EECS 225D
Audio Signal Processing in Humans and Machines

Lecture 17 – Source Separation

2012-3-19

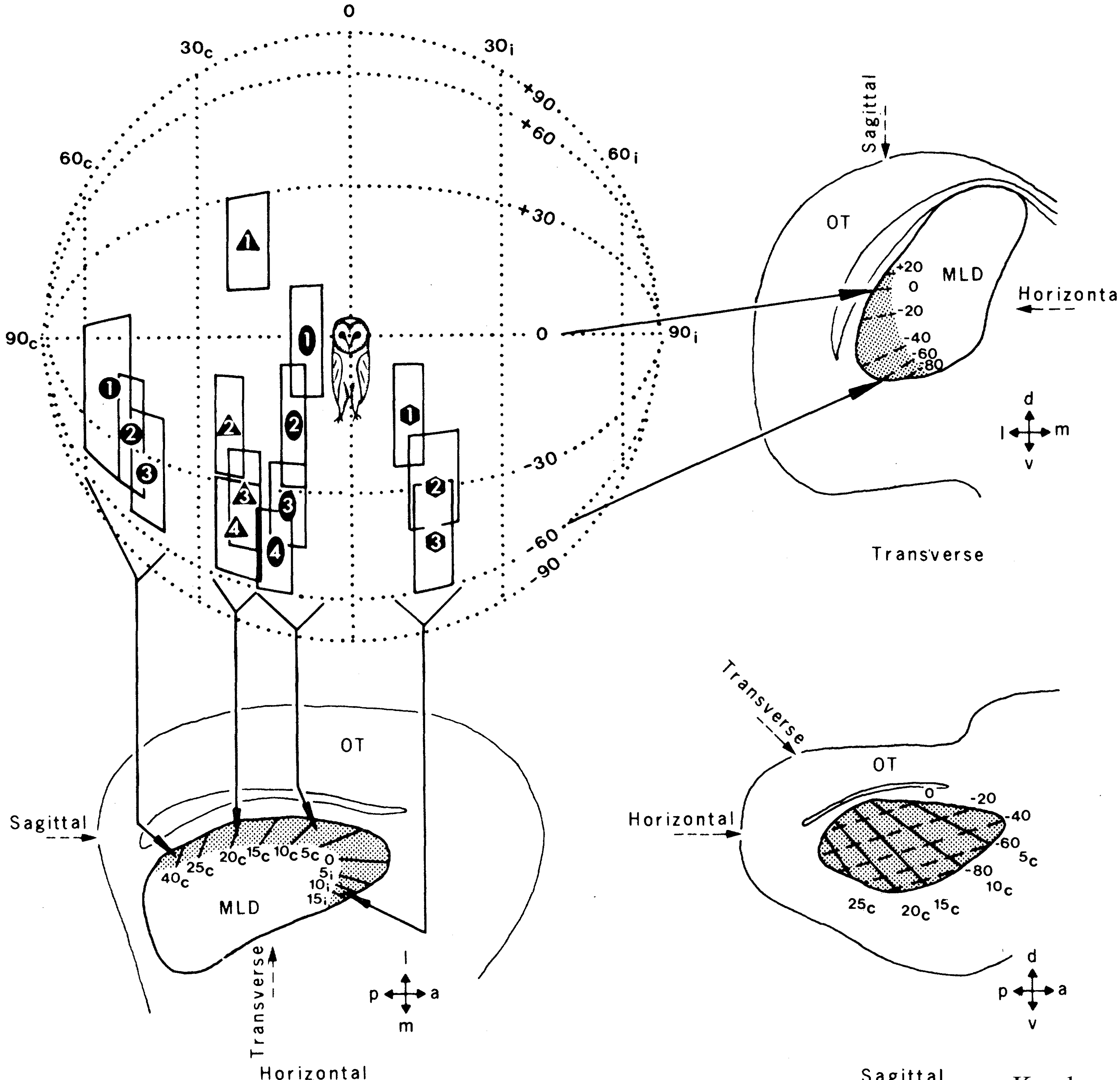
Professor Nelson Morgan

today's lecture by John Lazzaro

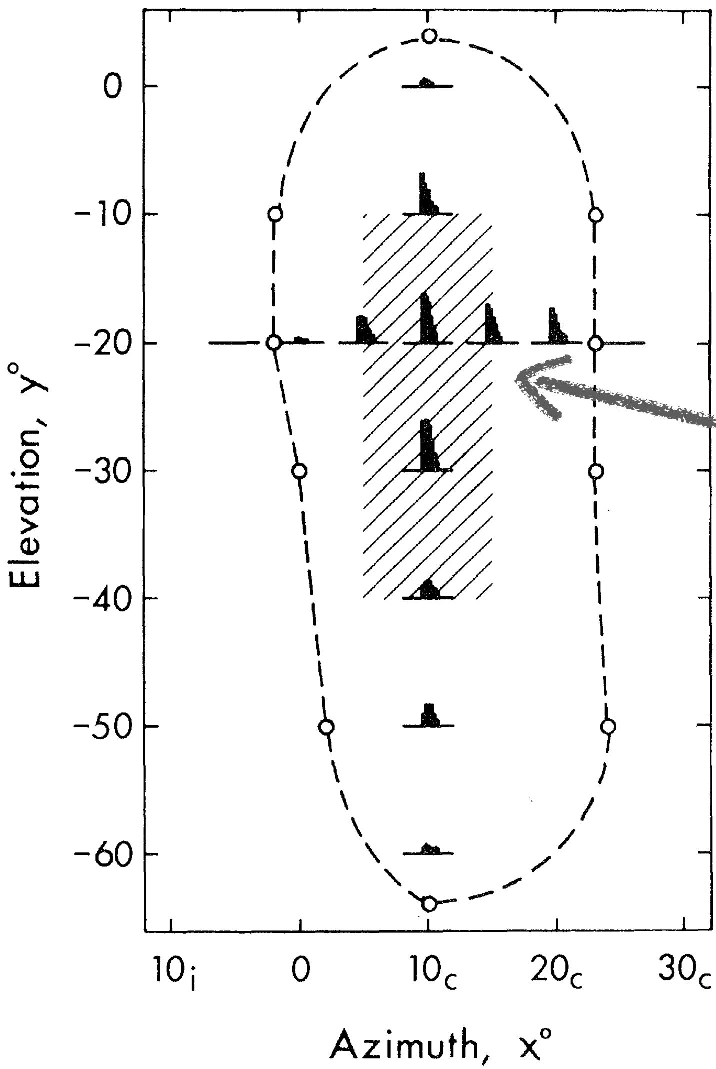
www.icsi.berkeley.edu/eecs225d/spr12/

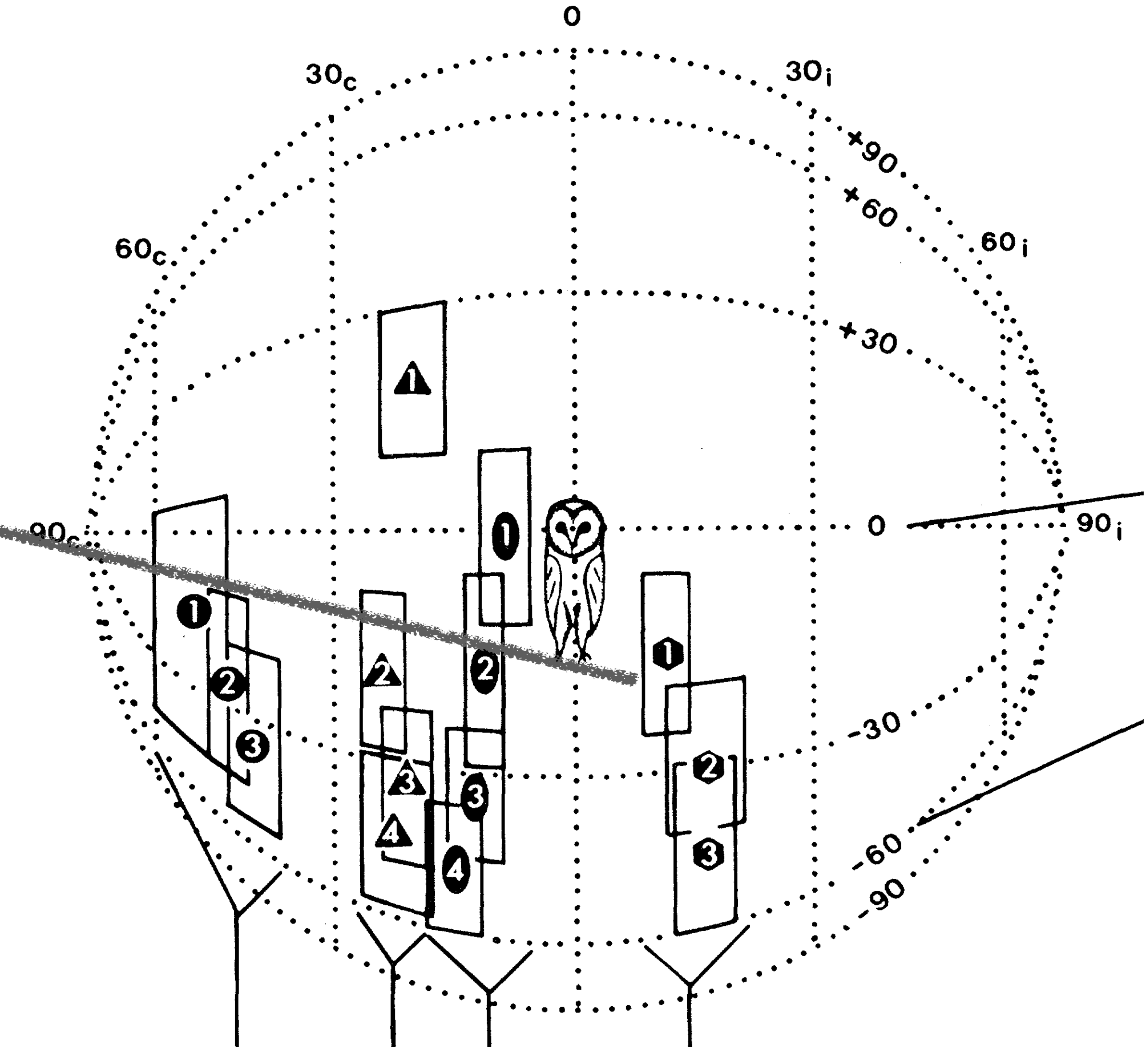
Today's lecture: Source Separation

- Two approaches to the problem ...
  - Auditory scene analysis
  - Microphone array techniques
- Research project ideas ...



Knudsen and Konishi

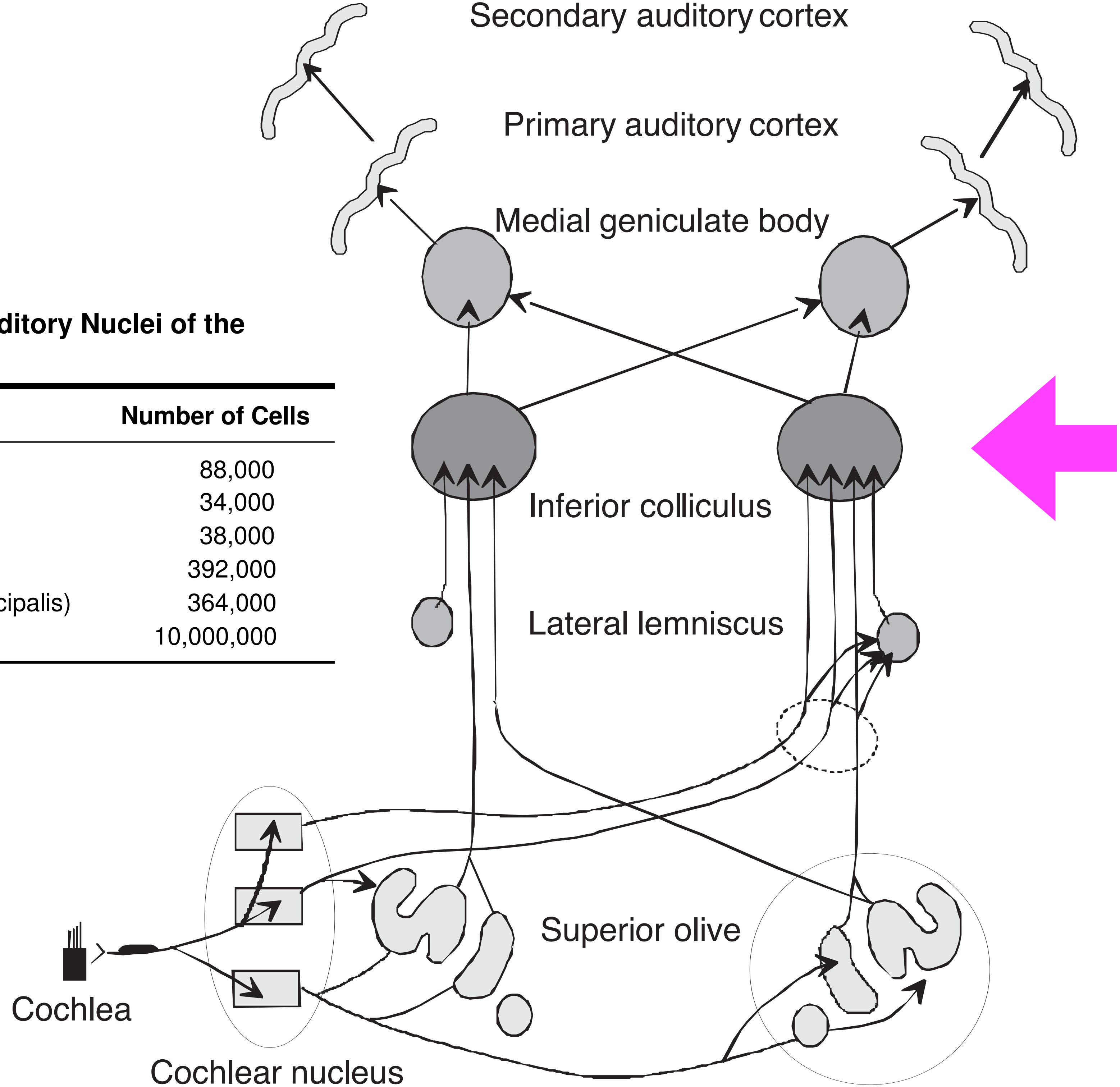




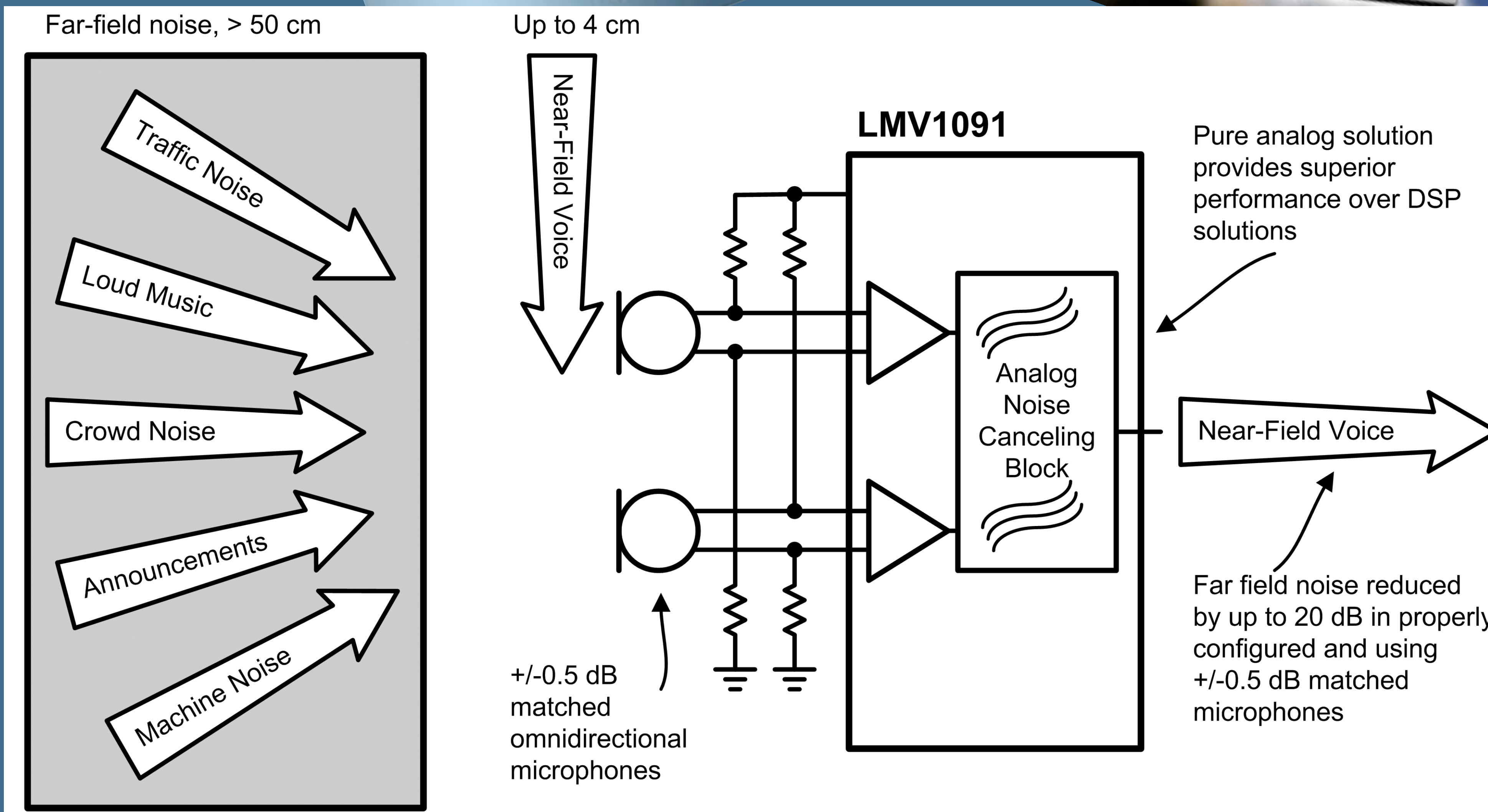
Knudsen and Konishi

TABLE 14.1 Cells in the Auditory Nuclei of the Monkey[a]

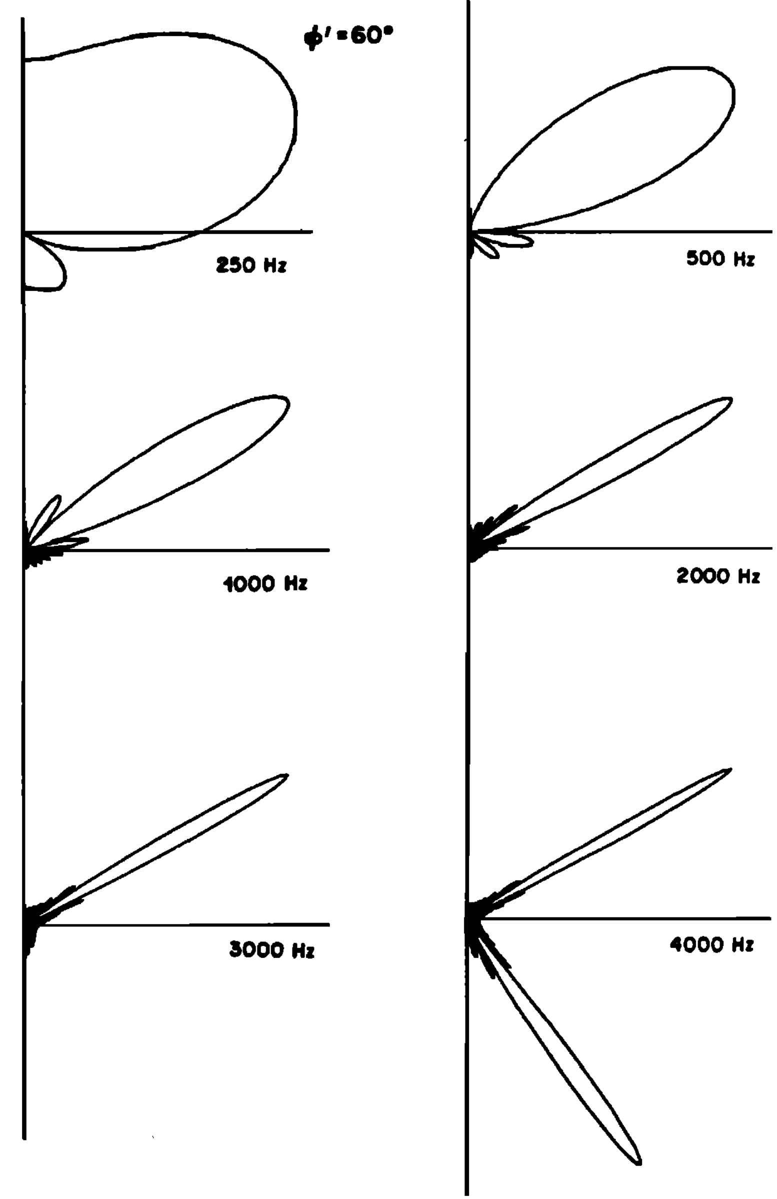
Central Auditory Nucleus	Number of Cells
Cochlear nuclei	88,000
Superior olivary complex	34,000
Nuclei of lateral lemniscus	38,000
Inferior colliculus	392,000
Medial geniculate body (pars principalis)	364,000
Auditory cortex	10,000,000

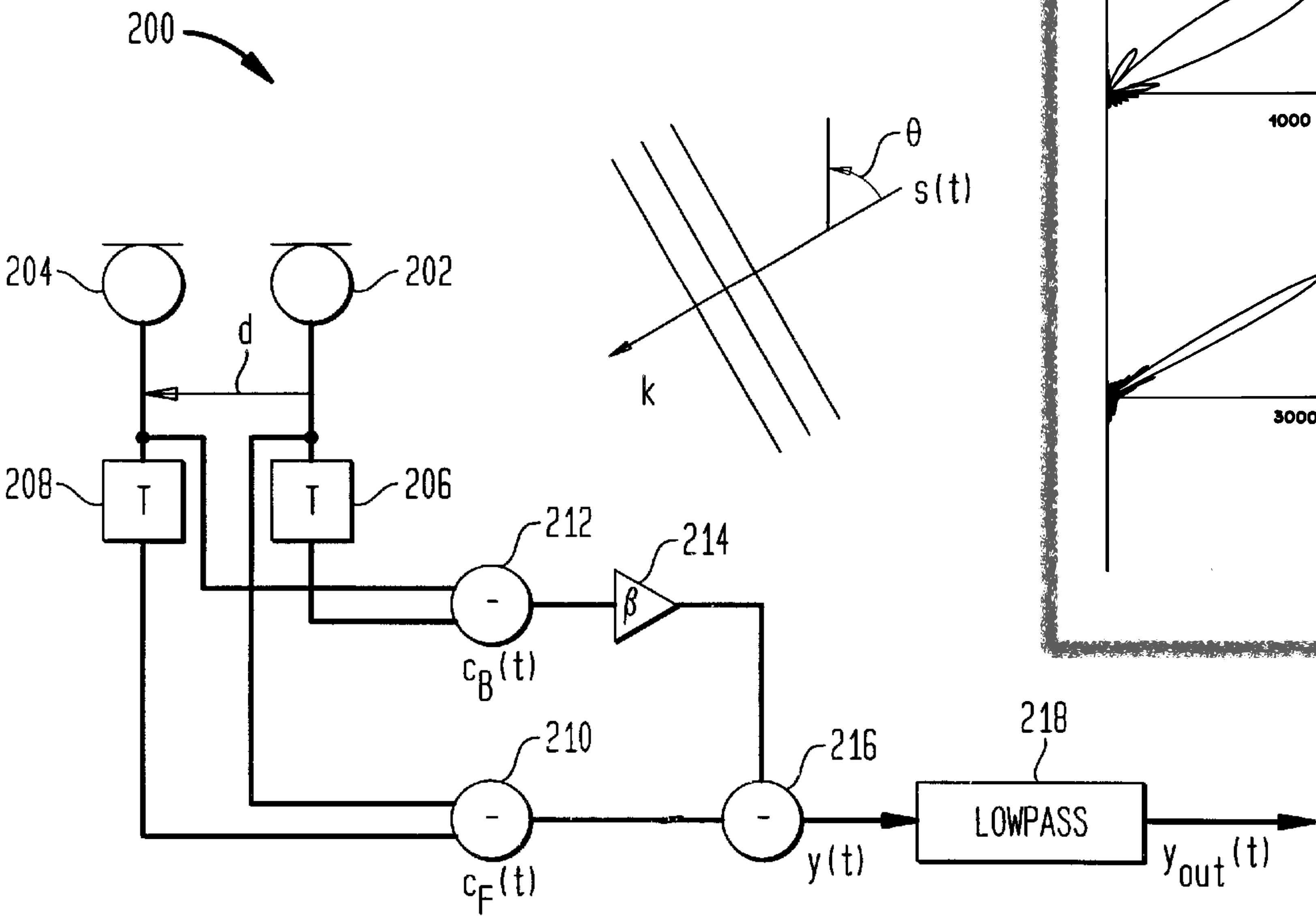



National Semiconductor
LMV1088


Far-field noise, > 50 cm
Traffic Noise
Loud Music
Crowd Noise
Announcements
Machine Noise
Up to 4 cm
Near-Field Voice
LMV1091
Analog Noise Canceling Block
Pure analog solution provides superior performance over DSP solutions
Near-Field Voice
+/-0.5 dB matched omnidirectional microphones
Far field noise reduced by up to 20 dB in properly configured and using +/-0.5 dB matched microphones

$$d^n_{mm'}(\beta) = \xi_{mm'}\sqrt{\frac{s!(s+\mu+\nu)!}{(s+\mu)!(s+\nu)!}}$$
$$\times \sin\left(\frac{\beta}{2}\right)^{\mu} \cos\left(\frac{\beta}{2}\right)^{\nu} P_s^{(\mu,\nu)}(\cos\beta)$$

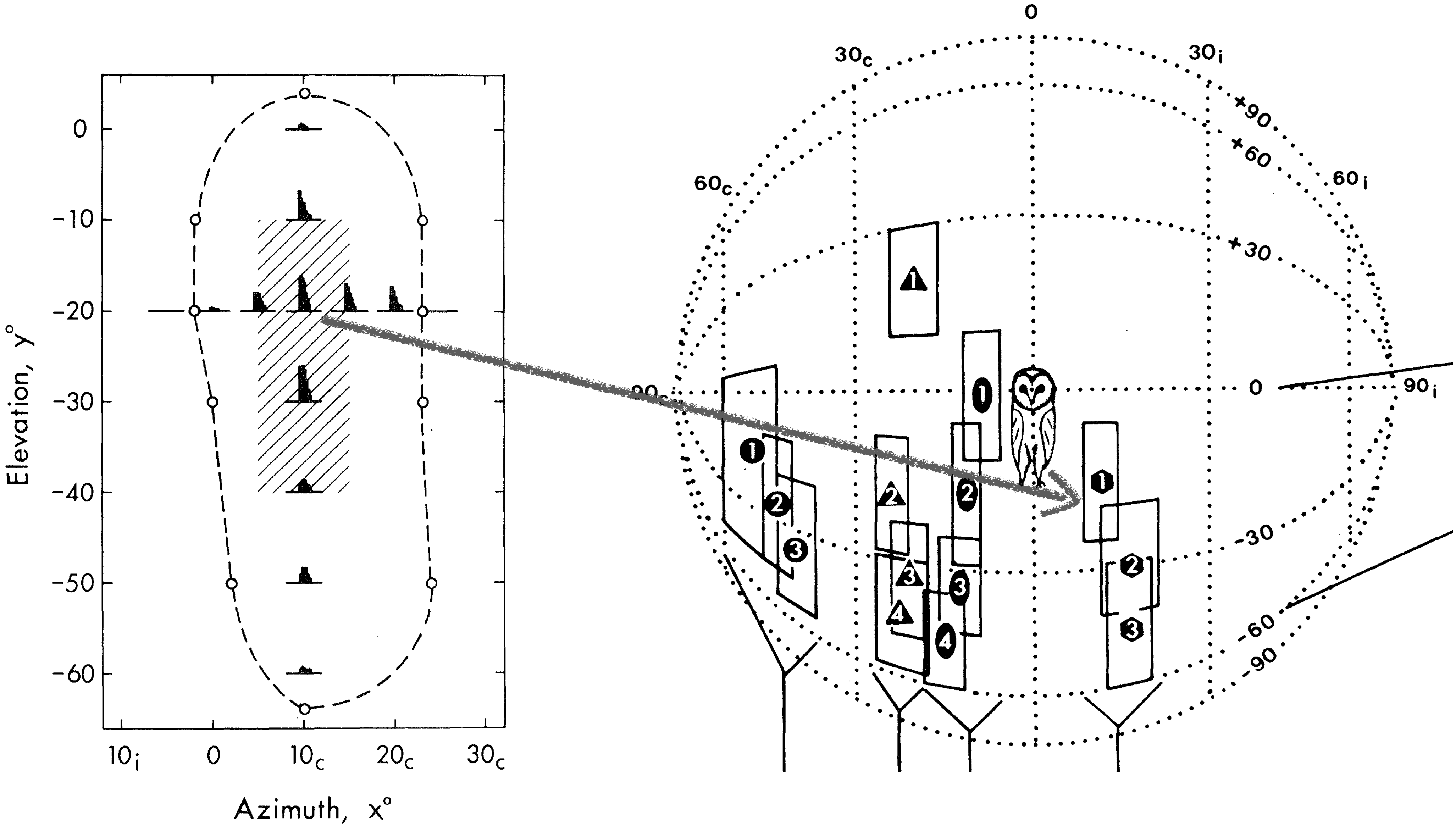




Today's lecture: Source Separation

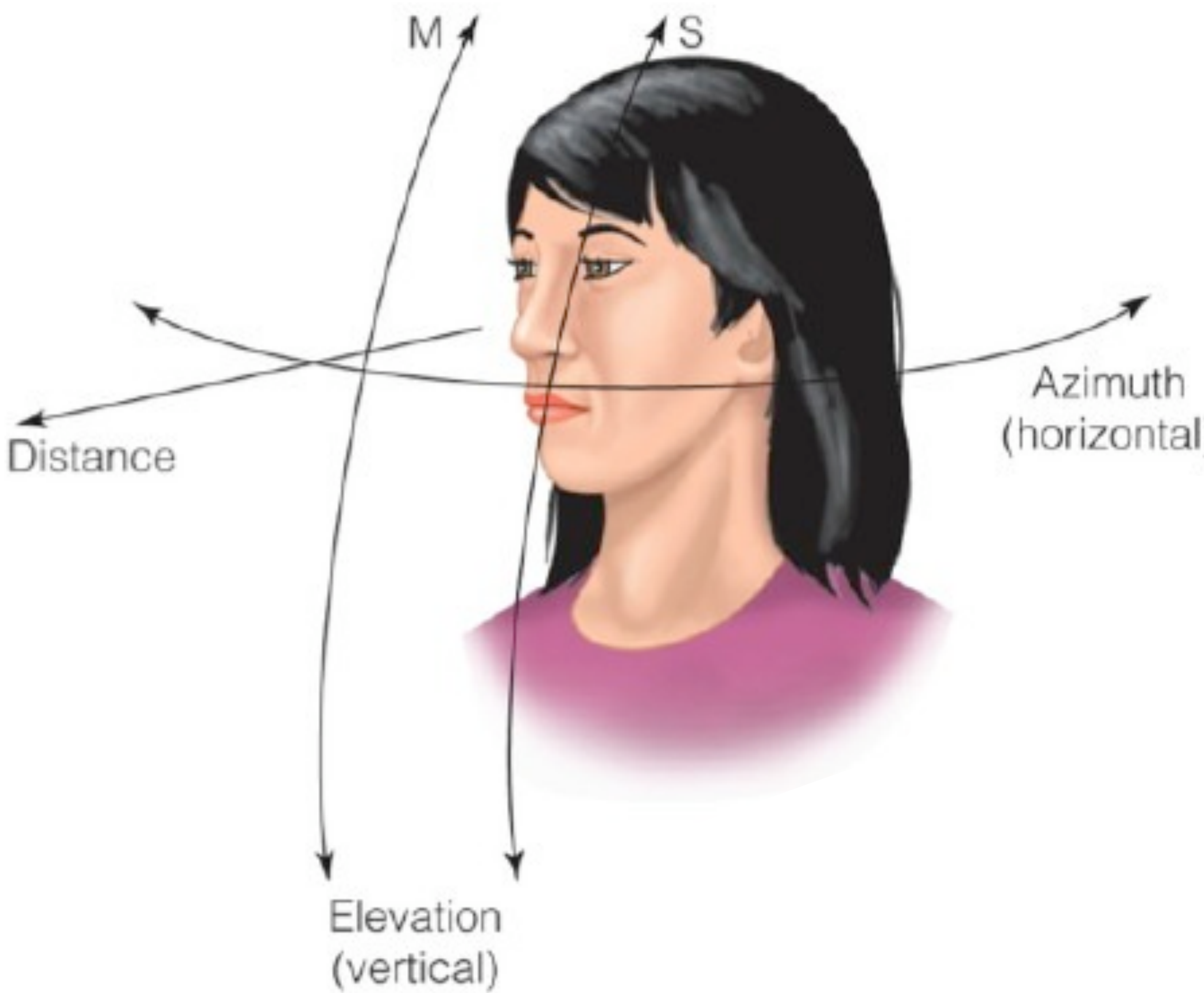
- Two approaches to the problem ...
  - **Auditory scene analysis**
  - Microphone array techniques
- Research project ideas ...

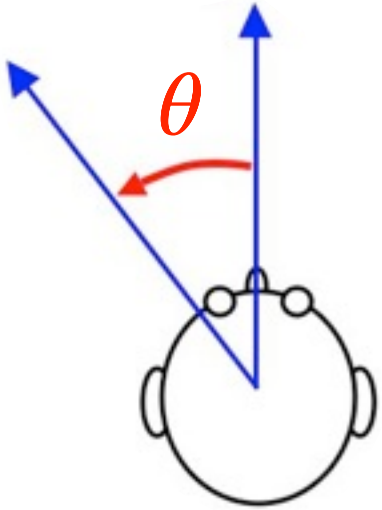
Where do we begin to understand how auditory spatial maps are computed in the brain?



Spatial hearing cues ...

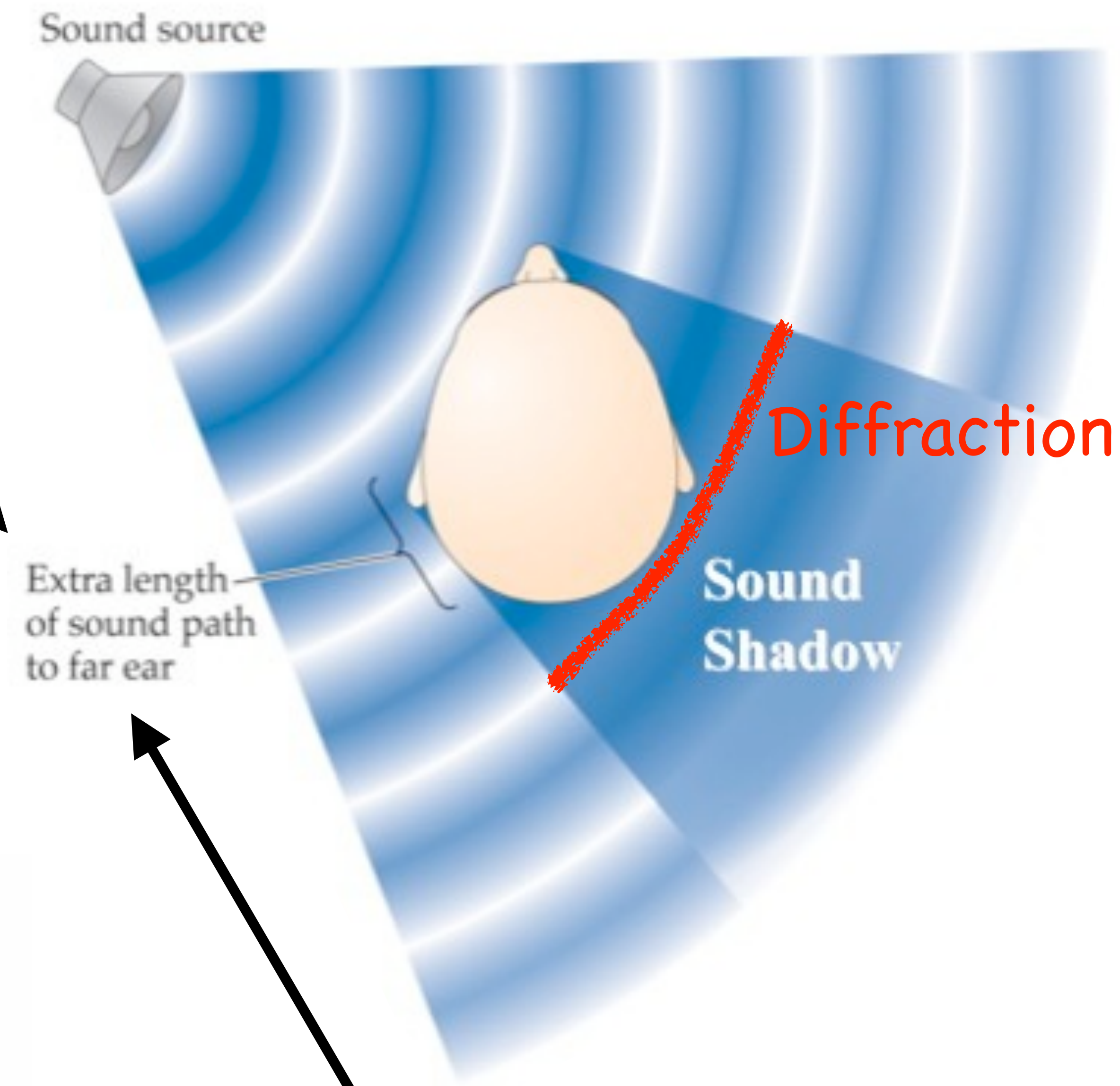
The story begins with acoustics of the head ...

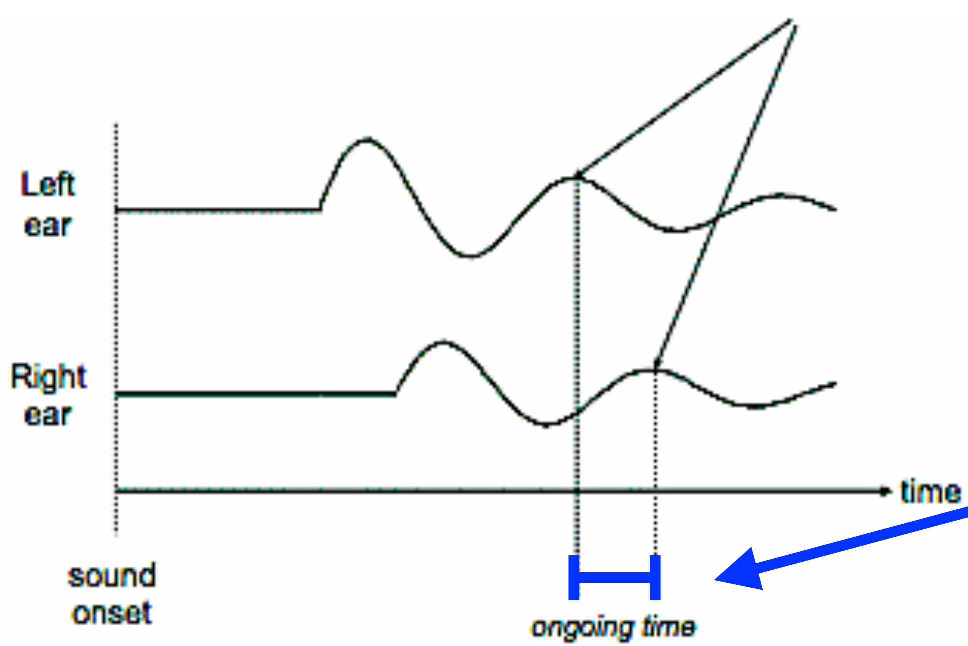




(Azimuth)

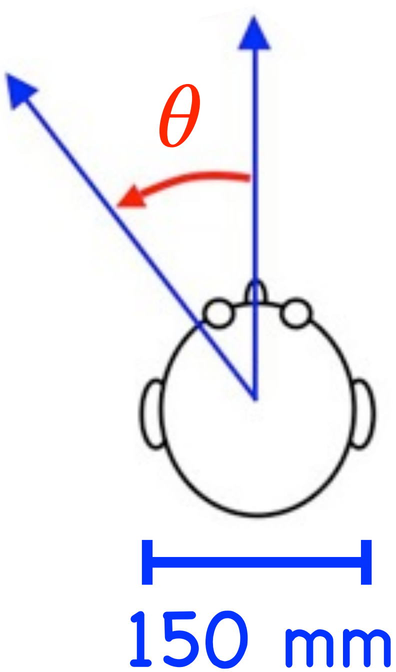
Trigonometry maps θ to "extra length"

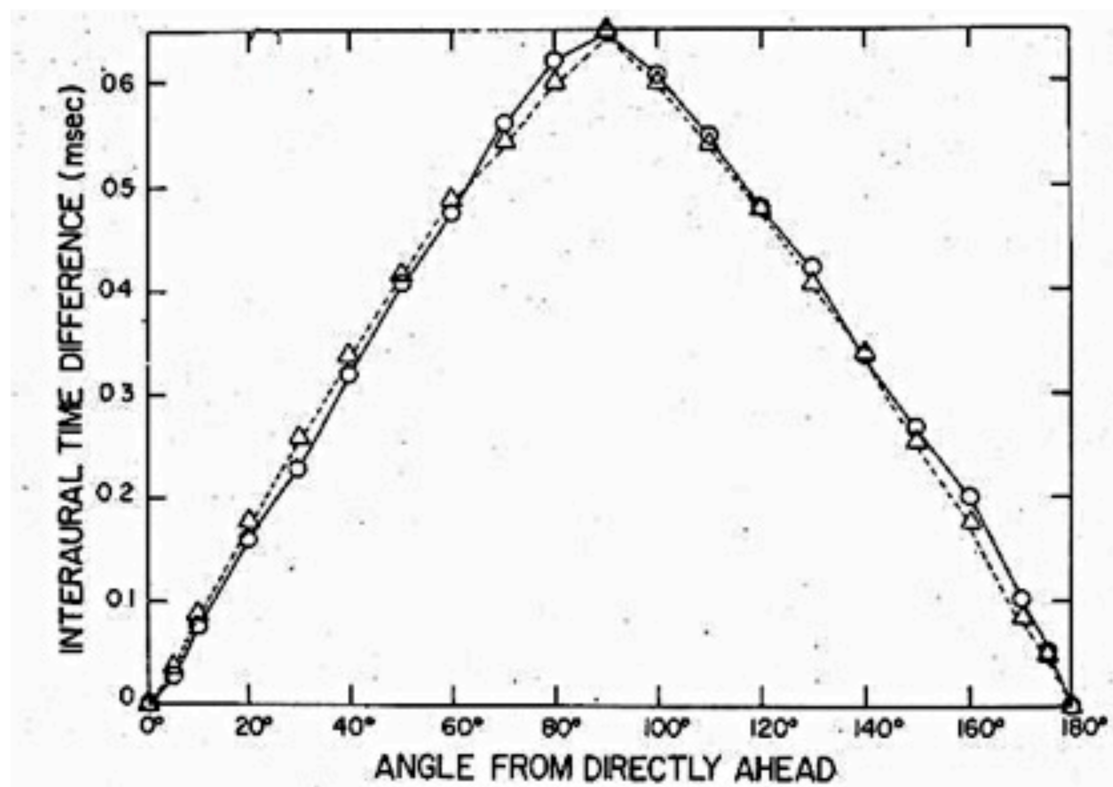




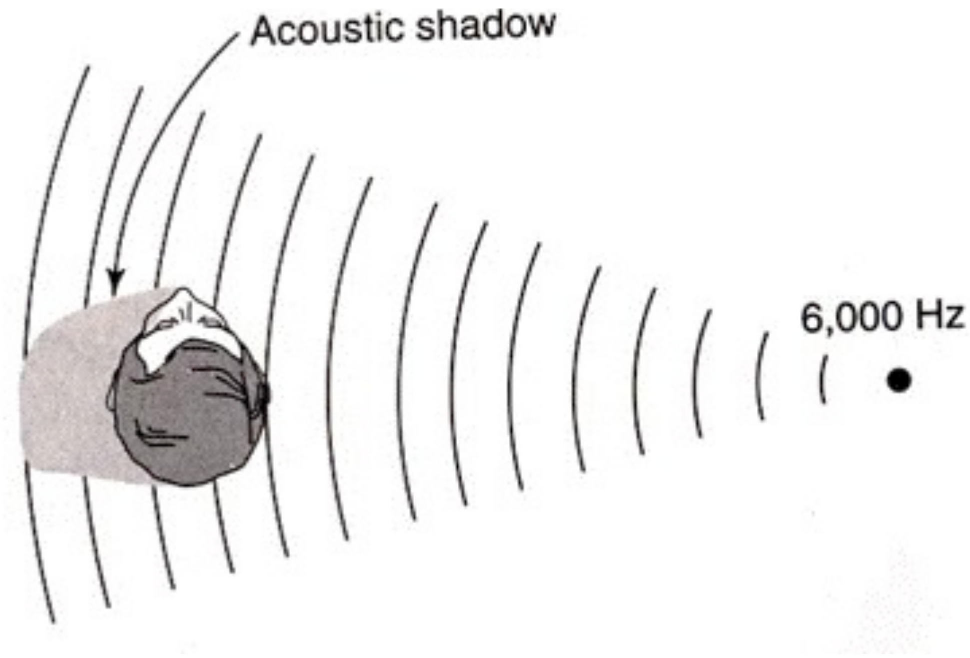
Speed of sound determines time difference that corresponds to the "extra length".

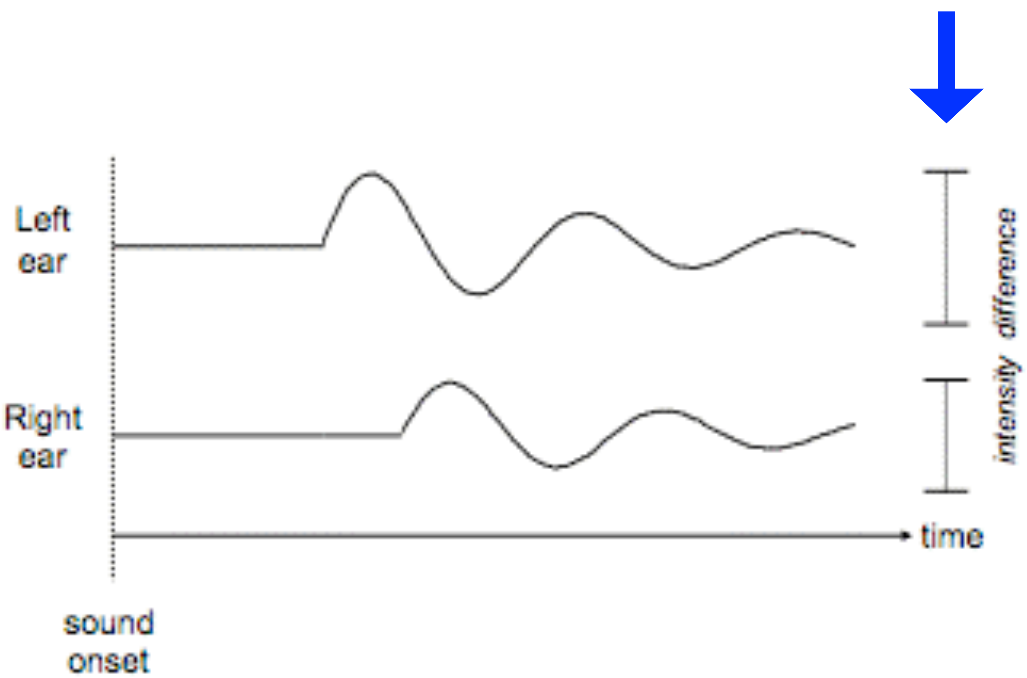
150 mm average human head size limits the interaural timing cue to 600 μs, corresponding to detecting the phase of a 1.7 kHz sine wave

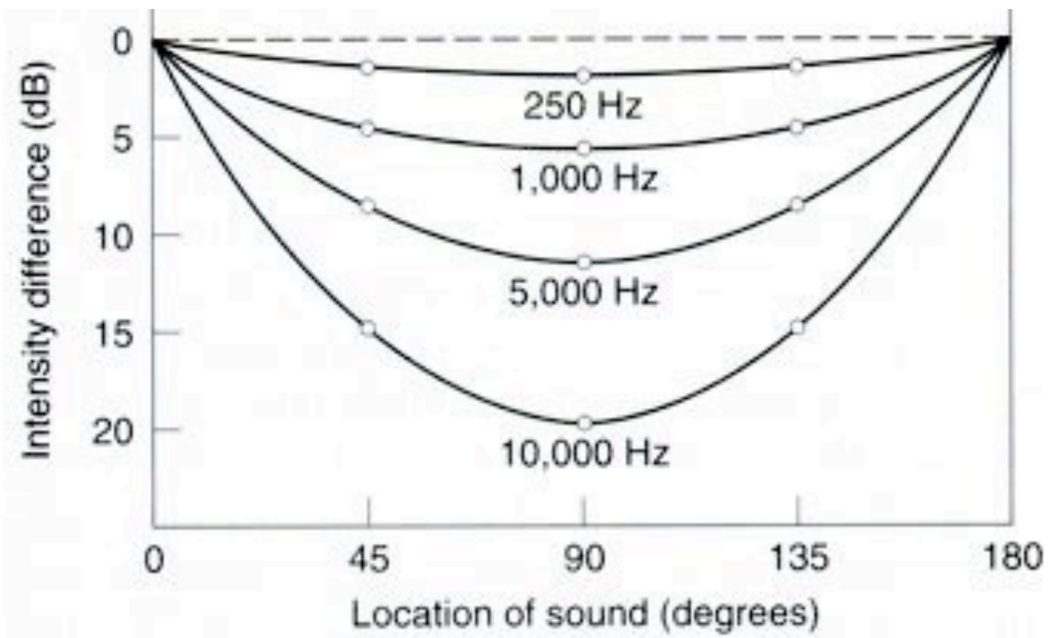


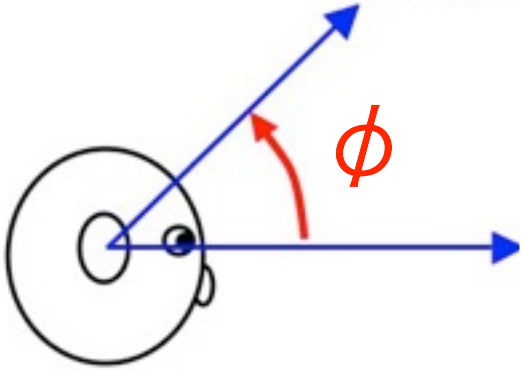


For kHz sound energy, azimuth can be computed by comparing interaural intensity cues generated by the head shadow effect.

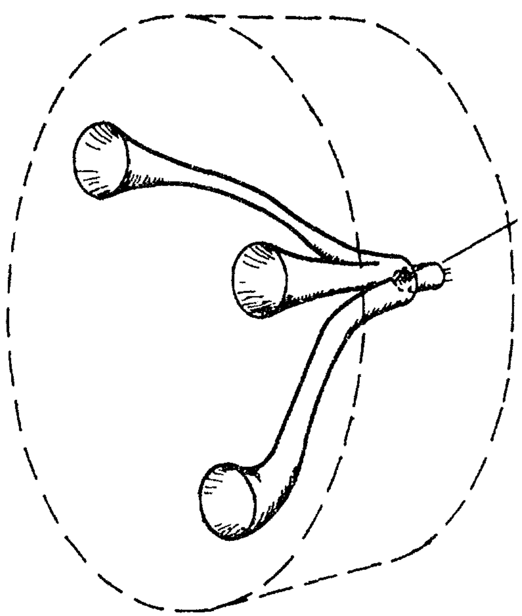


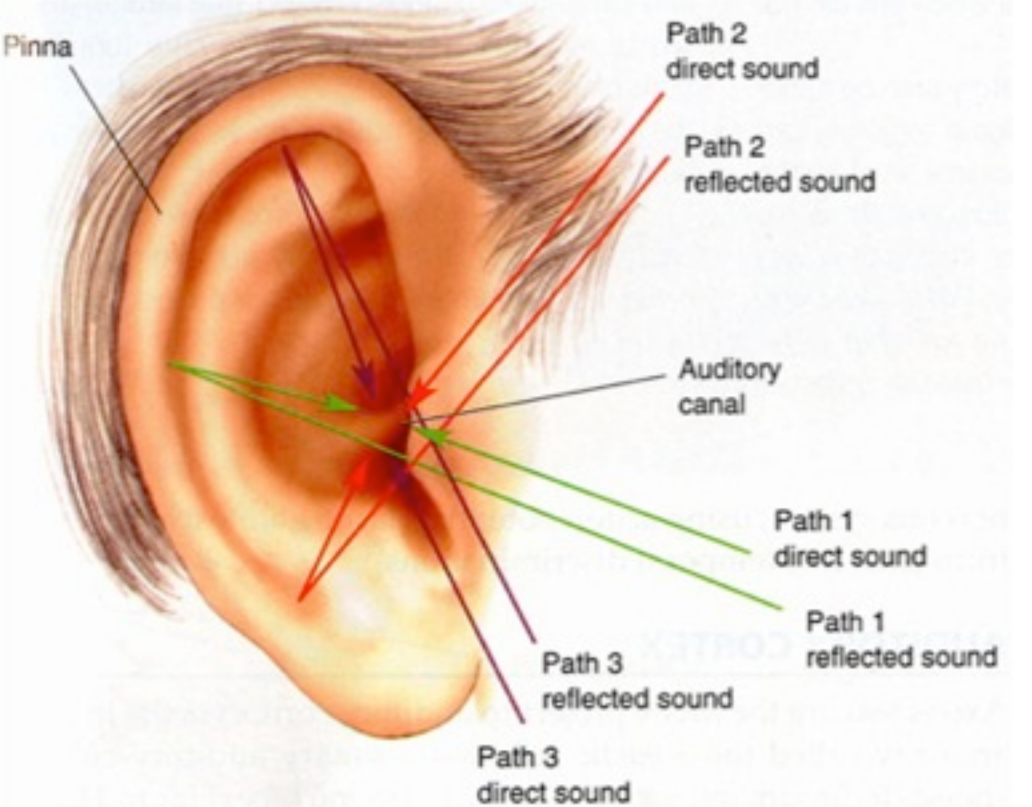


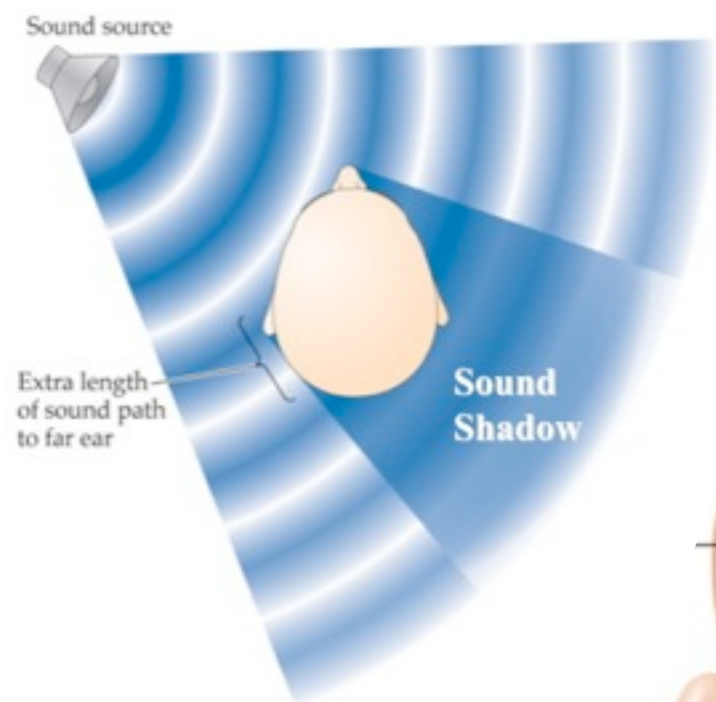




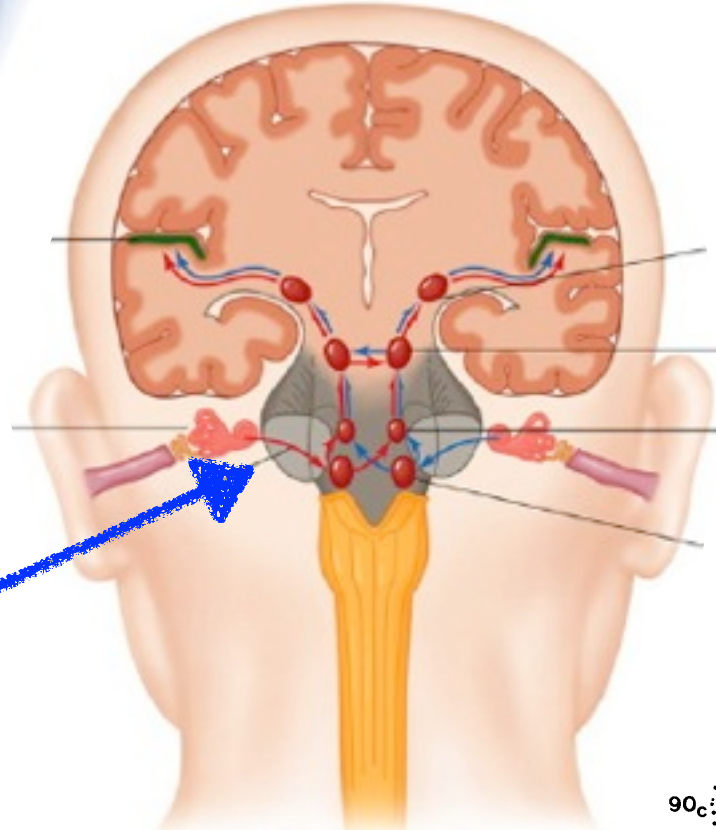
Elevation cues are coded by acoustic comb filtering in the outer ear.

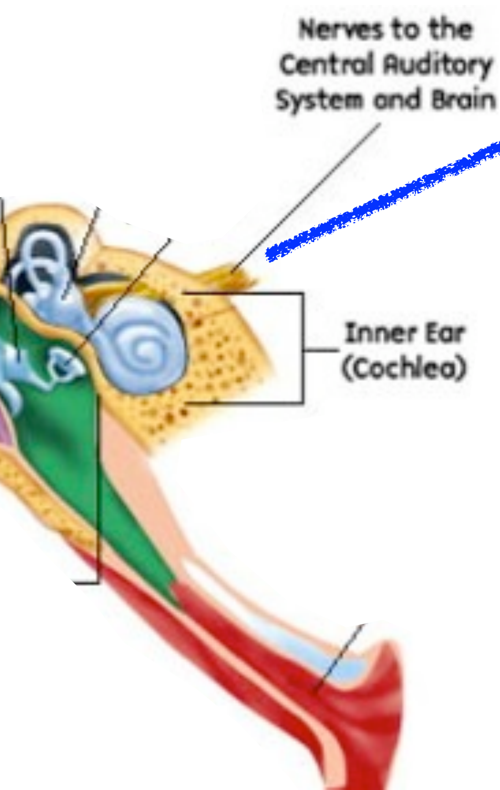




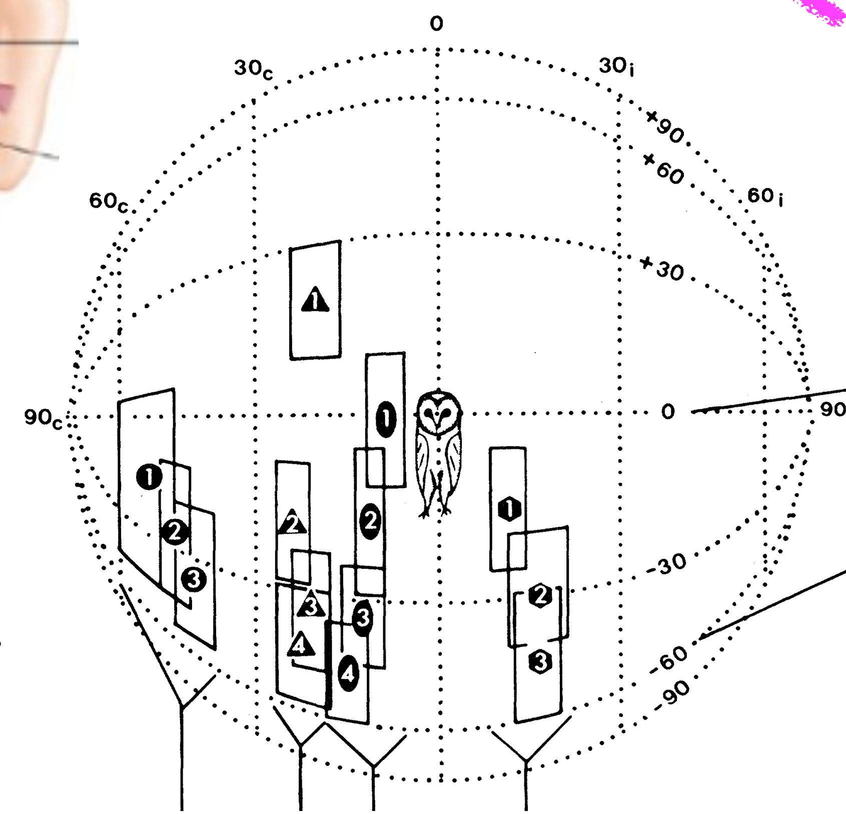


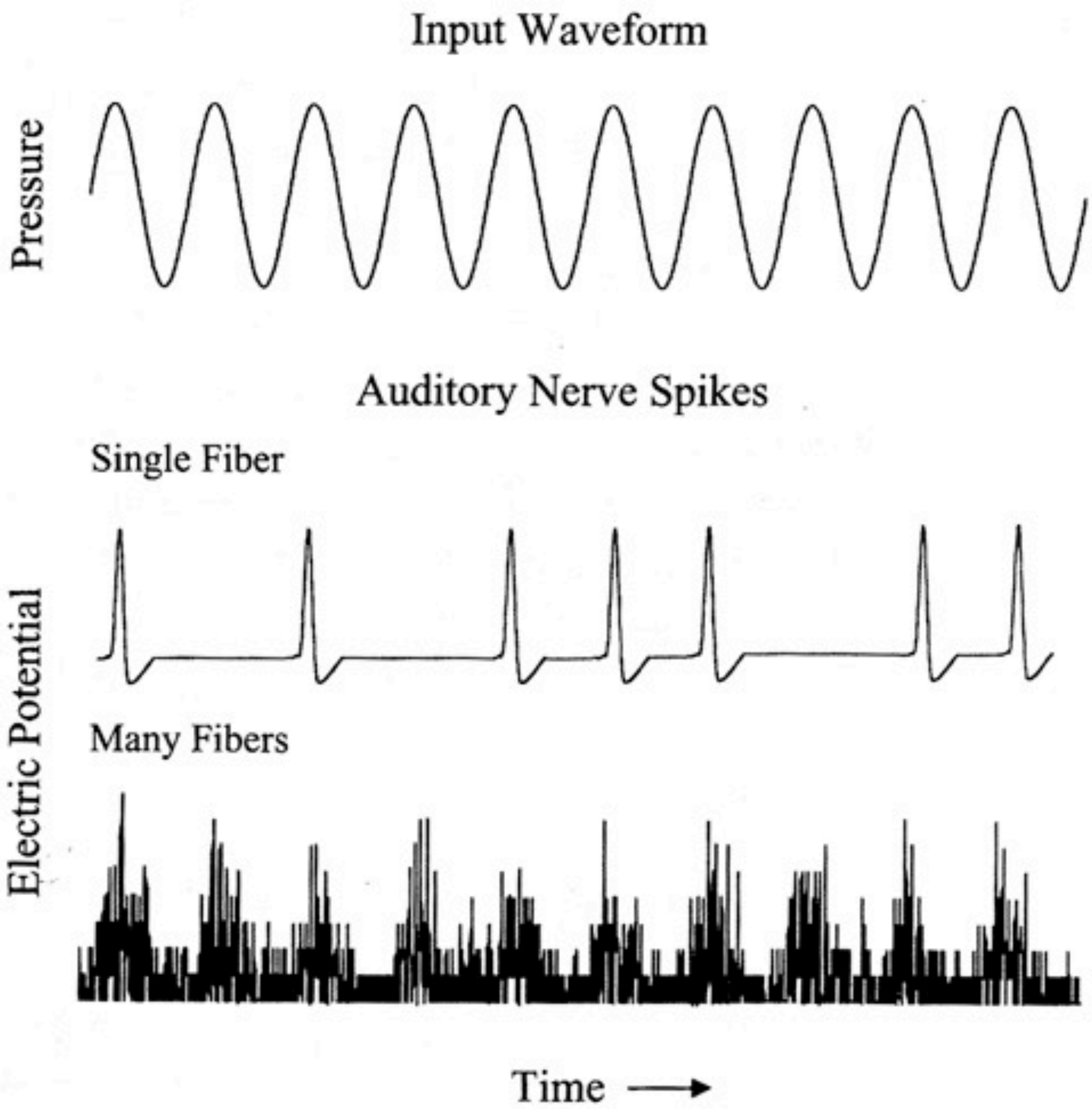
Q. How does the brain compute neural maps of space from acoustic cues ?





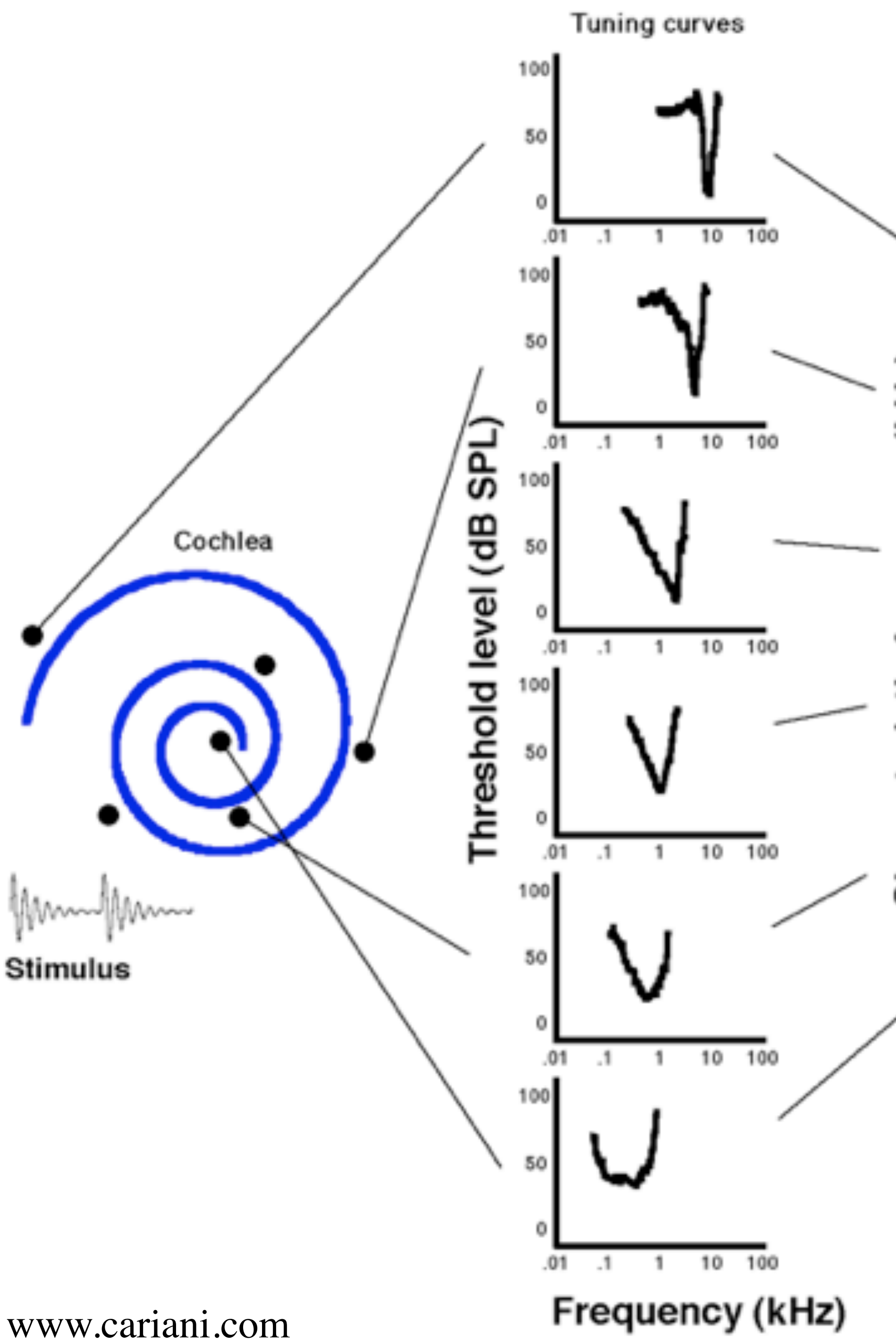
A. By processing the data coded by the auditory nerve.

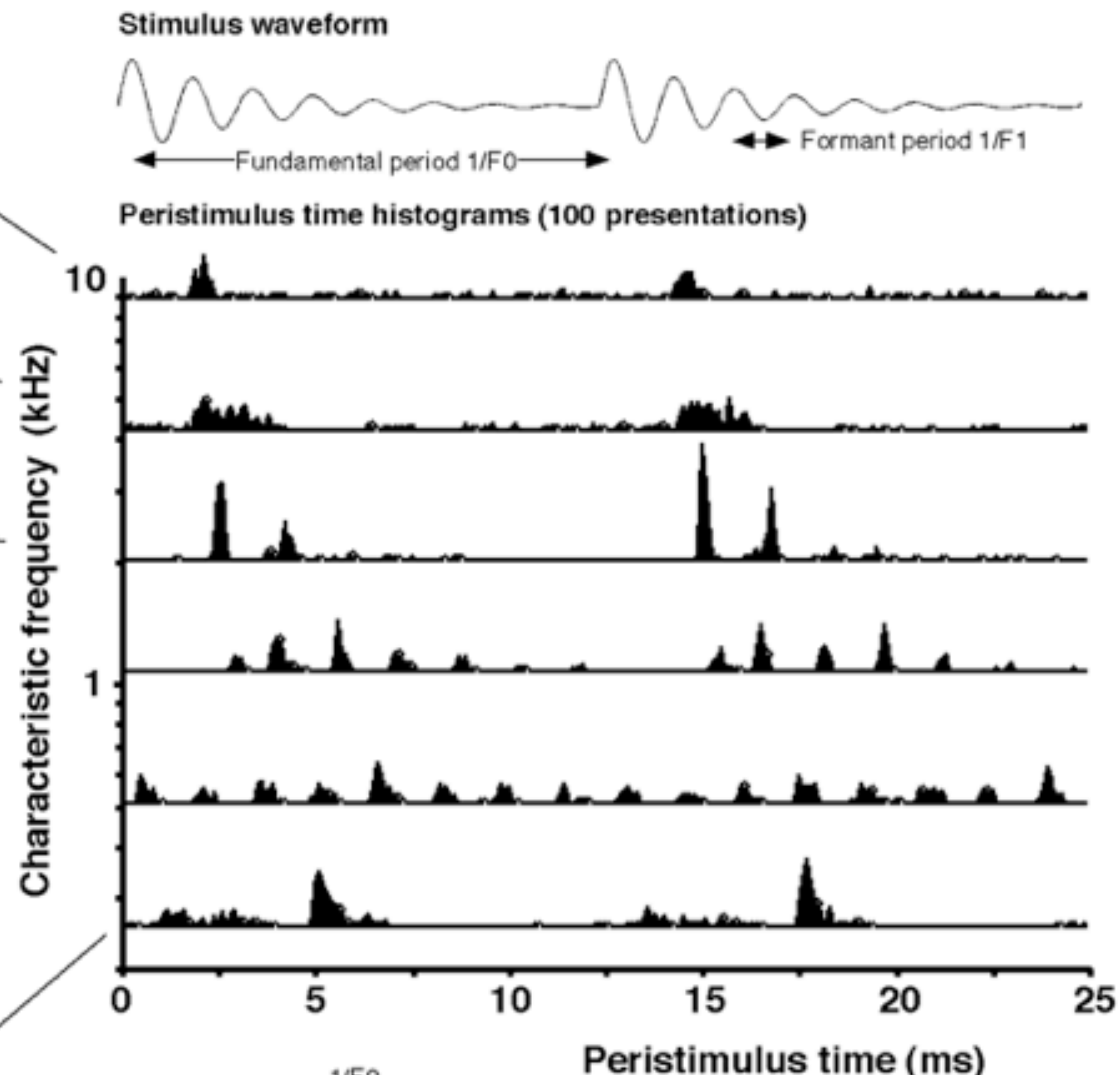




Cycle-by-cycle acoustic waveform shape can be reconstructed from the spike trains of multiple auditory nerve fibers.

Auditory Nerve: A "neural microphone" up to multi-kHz.





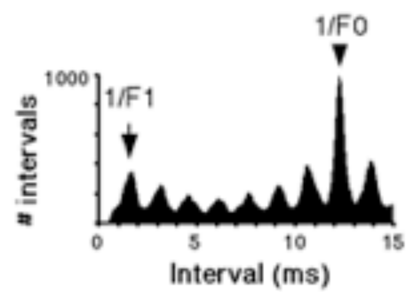
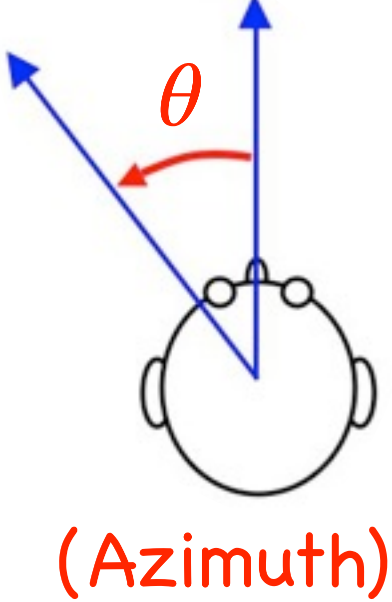
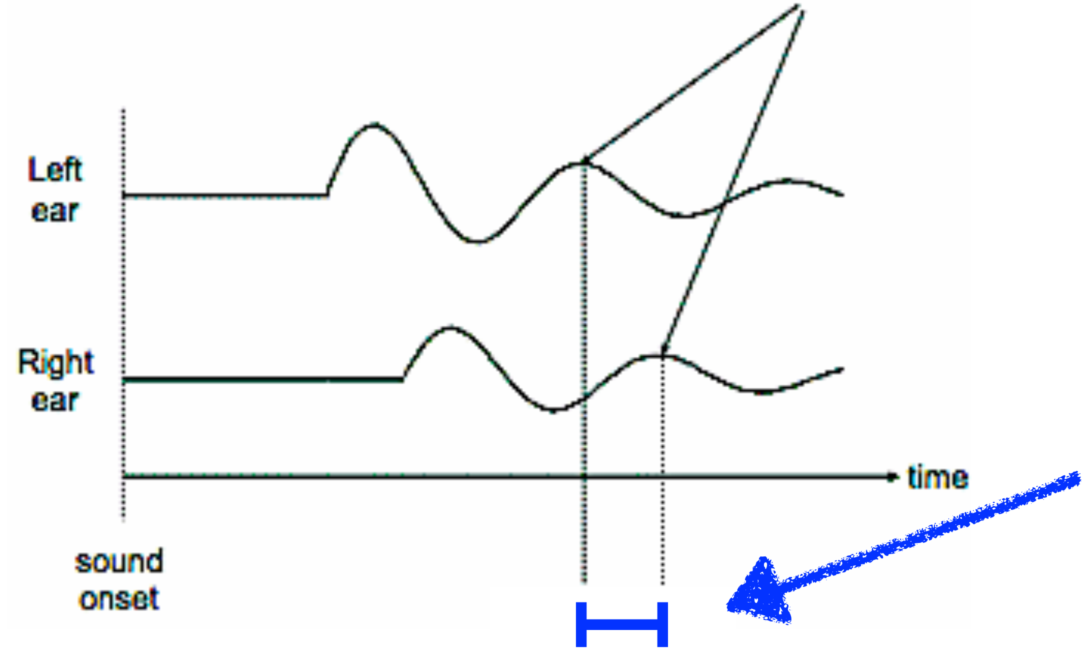
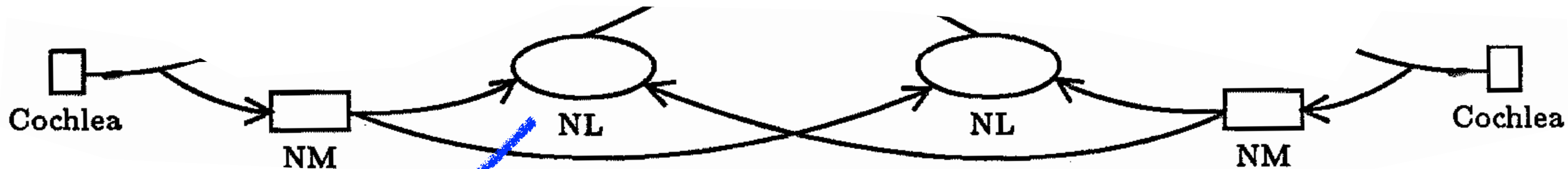


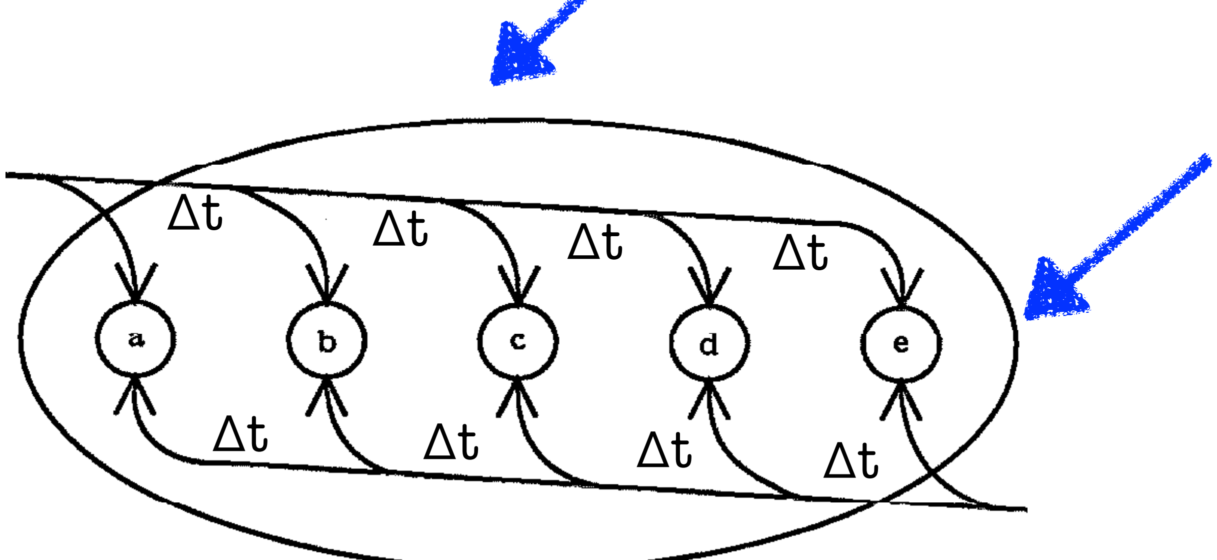
Figure 5. Temporal patterns in the auditory nerve. Tuning curves and peristimulus time (PST) histograms for six representative auditory nerve fibers with different characteristic frequencies. Left: All-order interspike interval histogram for a population of 79 auditory nerve fibers.



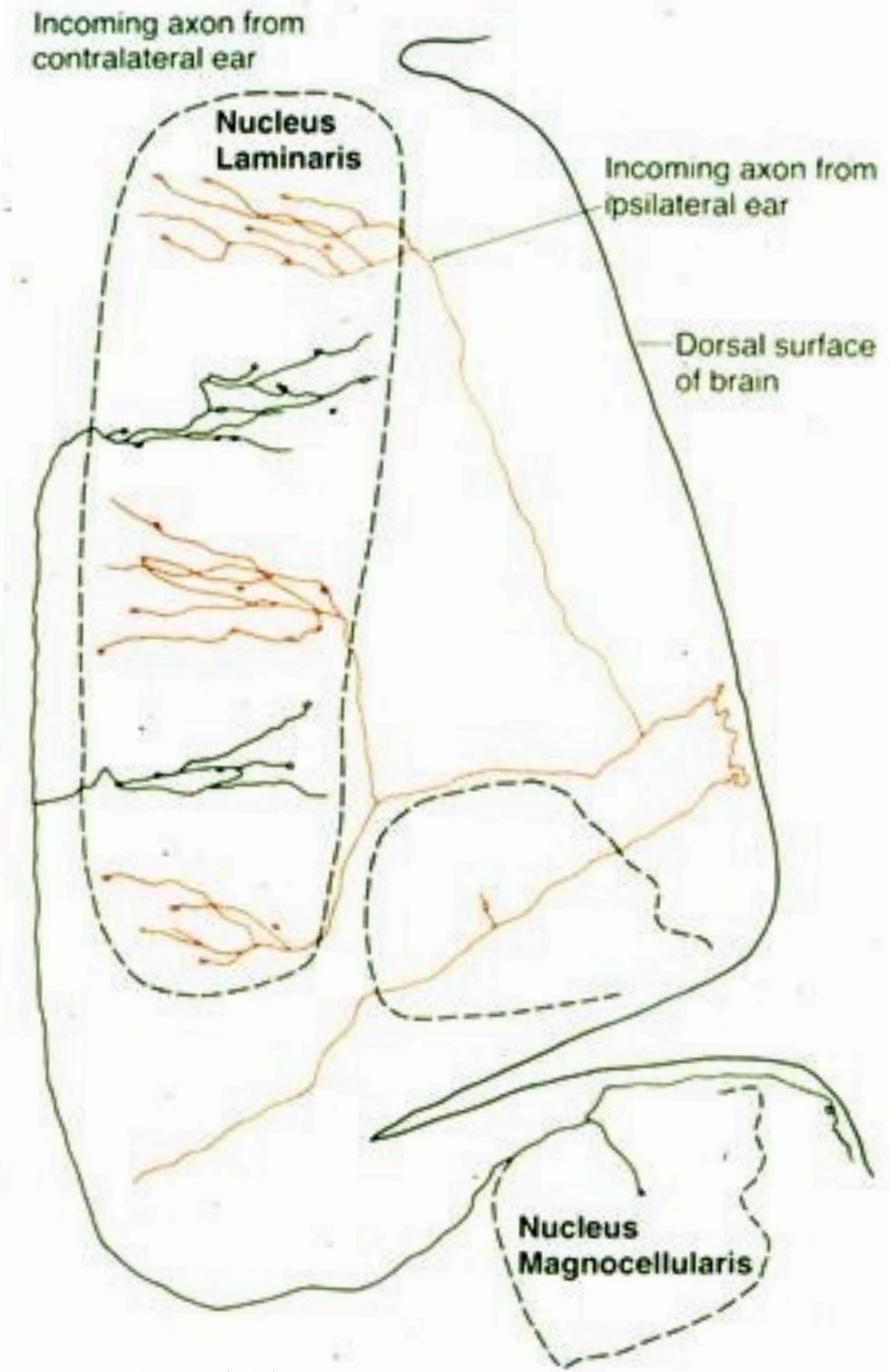


To compute the time delay, cross-correlate the waveforms encoded by the auditory nerve of the left and right cochlea.



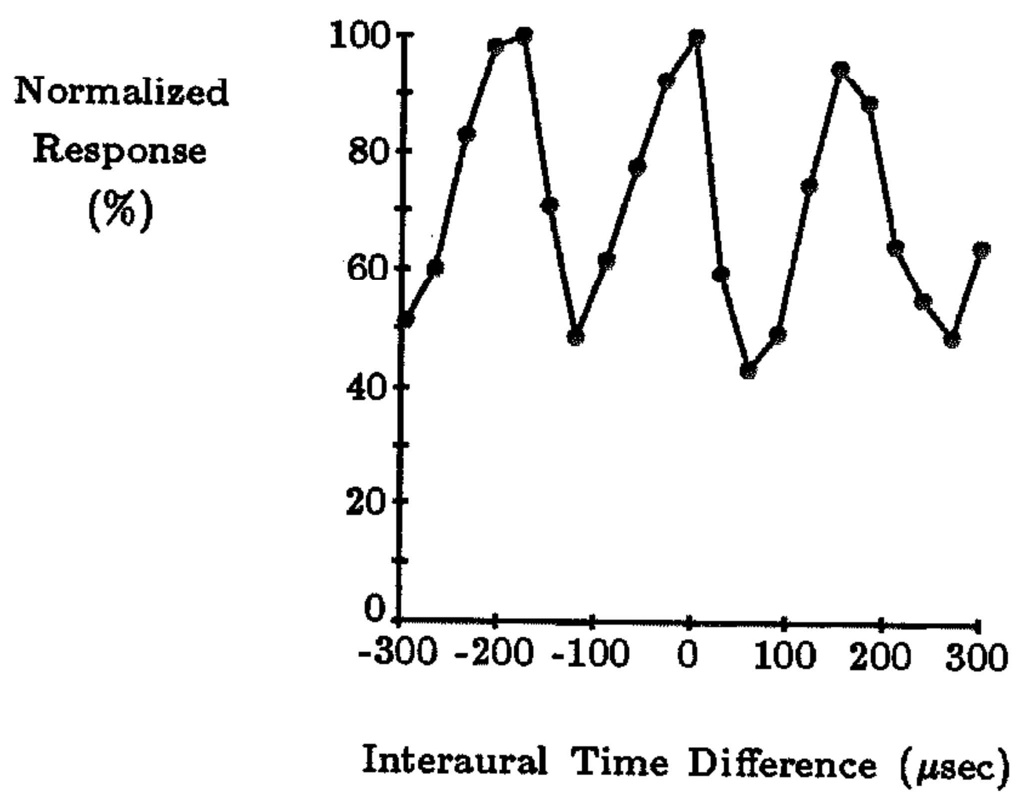


Nucleus Laminaris: Neural structure in owl auditory pathway that appears to be a spike cross-correlation module (Jeffress model).



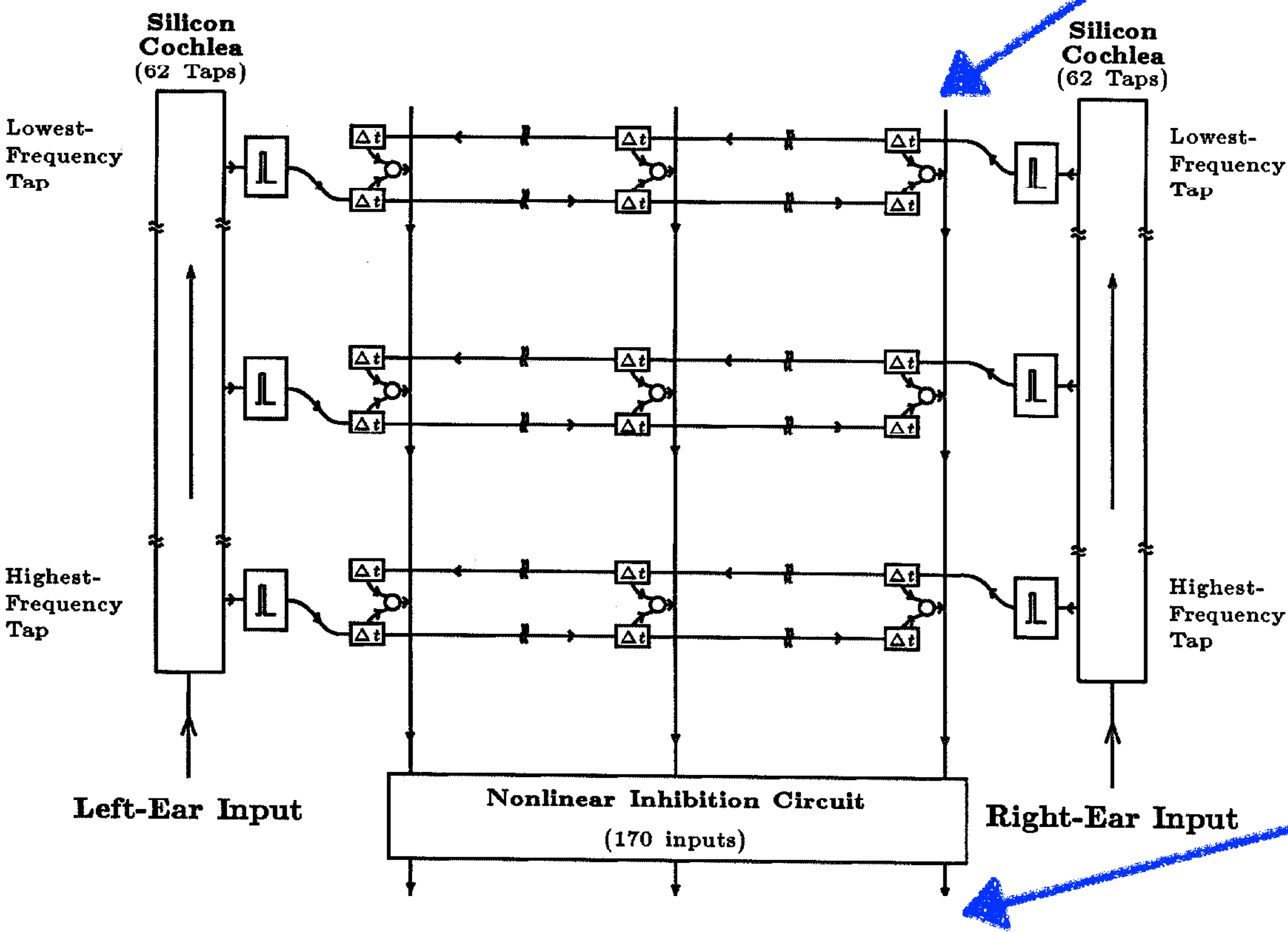
Carr and Konishi

Anatomy (left) and neurophysiology (below) that show the plausibility of cross-correlation structures in the auditory brainstem for interaural time delay computation.



Fujita and Konishi

Engineers borrowed this concept, and built computational auditory maps of interaural time differences ...



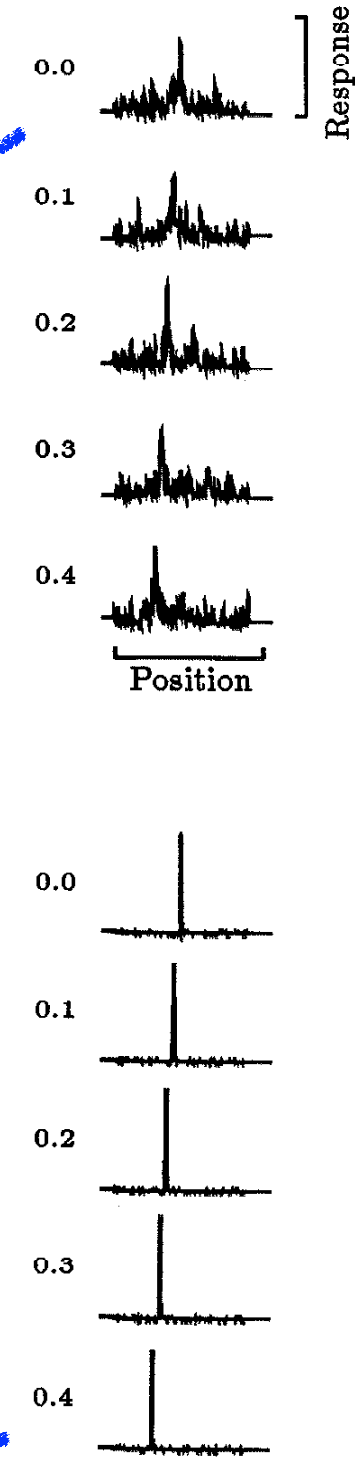
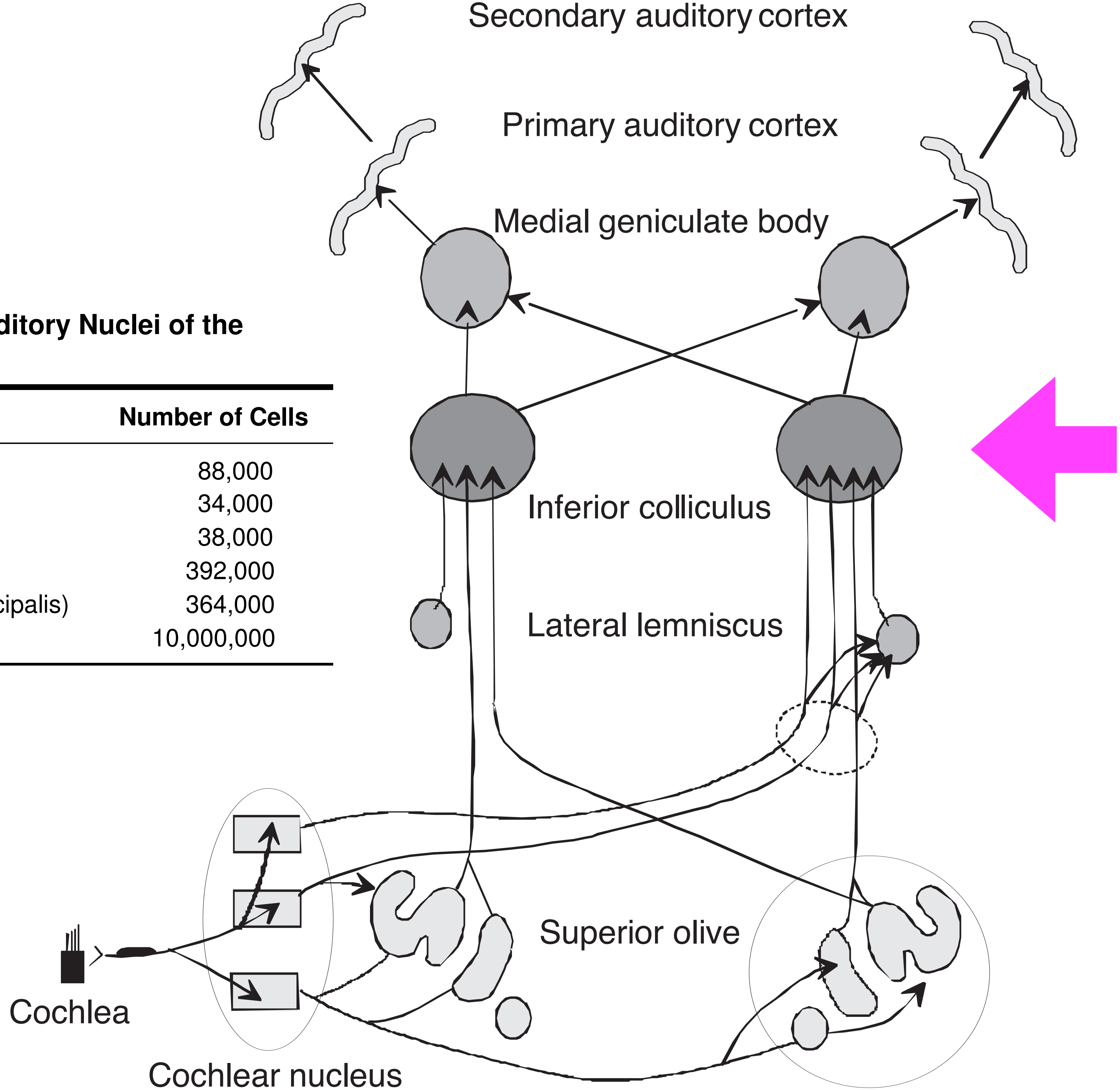


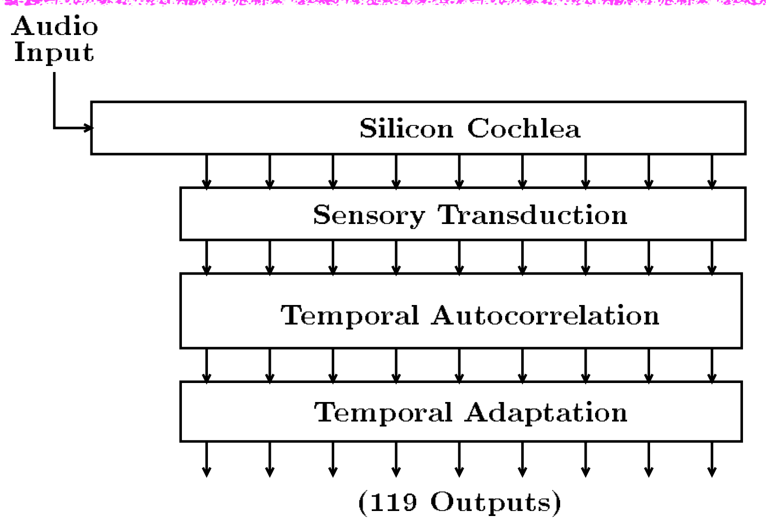
TABLE 14.1 Cells in the Auditory Nuclei of the Monkey[a]

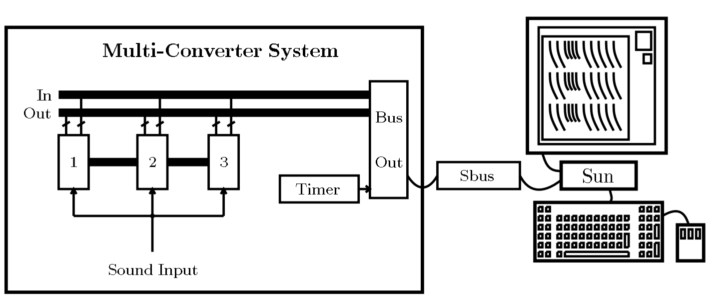
Central Auditory Nucleus	Number of Cells
Cochlear nuclei	88,000
Superior olivary complex	34,000
Nuclei of lateral lemniscus	38,000
Inferior colliculus	392,000
Medial geniculate body (pars principalis)	364,000
Auditory cortex	10,000,000

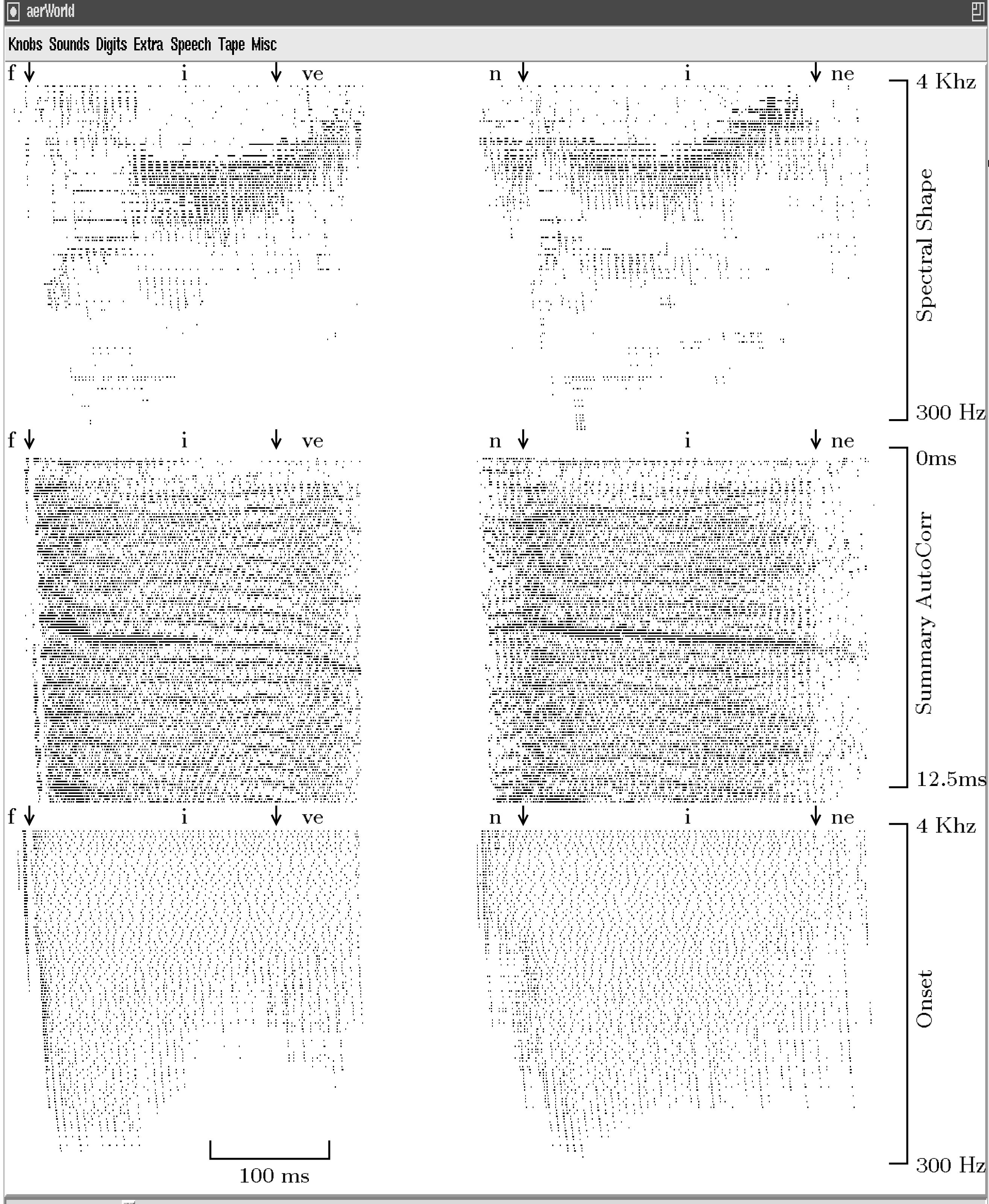


... and multi-map systems to compute periodicity, onsets, spectral shape, etc.

"Turn audition into the vision problem"

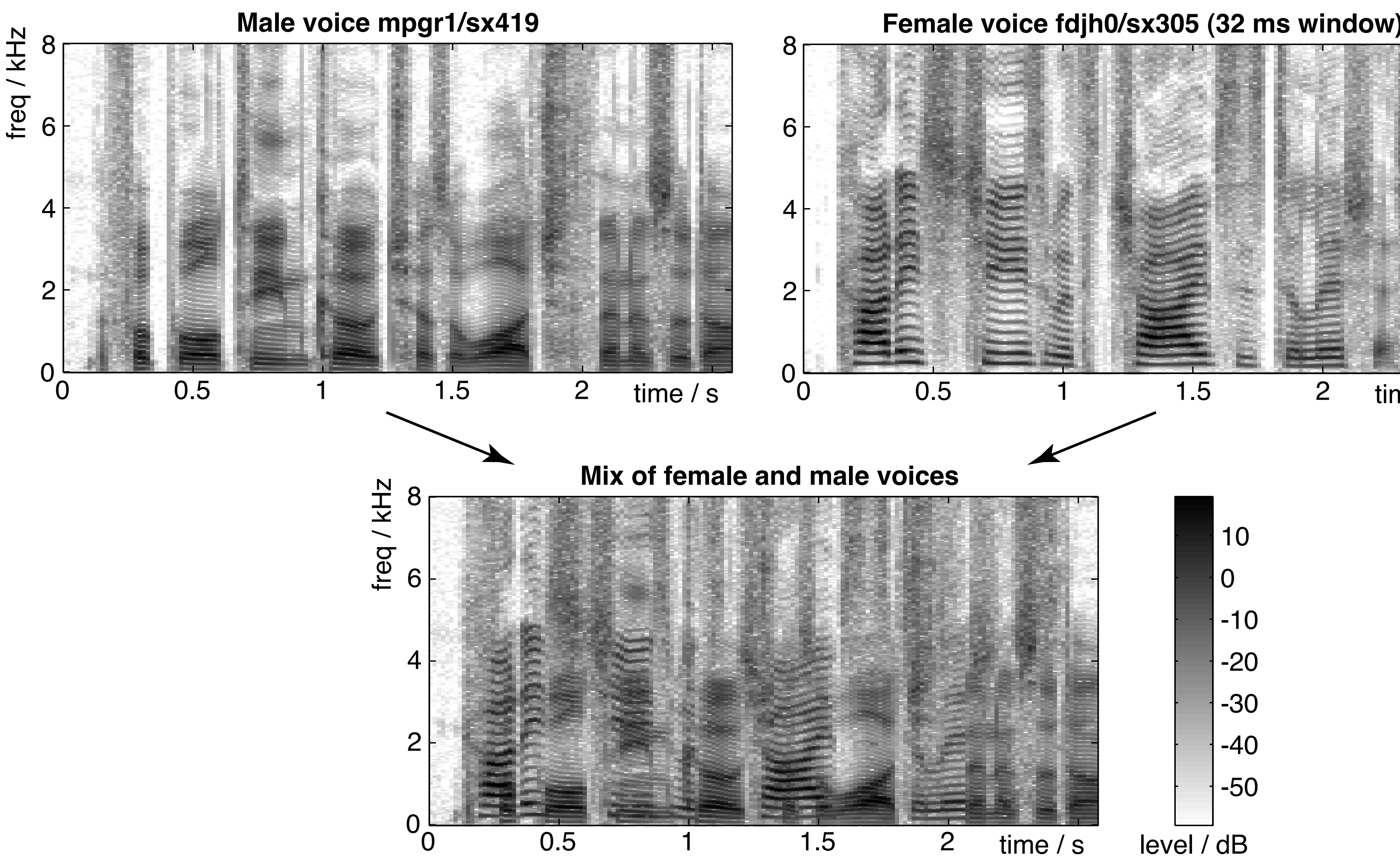




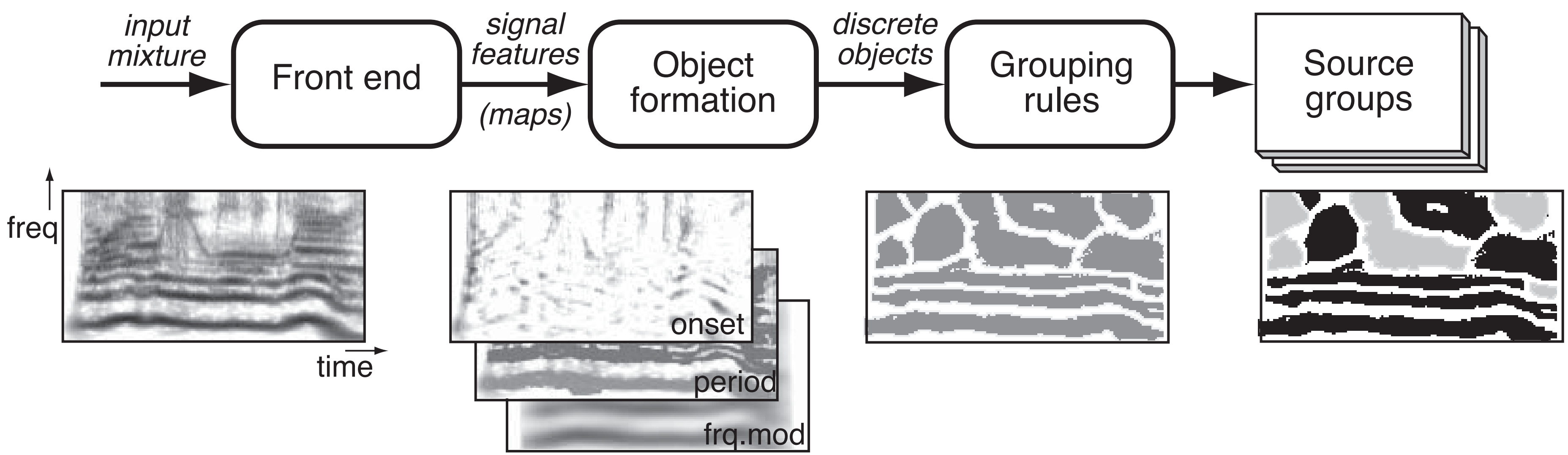


Computational Auditory Scene Analysis

Typical problem: Separate 2 voices captured by 1 microphone.



… by using a auditory scene analysis pipeline that is inspired by the Gestalt school of visual processing.



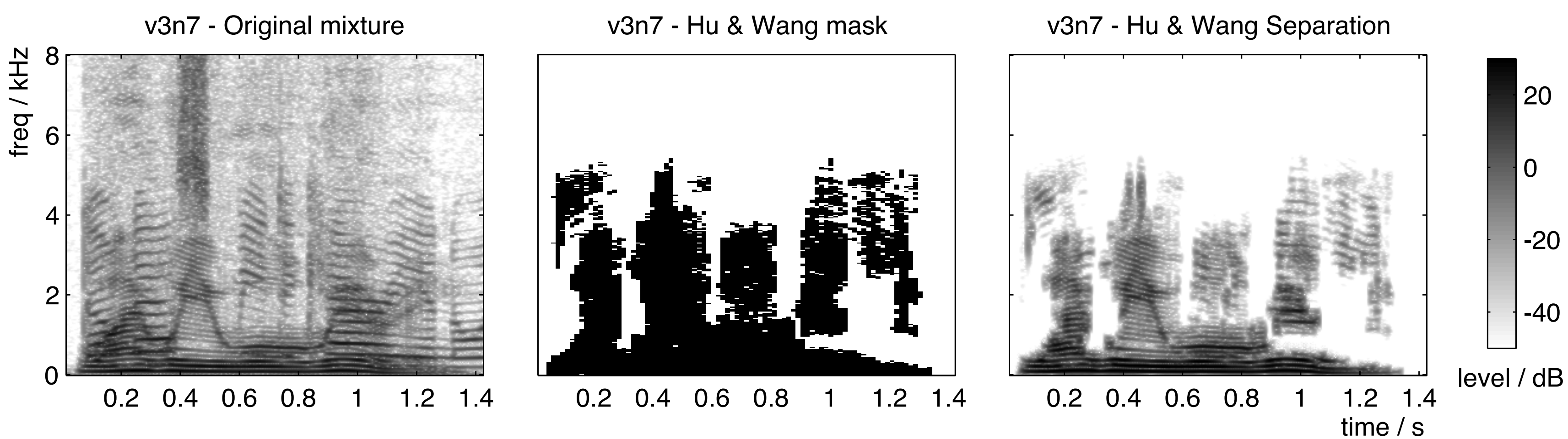


FIGURE 39.7 Example of CASA signal separation via time-frequency masking. Left pane is a spectrogram of a two-voice mixture. Middle pane shows the mask indicating cells dominated by the target voice on the basis of detected harmonicity cues by Hu & Wang [21]. Right pane shows reconstructed target voice.

Other approaches to auditory scene analysis use machine learning techniques, such as factorial HMMs.

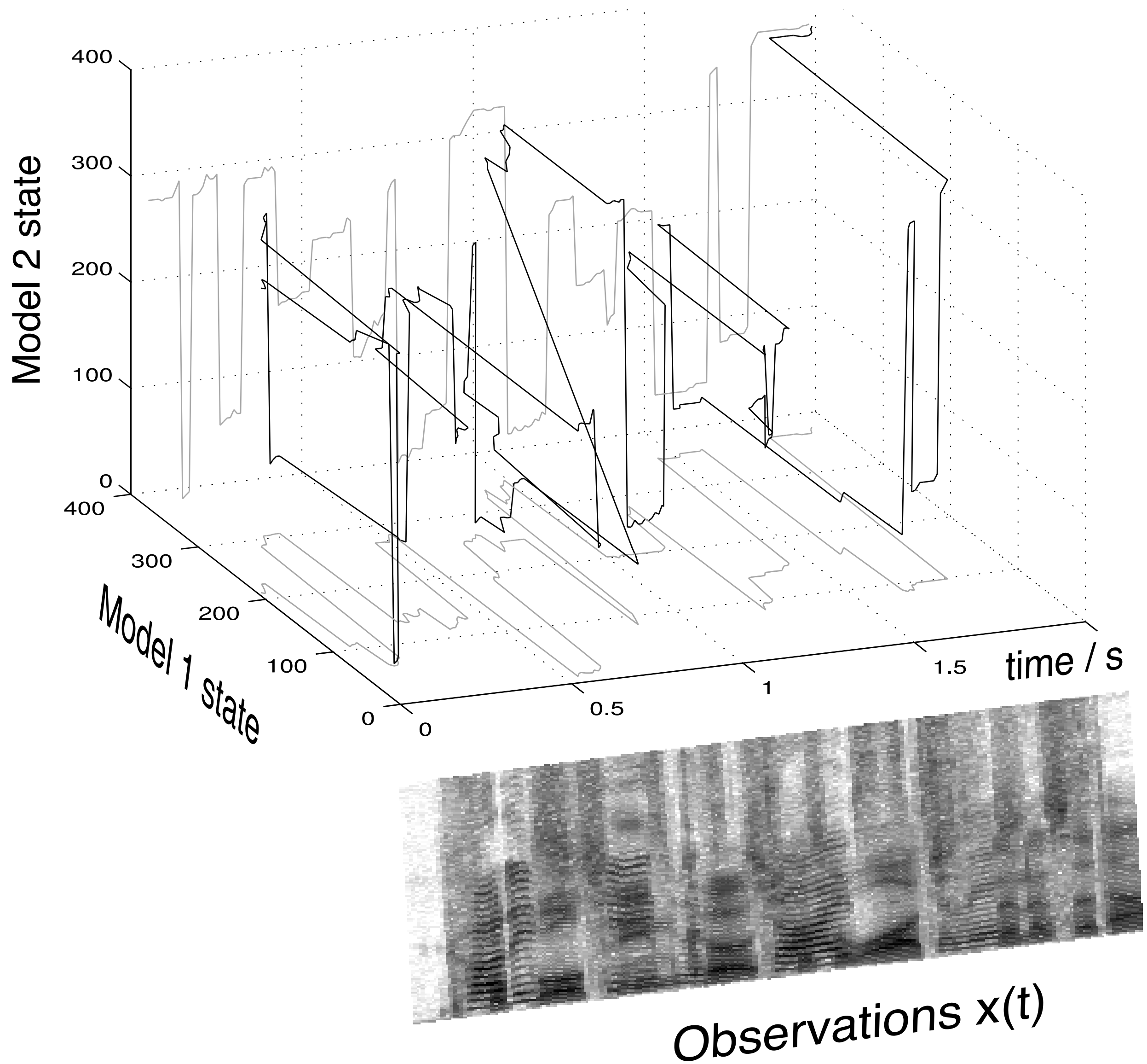


FIGURE 39.8 Illustration of a factorial HMM. The observed mixture is modeled as the combination of two, independent hidden Markov models; the best state sequence is thus a trajectory in a 3-dimensional volume with axes model 1 state, model 2 state, and time. (Figure drawn by Ron Weiss.)

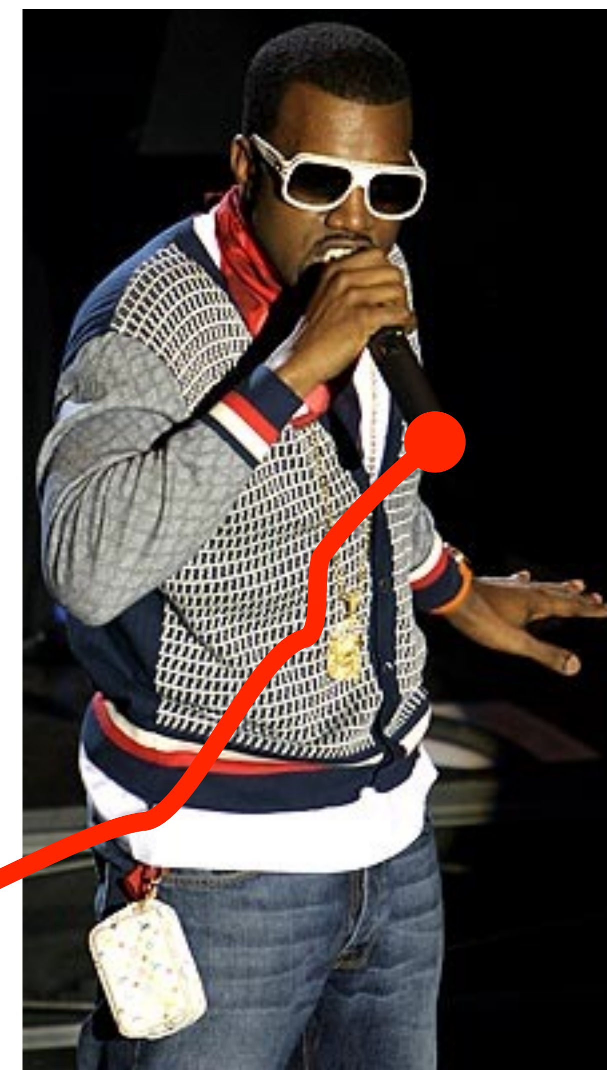
Today's lecture: Source Separation

- Two approaches to the problem ...
  - Auditory scene analysis
  - **Microphone array techniques**
- Research project ideas ...

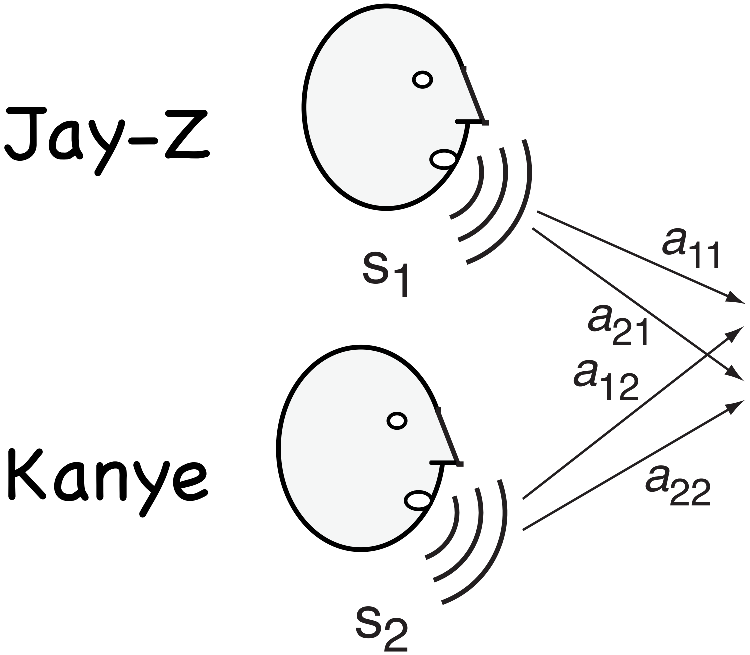
(Close-mic'd voices, different L/R panning for each voice)

Kanye West

Jay-Z



Task: Separate Jay-Z & Kanye from headphone outs



$a_{11}/a_{21}/a_{12}/a_{22}$ are derived from panpot and volume settings.

$$\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} s_1(t) \\ s_2(t) \end{bmatrix}$$

$$\mathbf{x} = \mathbf{A}\mathbf{s}$$

To separate:

$$\hat{\mathbf{s}} = \mathbf{W}\mathbf{x}$$

Where:

$$\mathbf{W} = \mathbf{A}^{-1}$$

What's the catch?

Left = x1 Right = x2

What's the catch?

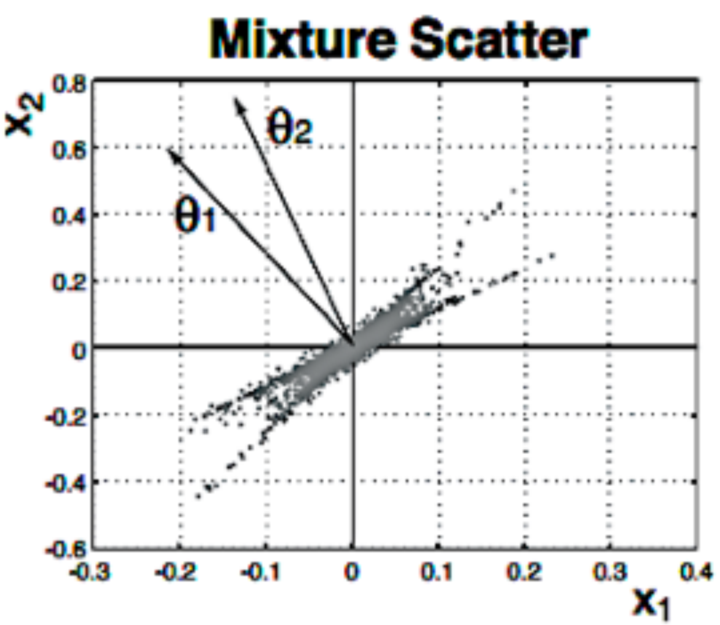
What if matrix inverse of **A** doesn't exist?

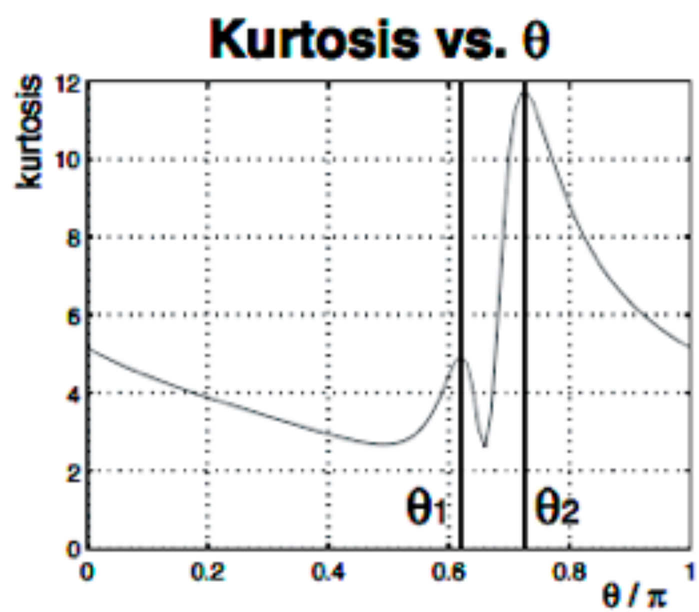
$$\mathbf{W} = \mathbf{A}^{-1}$$

$$\begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{bmatrix} = \frac{1}{\det \mathbf{A}} \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix}$$

Also, numerical stability issues, etc ...

$a_{11}/a_{21}/a_{12}/a_{22}$ are generally not known. We need to "learn" them from the signals over time.

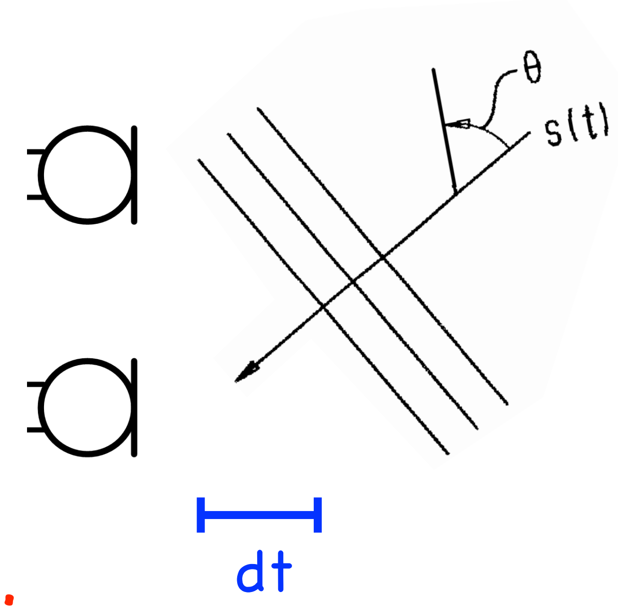




Independent Component Analysis (ICA) can be used to learn **A**

What's the catch?

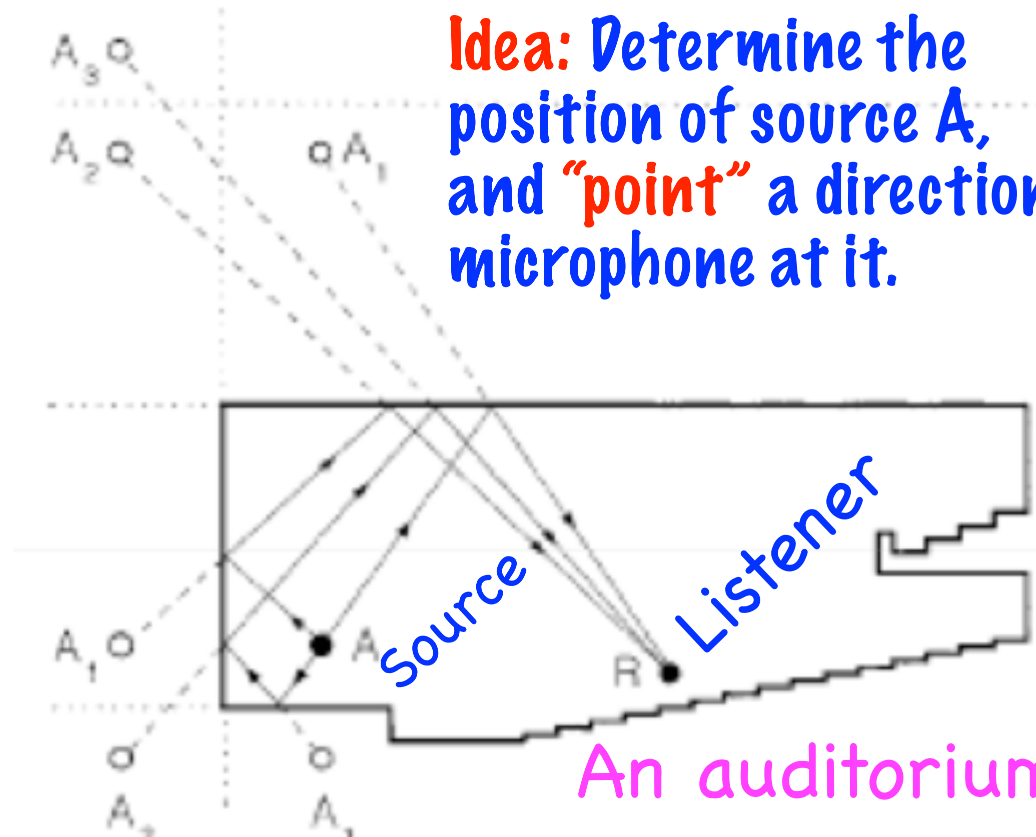
Real-world voices combine "in the air", not a mixer. The speed of sound is finite, and so each mic hears each voice with a (relative) delay.



Doing "un-mixing" AND "un-delaying" is harder ...

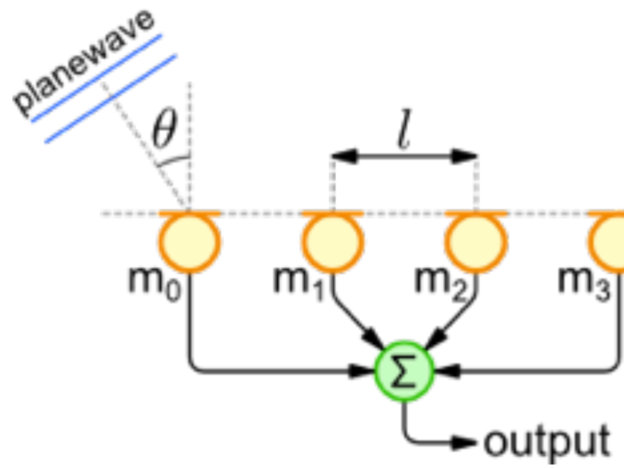
Real rooms are not anechoic chambers. "Early reflections" of a source A act like "virtual sources," (labelled $A_1 \ldots A_6$) that confuse simple unmixing algorithms.

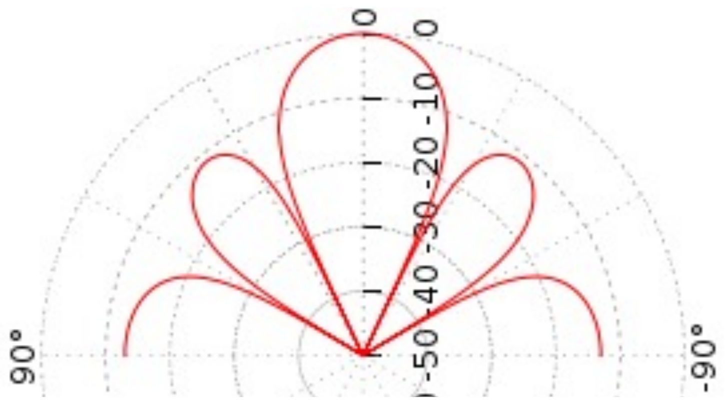
Idea: Determine the position of source A, and "point" a directional microphone at it.

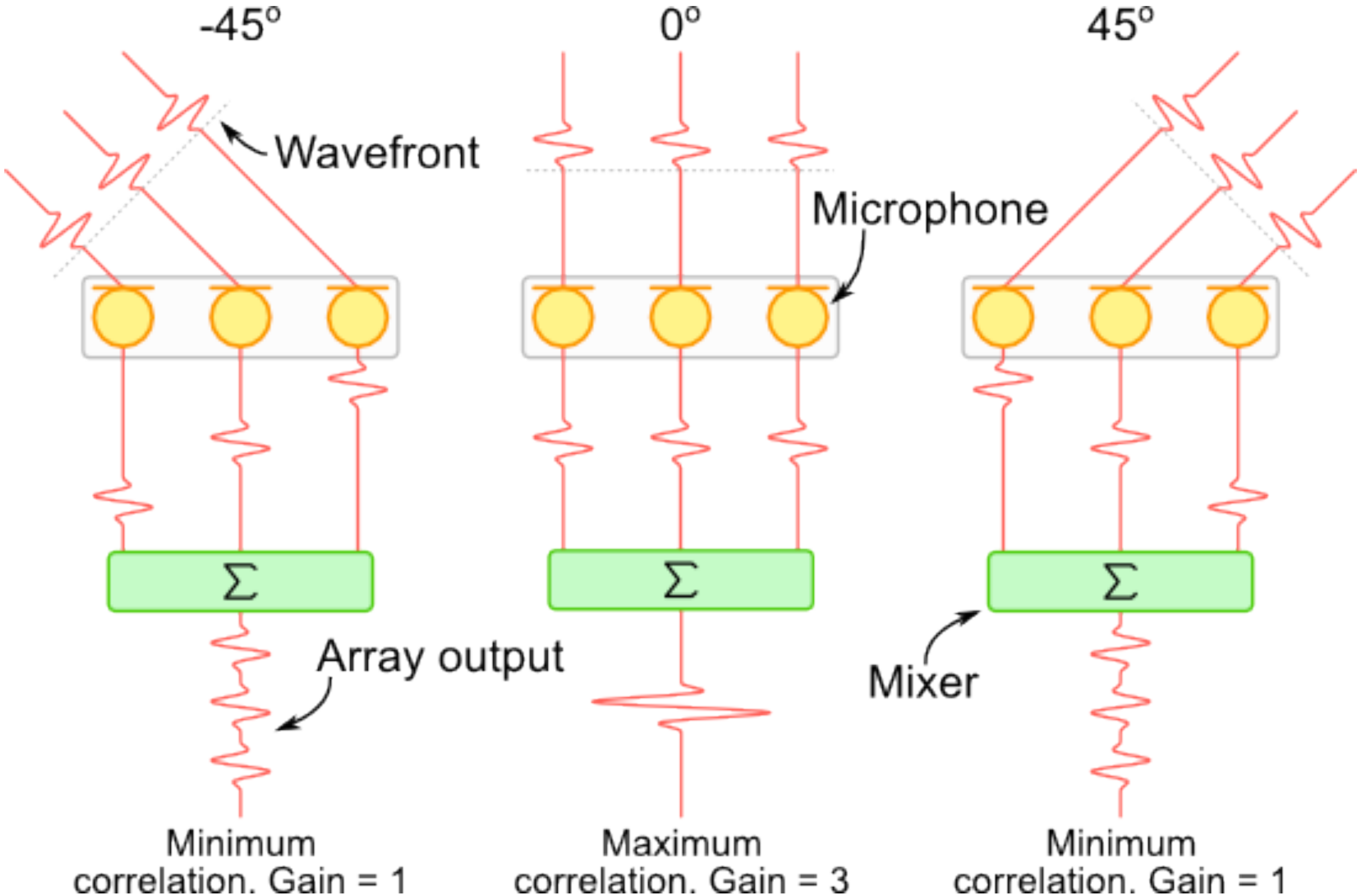


An auditorium

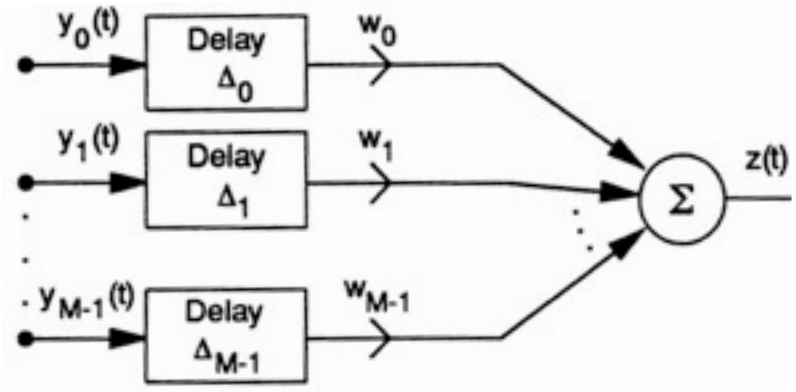
Broadside array



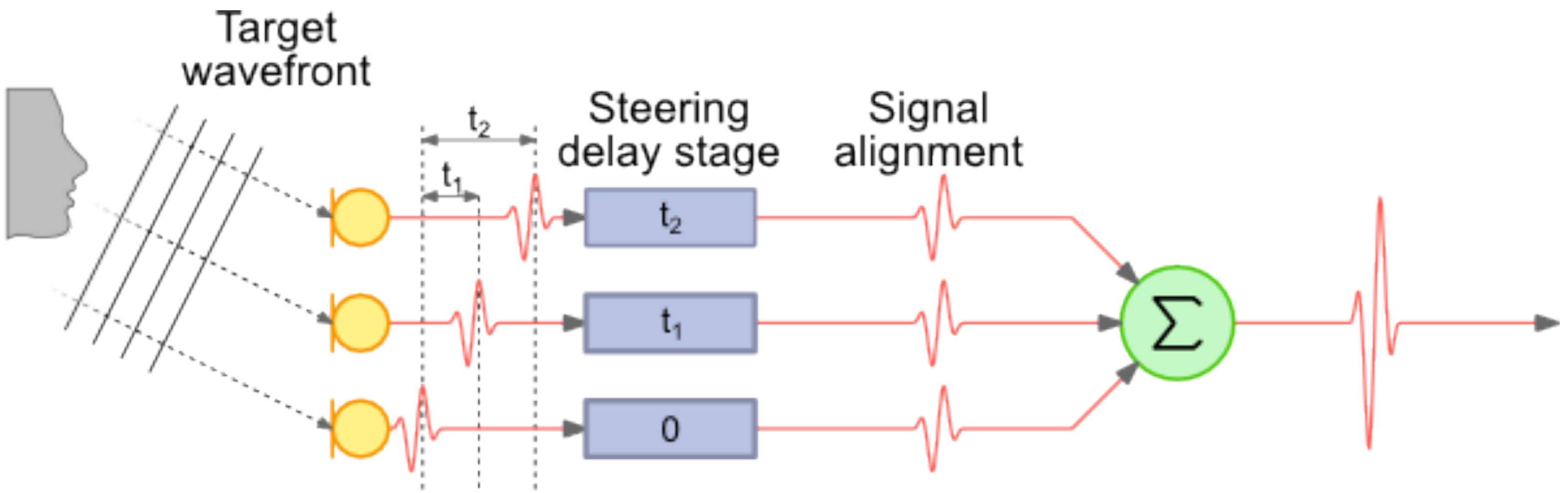


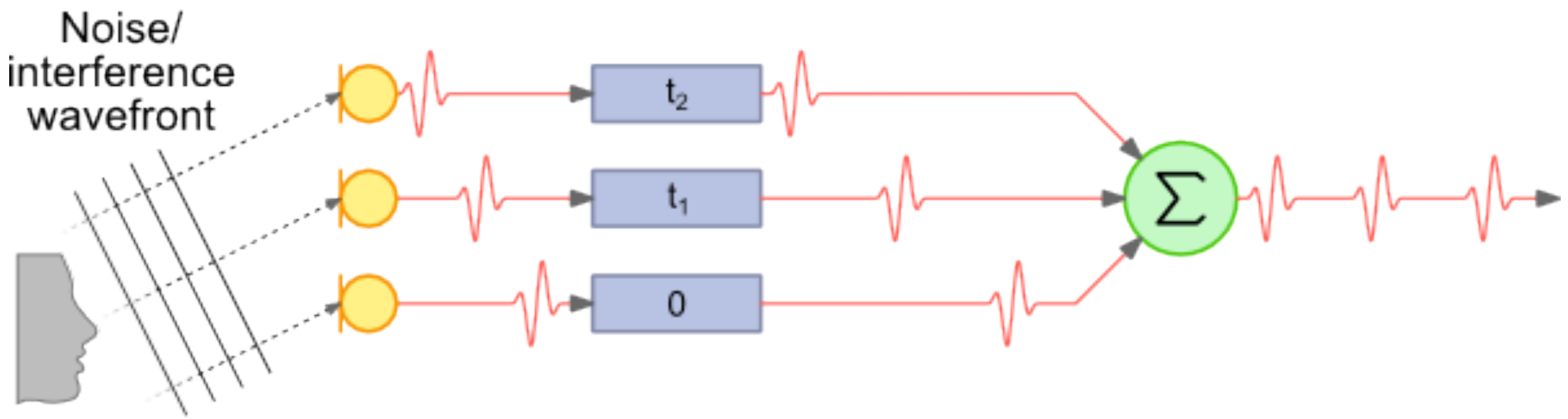


Steerable beam …



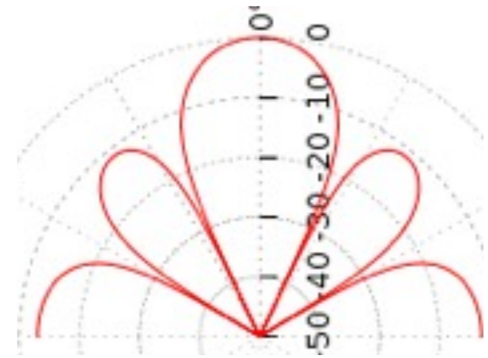
By changing the delays, we can steer the beam to track a target.

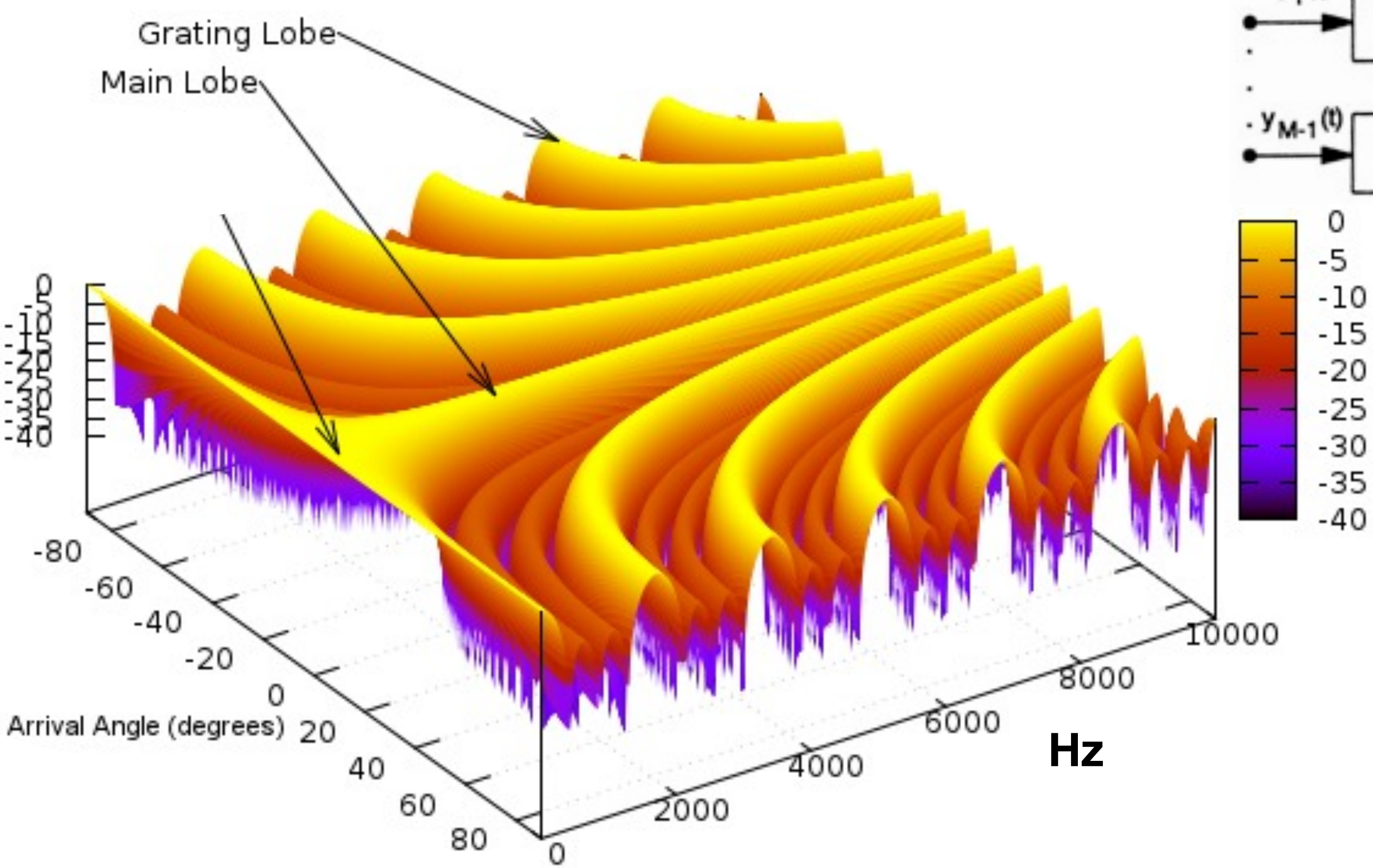


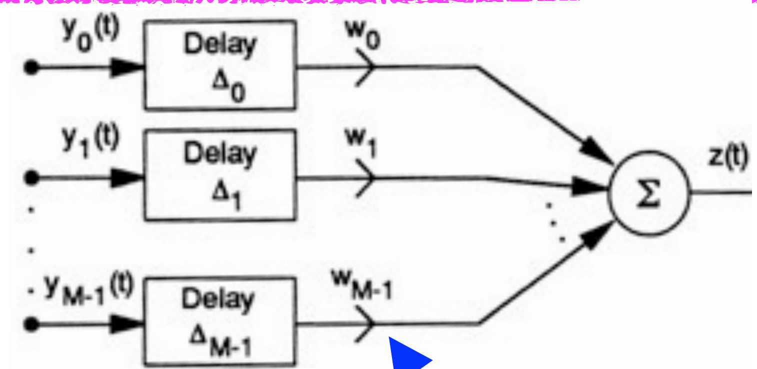


Off-axis response

The main lobe has a flat frequency response, but the side lobes are comb filters … yielding an unnatural off-axis sound







The weight values can be chosen to minimize the combing effects … see book for details.

Adaptive filters ...

A more flexible approach is to add an adaptive filter to the architecture, so that signal quality can be optimized on-line.

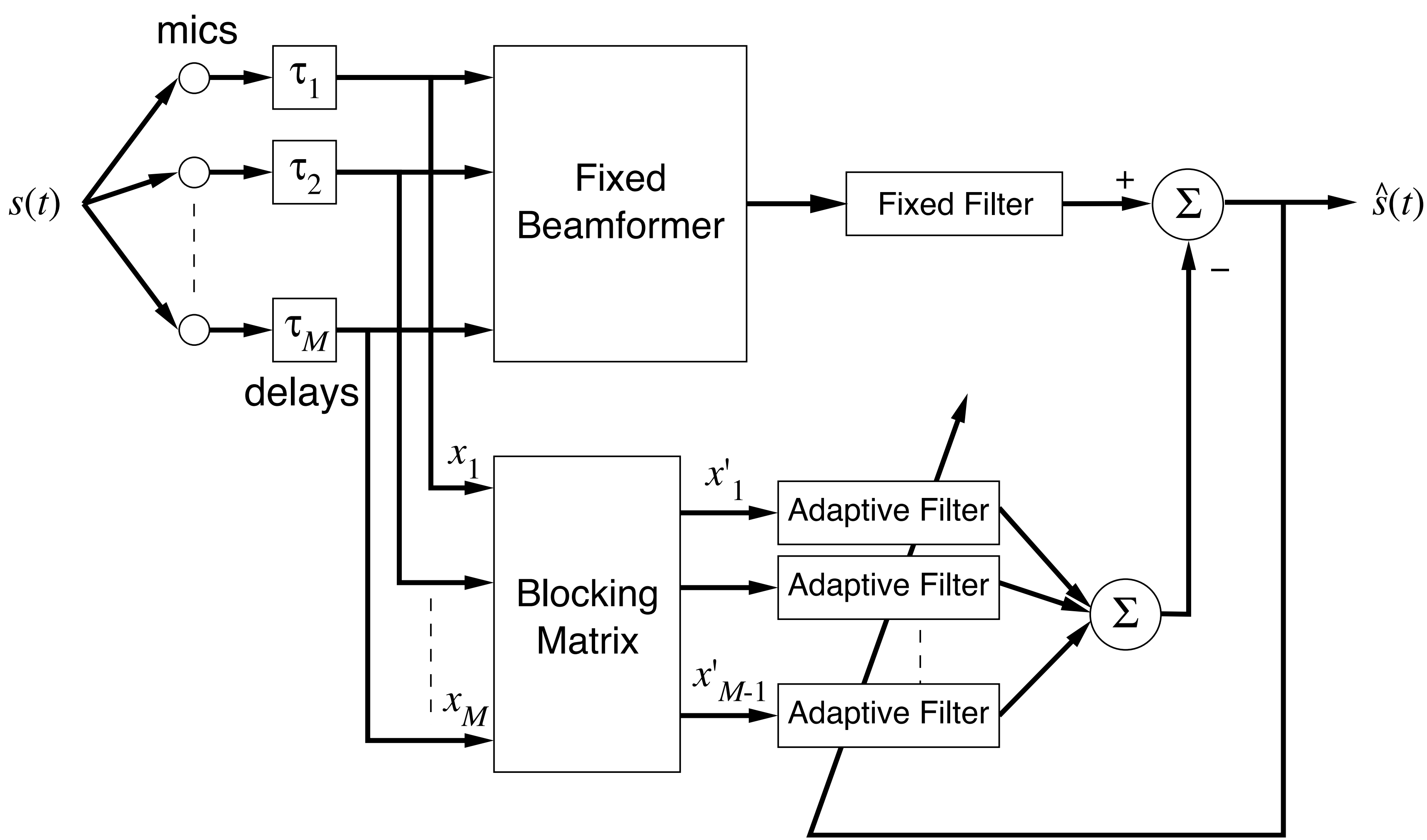
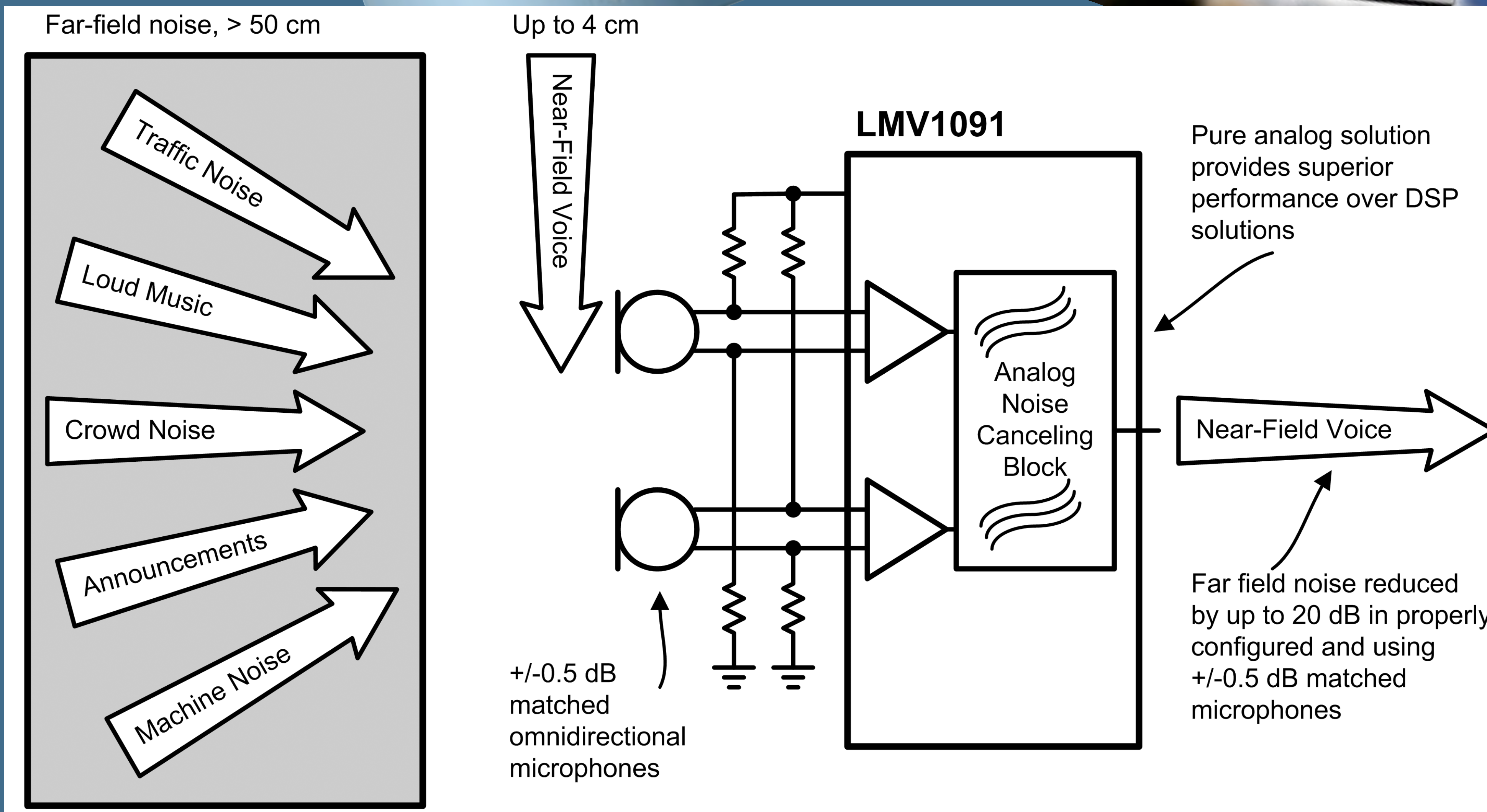


FIGURE 39.4 The Generalized Sidelobe Canceller.

Recall ...


National
Semiconductor
LMV1088

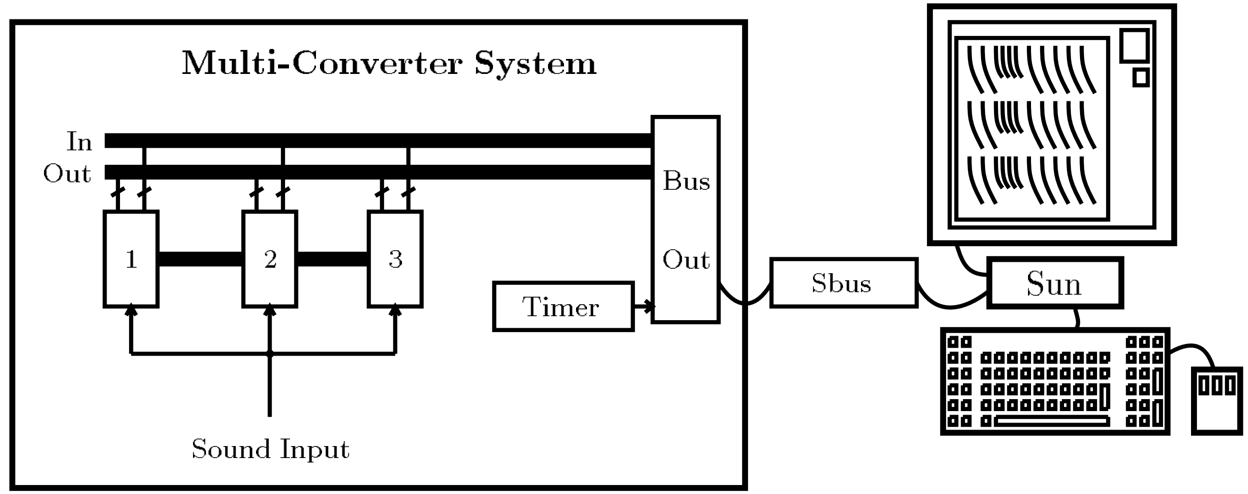

Far-field noise, > 50 cm
Traffic Noise
Loud Music
Crowd Noise
Announcements
Machine Noise
Up to 4 cm
Near-Field Voice
LMV1091
Analog Noise Canceling Block
Pure analog solution provides superior performance over DSP solutions
Near-Field Voice
+/-0.5 dB matched omnidirectional microphones
Far field noise reduced by up to 20 dB in properly configured and using +/-0.5 dB matched microphones

Today's lecture: Source Separation

- Two approaches to the problem ...
  - Auditory scene analysis
  - Microphone array techniques
- **Research project ideas ...**

Auditory scene analysis ... why isn't the future here yet?

The Recognizer – Representation Gap.



Auditory Models	Speech Recognition
Adaptive Sampling	Uniform Sampling
Specialized Features	General-Purpose Features
Multiple Representations	Single Representation
High-Dimensional	Low-Dimensional
Correlated Features	Uncorrelated Features