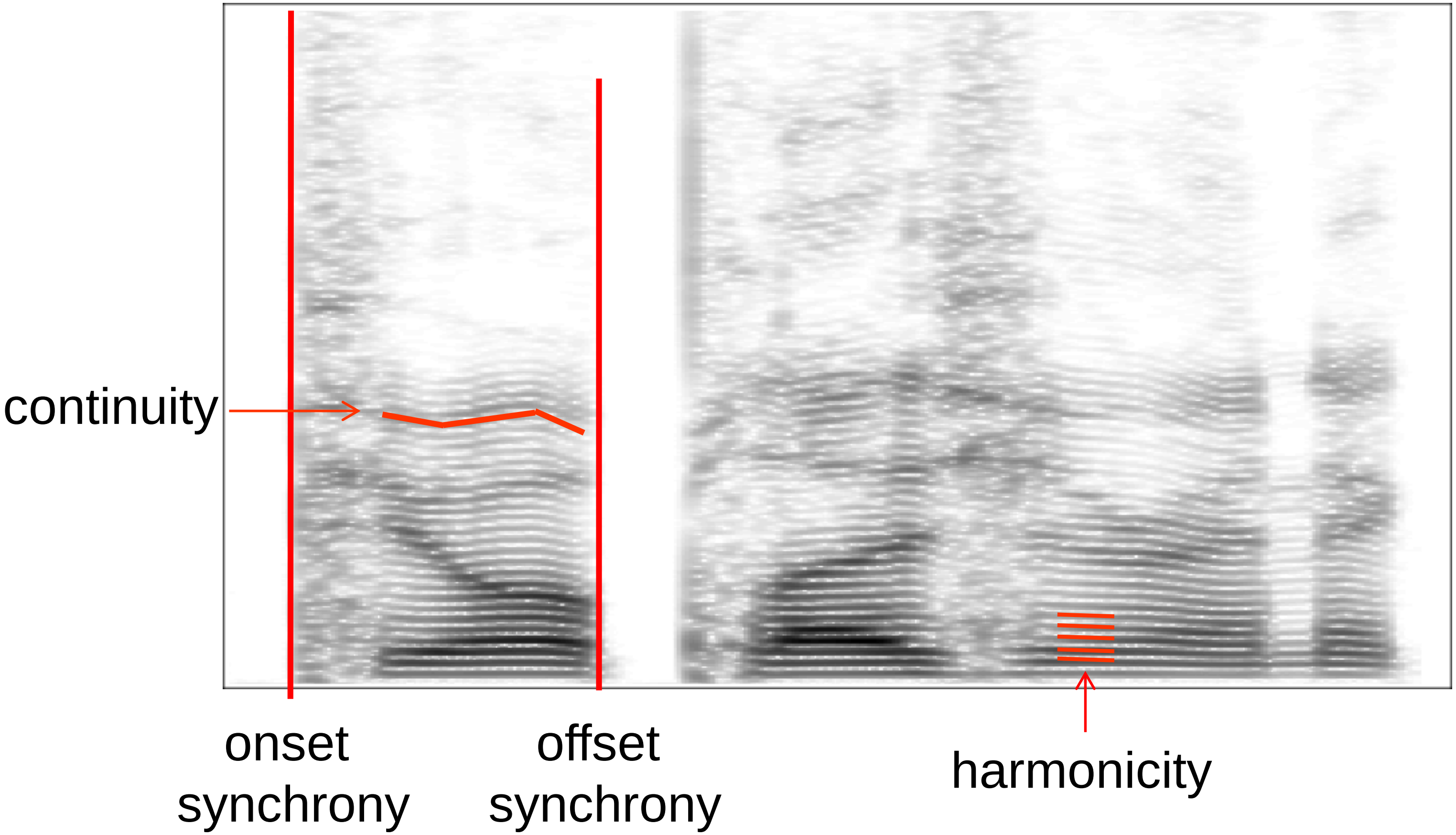
Response to a pure tone

- **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**



Organization in speech: Spectrogram

"… pure pleasure … "



Cochleagram: Auditory spectrogram

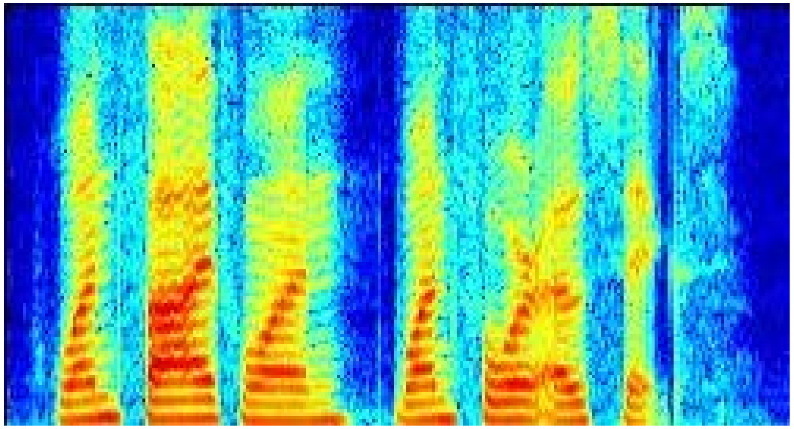
Spectrogram

- Plot of log energy across time and frequency (linear frequency scale)

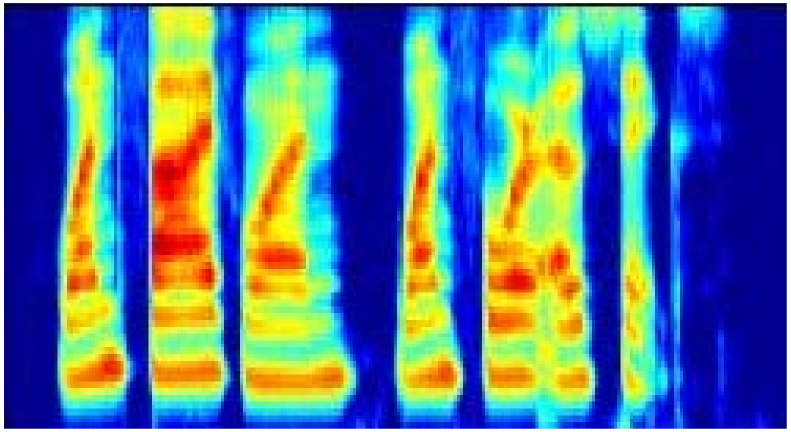
Cochleagram

- Cochlear filtering by the gammatone filterbank (or other models of cochlear filtering), followed by a stage of nonlinear rectification; the latter corresponds to hair cell transduction by either a hair cell model or simple compression operations (log and cube root)
- Quasi-logarithmic frequency scale, and filter bandwidth is frequency-dependent
- A waveform signal can be constructed (inverted) from a cochleagram

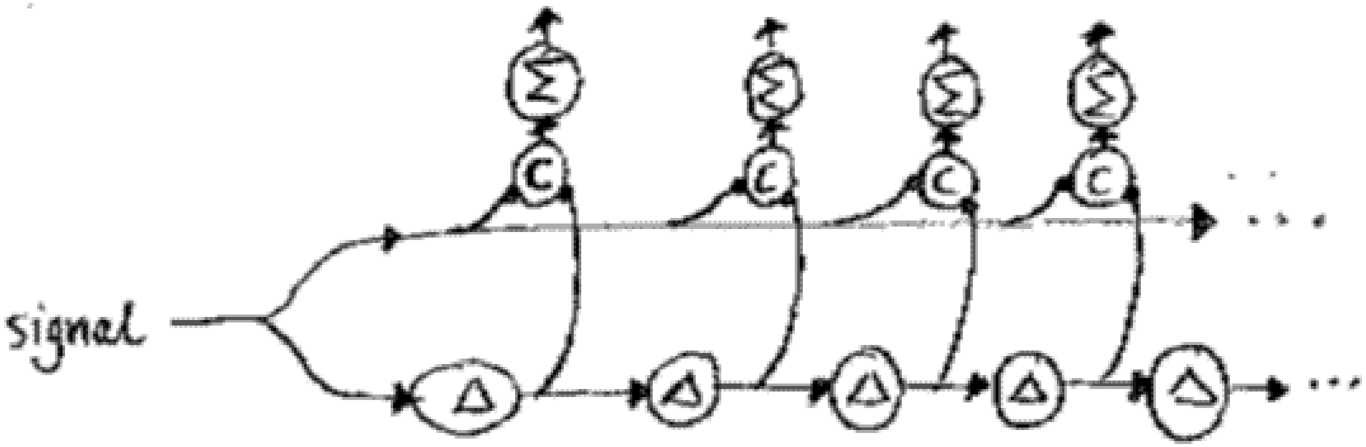
Spectrogram



Cochleagram



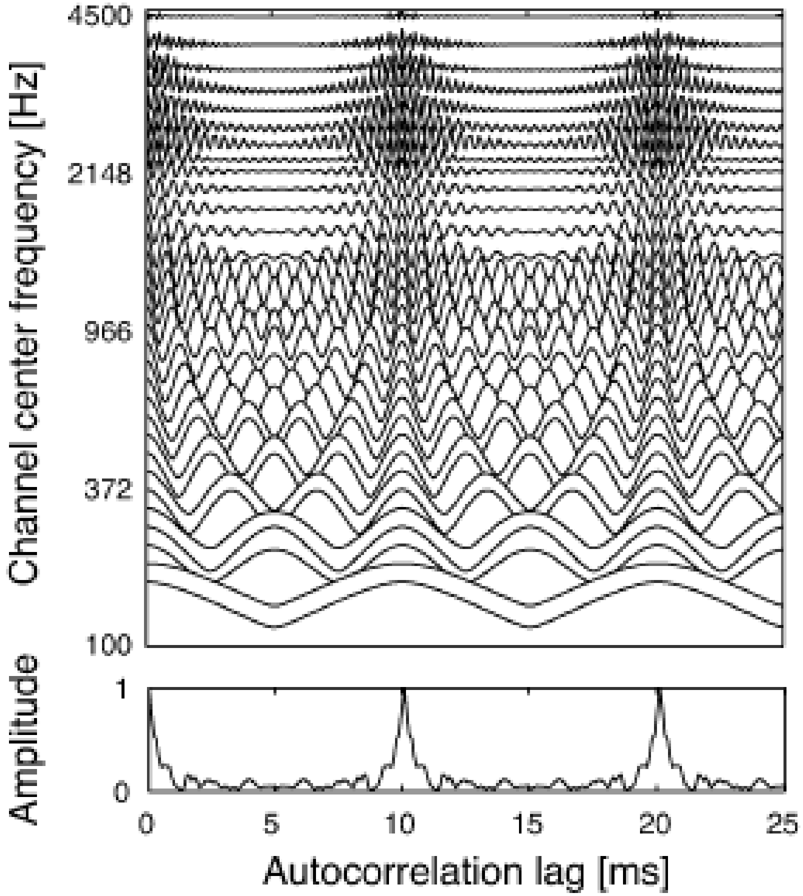
Neural autocorrelation for pitch perception



Licklider (1951)

Correlogram

- **Short-term autocorrelation of the output of each frequency channel of the cochleagram**
- **Peaks in summary correlogram indicate pitch periods (F0)**
- **A standard model of pitch perception**



Correlogram & summary correlogram of a vowel with F0 of 100 Hz

Onset and offset detection

- **An onset (offset) corresponds to a sudden intensity increase (decrease), which can be detected by taking the time derivative of the intensity**
- **To reduce intensity fluctuations, Gaussian smoothing (low-pass filtering) is typically applied (as in edge detection for image analysis):**

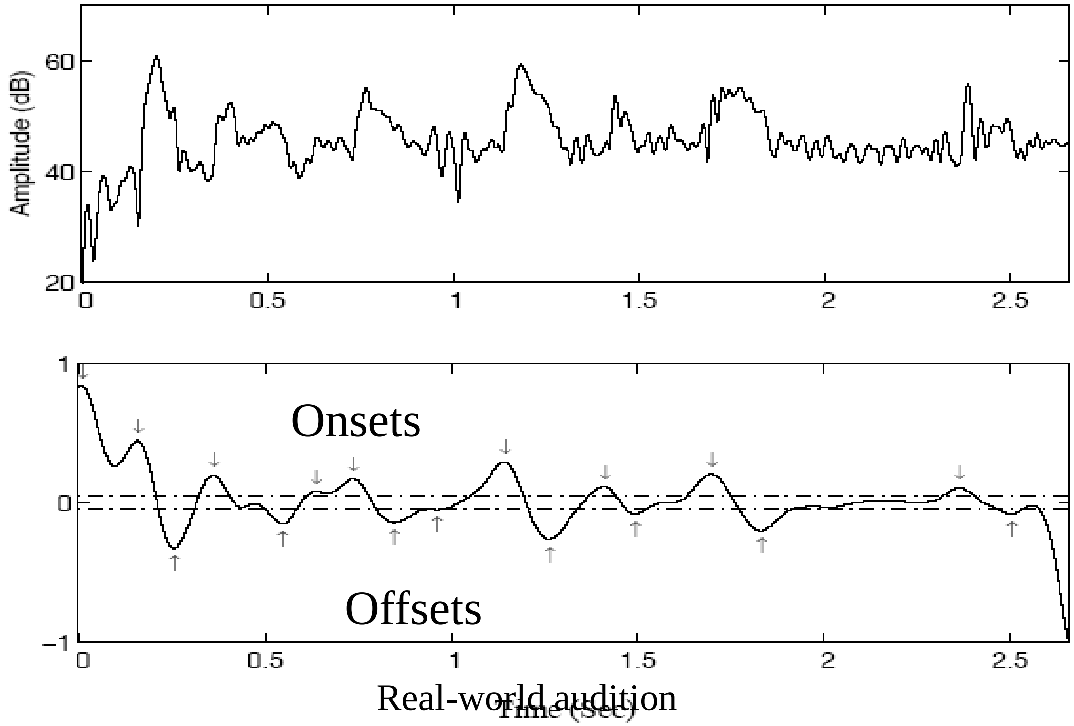
$$G(t,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{t^2}{2\sigma^2})$$

- **Note that** $(s(t) * G(t,\sigma))' = s(t) * G'(t,\sigma)$**, where** $s(t)$ **denotes intensity and**

$$G'(t,\sigma) = \frac{-t}{\sqrt{2\pi}\sigma^3} \exp(-\frac{t^2}{2\sigma^2})$$

Onset and offset detection (cont.)

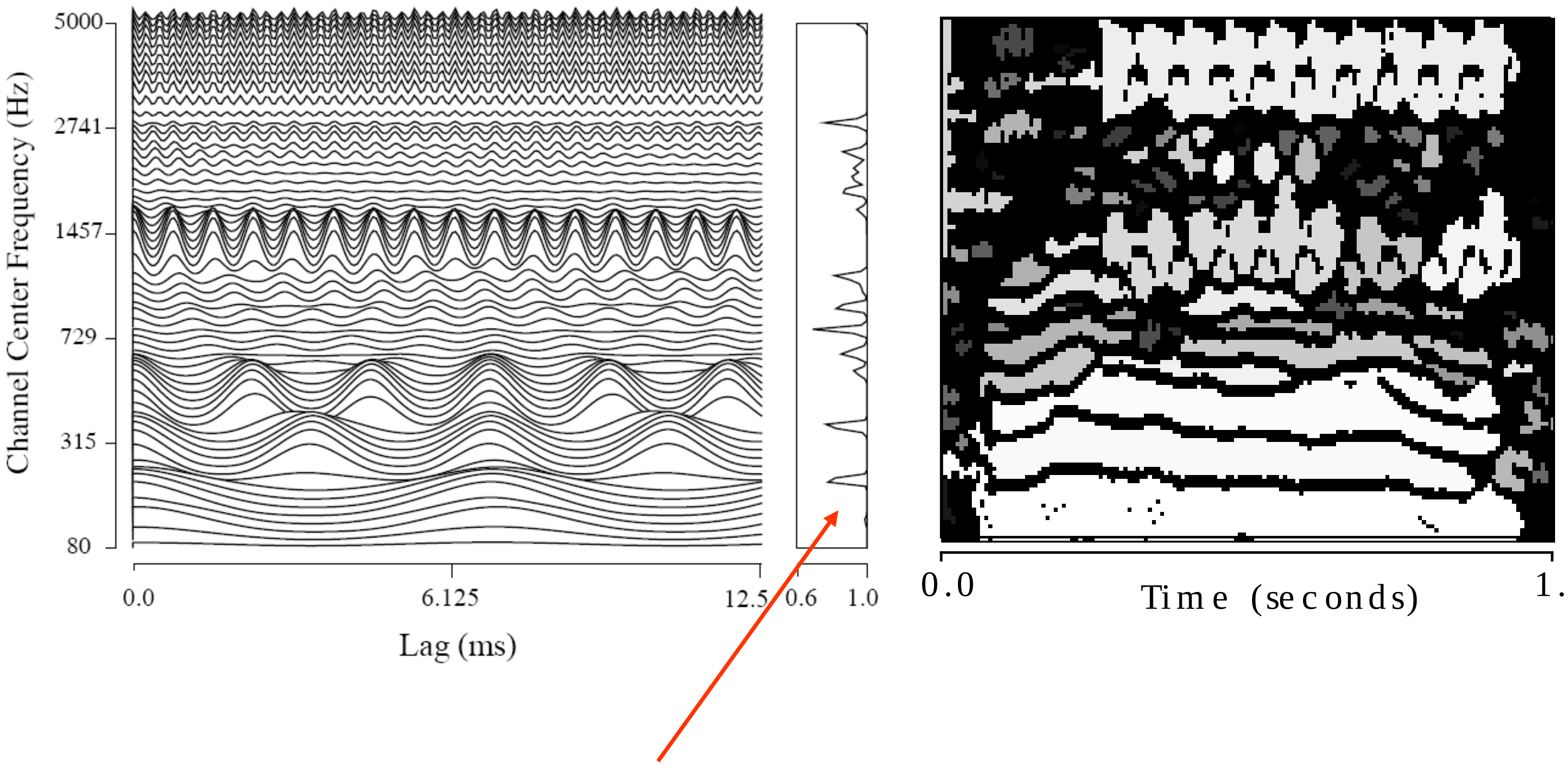
- **Hence onset and offset detection is a three-step procedure**
  - Convolve the intensity $s(t)$ with G' to obtain $O(t)$
  - Identify the peaks and the valleys of $O(t)$
  - Onsets are those peaks above a certain threshold, and offsets are those valleys below a certain threshold



Segmentation versus grouping

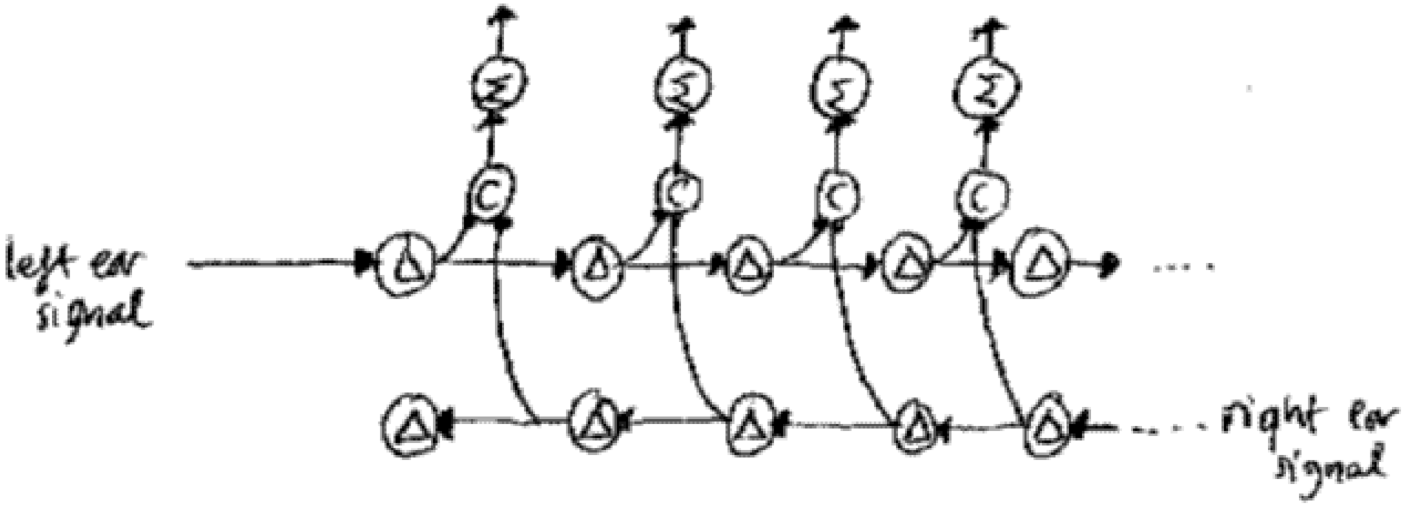
- **Mirroring Bregman's two-stage conceptual model, a CASA model generally consists of a segmentation stage and a subsequent grouping stage**
- **Segmentation stage decomposes an acoustic scene into a collection of segments, each of which is a contiguous region in the cochleagram with energy primarily from one source**
  - Based on cross-channel correlation that encodes correlated responses (temporal fine structure) of adjacent filter channels, and temporal continuity
  - Based on onset and offset analysis
- **Grouping aggregates segments into streams based on various ASA cues**

Cross-channel correlation for segmentation



- Correlogram and cross-channel correlation for a mixture of speech and trill telephone
- Segments generated based on cross-channel correlation and temporal continuity

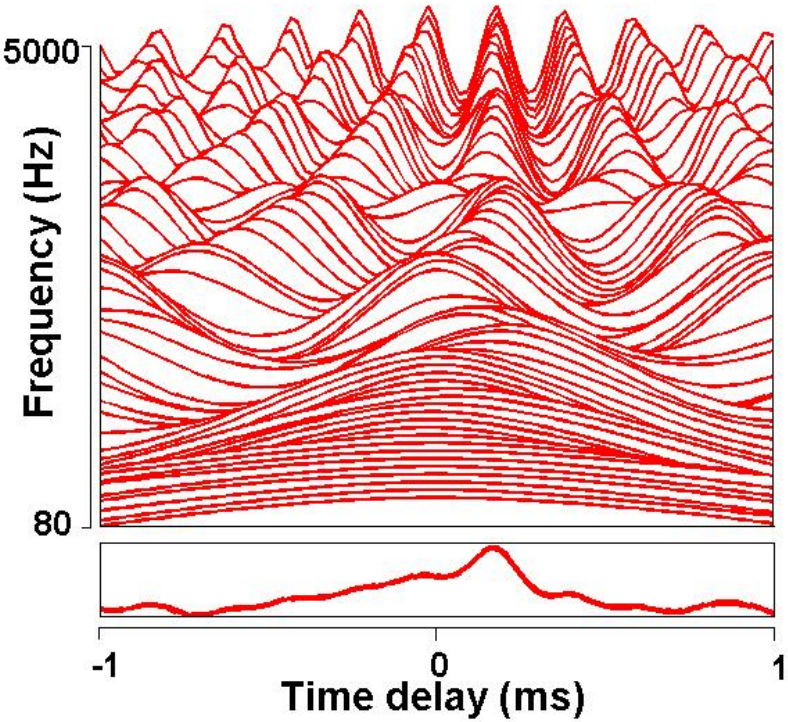
Neural cross-correlation



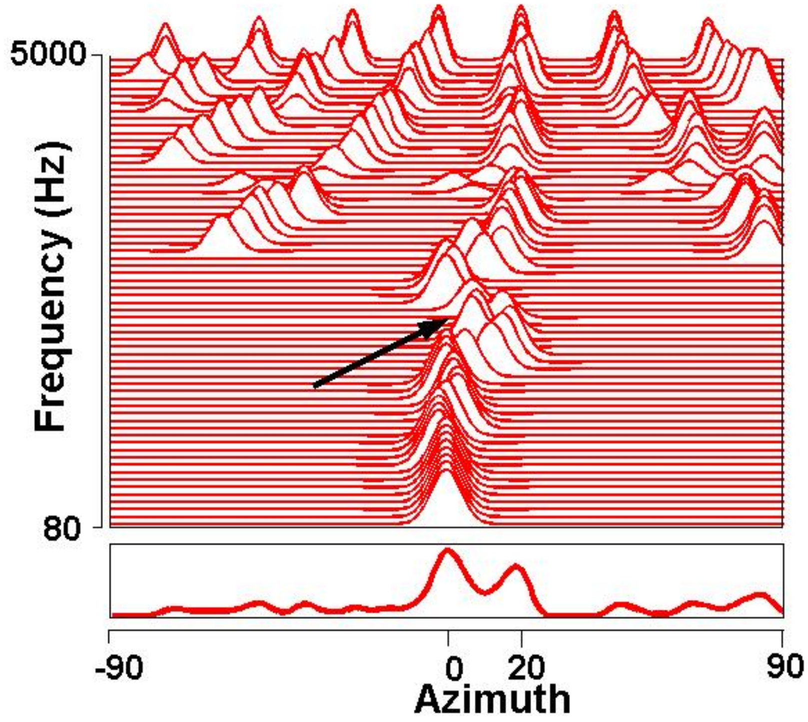
Jeffress (1948)

- Cross-correlogram: Cross-correlation (or coincidence) between the left ear signal and the right ear signal
- Strong physiological evidence supporting this neural mechanism for sound localization (more specifically azimuth localization)

Azimuth localization example (Target: 0°, Noise: 20°)



Cross-correlogram within one frame

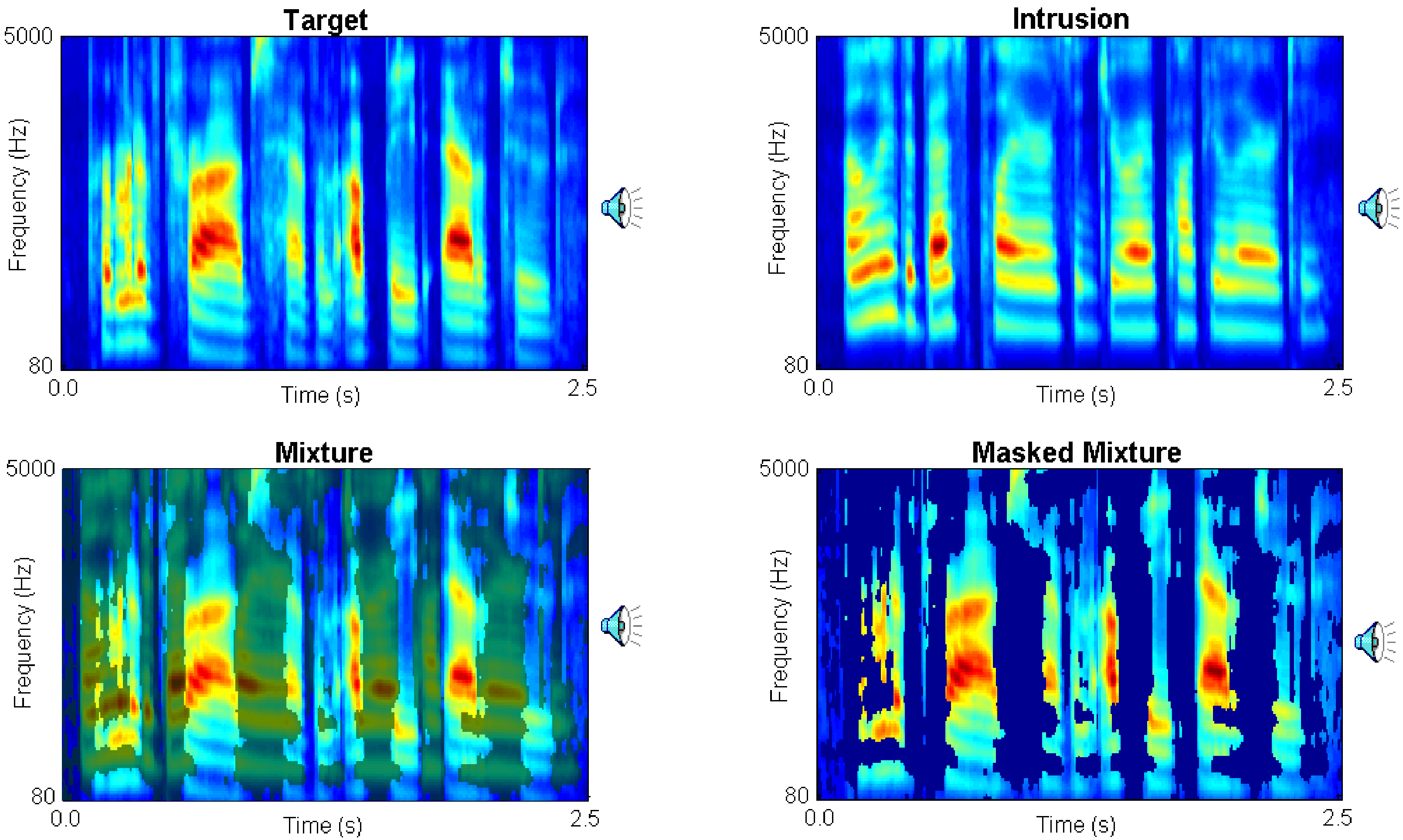


Skeleton cross-correlogram sharpens cross-correlogram, making peaks in the azimuth axis more pronounced

Ideal binary mask

- **A main CASA goal is to retain the parts of a mixture where target sound is stronger than the acoustic background (i.e. to mask interference by the target), and discard the other parts (Hu & Wang, 2001; 2004)**
  - What a target is depends on intention, attention, etc.
- **In other words, the goal is to identify the ideal binary mask (IBM), which is 1 for a time-frequency (T-F) unit if the SNR within the unit exceeds a threshold, and 0 otherwise**
  - It does not actually separate the mixture!

IBM illustration



Properties of the IBM

- **Consistent with the auditory masking phenomenon: A stronger signal masks a weaker one within a critical band**
- **Optimality: Under certain conditions the ideal binary mask with 0 dB local SNR criterion is the optimal binary mask for SNR gain (Li and Wang, 2009)**
- **The ideal binary mask is very effective for human speech intelligibility (Brungart et al., 2006; Li and Loizou, 2008)**
- **The IBM provides an excellent front-end for robust automatic speech recognition**