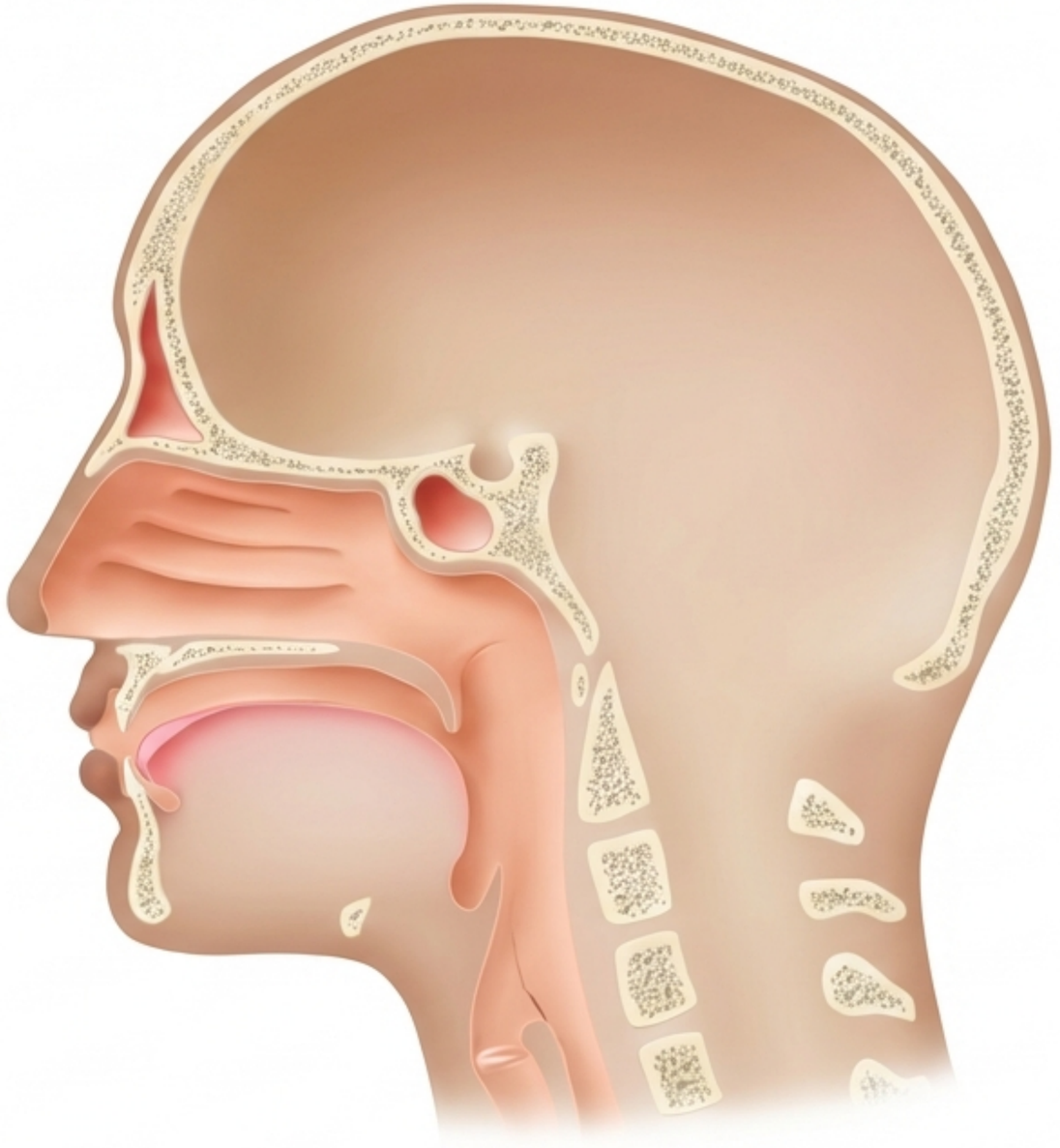
Perceiving Speech: The Decoder's Journey

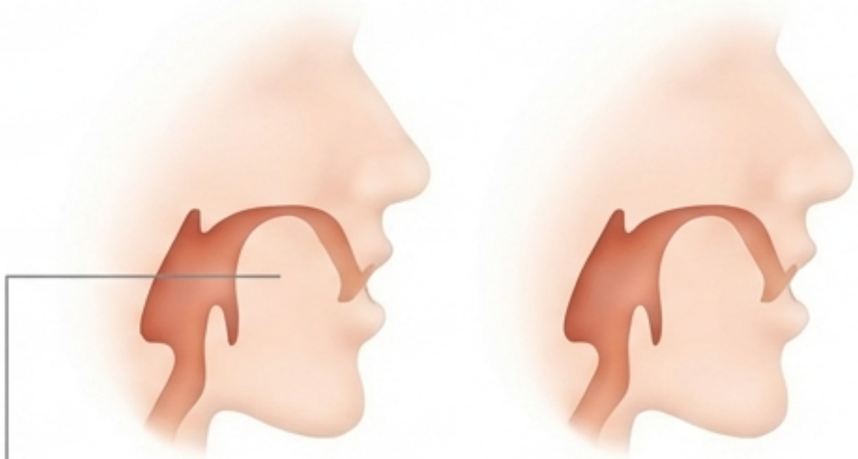
From Acoustic Signal to Neural Meaning

Speech perception is a feat of reverse engineering. We begin with a physical signal—pressure changes in the air—and transform it into abstract thoughts. This deck explores the mechanisms, the challenges of variability, and the neural architecture that makes human communication possible.

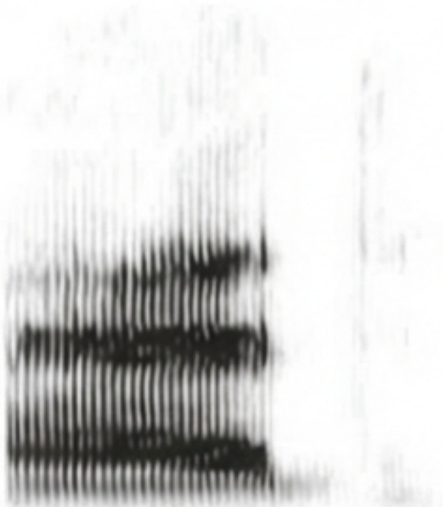


The Acoustic Signal

Speech is produced by air pushed from the lungs past the vocal cords. The vocal tract acts as a filter, concentrating energy at specific frequencies to create ‘formants’.

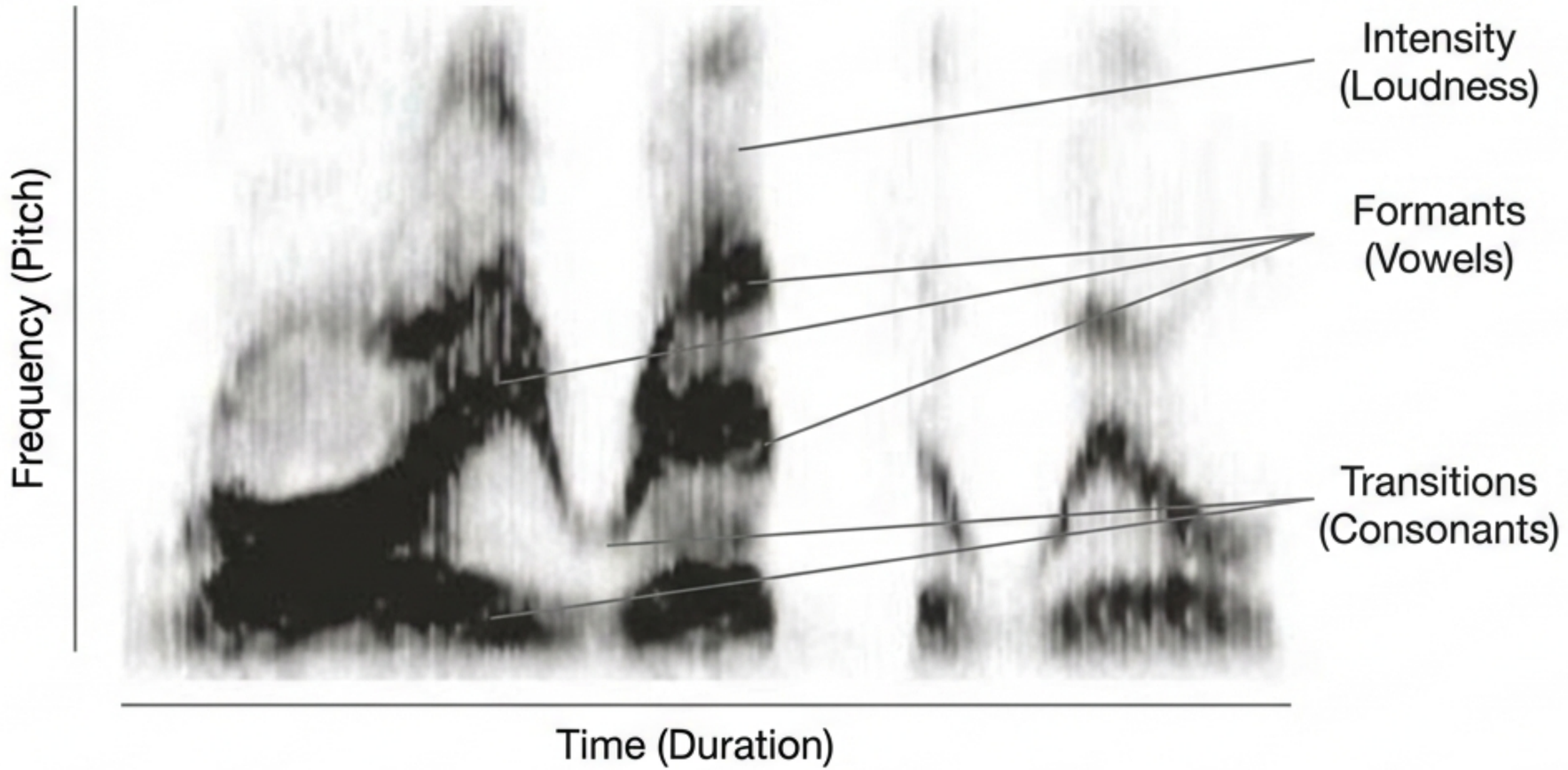


Articulatory Shapes



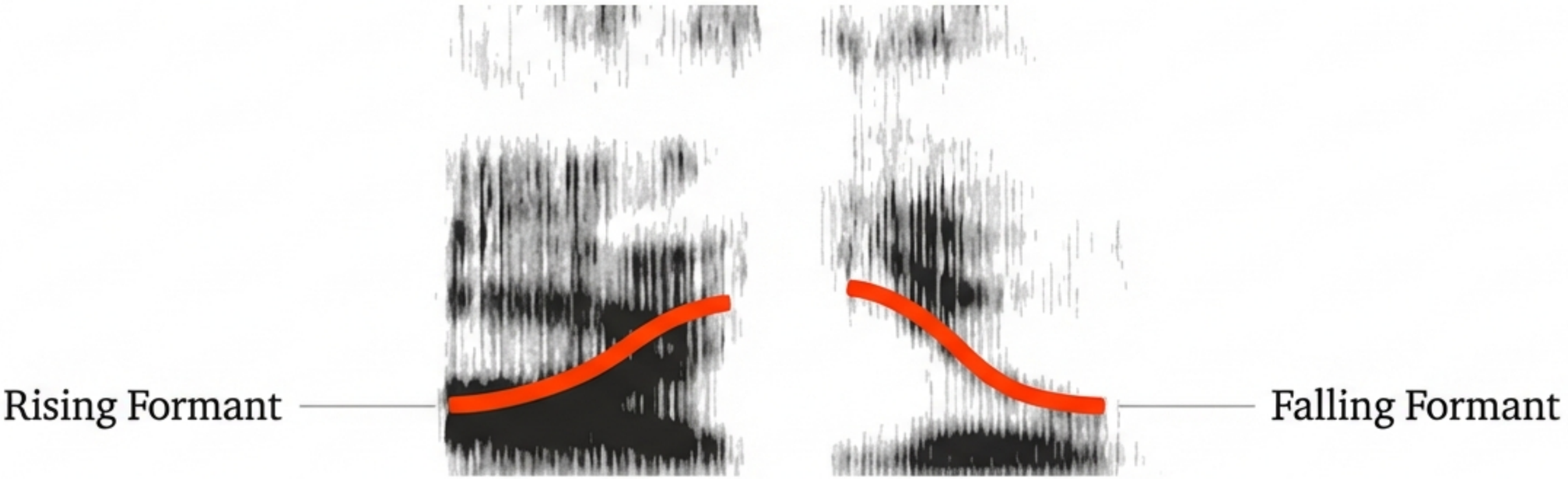
The Acoustic Output

Visualizing the Code: The Spectrogram



The spectrogram maps speech energy over time. Notice the rapid “Formant Transitions” (T2, T3) that signal the consonants, contrasted with the steady bands of the vowels.

The Variability Problem

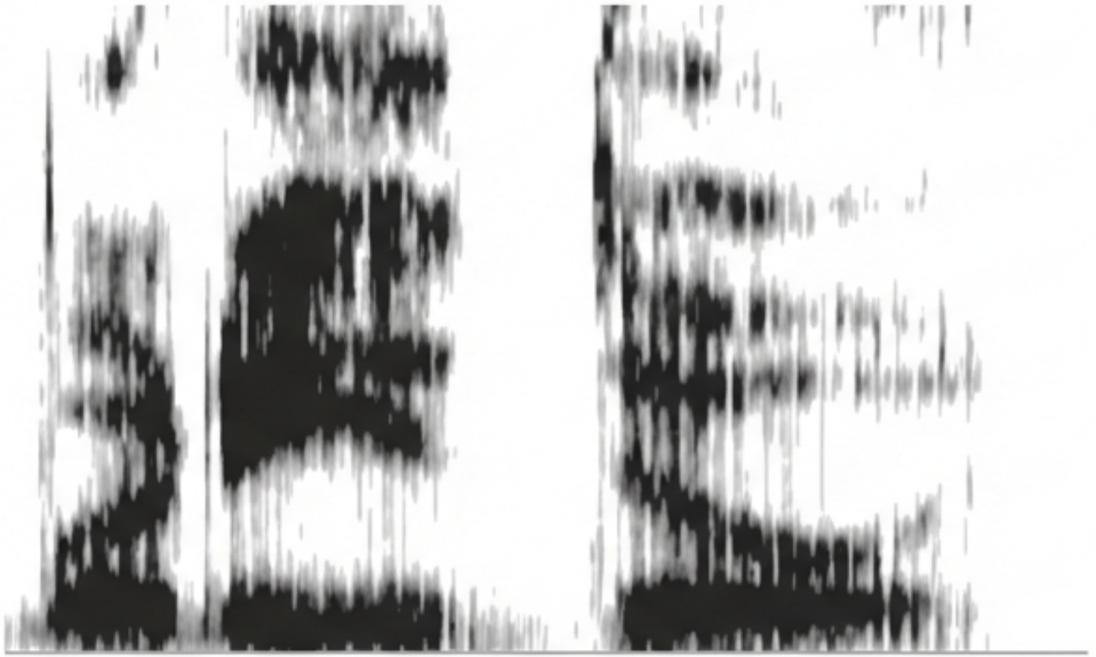


A phoneme is the smallest unit of meaning. However, the acoustic signal for a single phoneme is not constant. In ‘/di/’ the signal rises. In ‘/du/’ the signal falls. Yet, we perceive the same ‘d’. This is Perceptual Constancy.

Coarticulation & Sloppy Speech

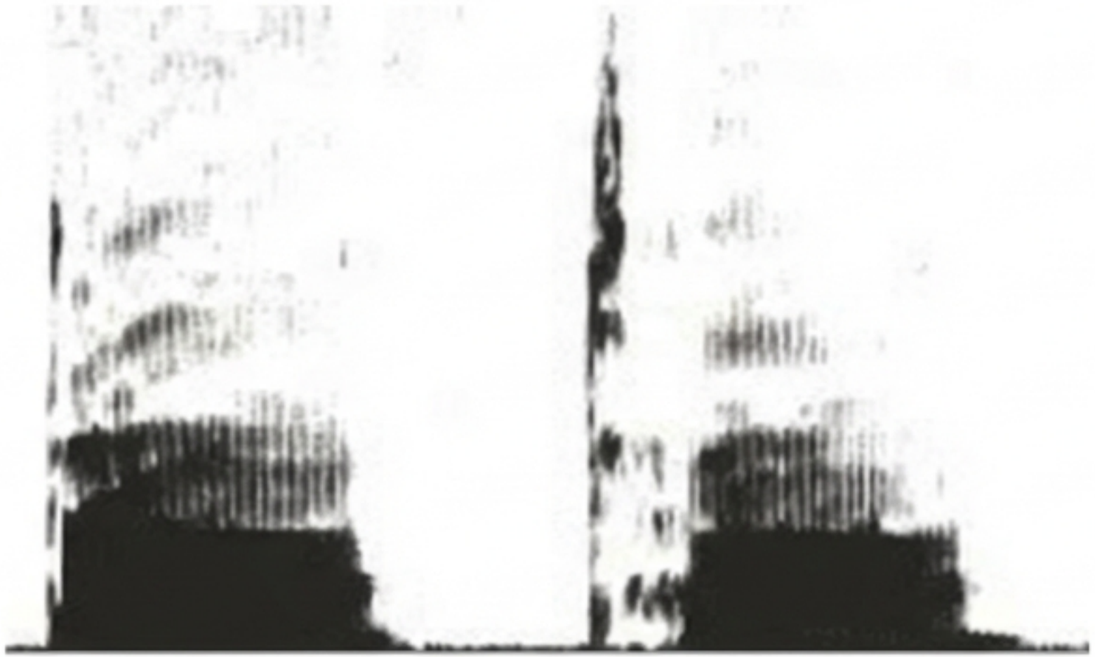
We do not speak like robots. We slur and merge sounds.

Formal Articulation



Clear separation between words.

Conversational Speech



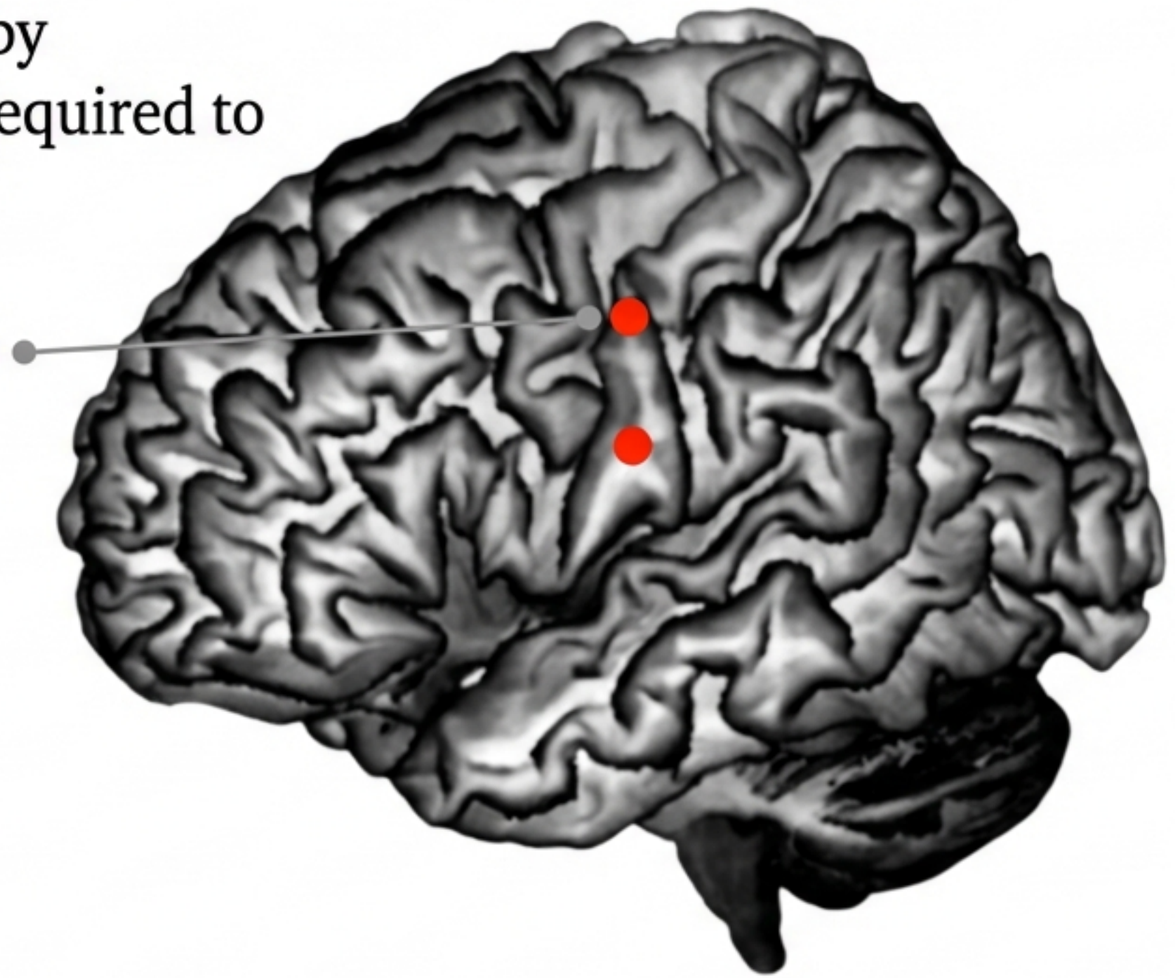
Coarticulation: Sounds overlap and pauses disappear.

Despite the massive loss of acoustic detail in conversational speech, the brain successfully segments the meaning.

Solution 1: The Motor Theory

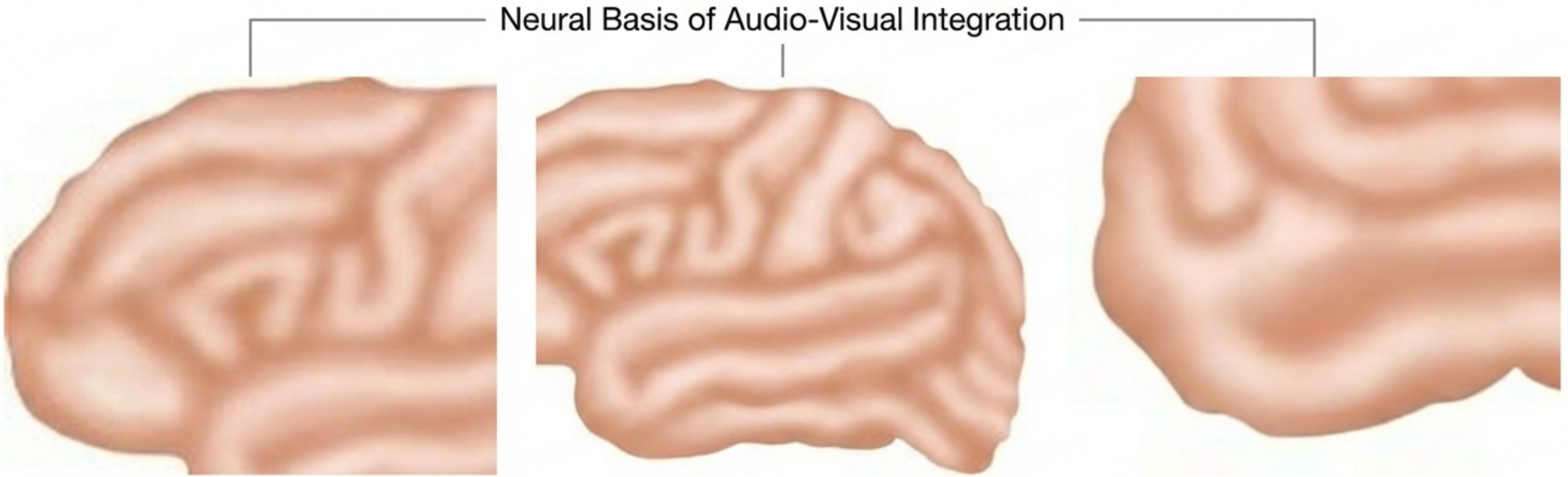
Proposed in the 1960s: We perceive speech by mentally simulating the motor movements required to produce it. “Hearing is simulated speaking.”

- **Categorical Perception:** We hear sharp boundaries (e.g., /da/ vs /ta/) rather than a continuum.
- **Voice Onset Time (VOT):** The time delay between sound onset and vocal cord vibration.
- **Phonetic Boundary:** The exact VOT where perception flips from one category to another.



Solution 2: Multisensory Integration

The McGurk Effect: Visual input alters auditory perception.

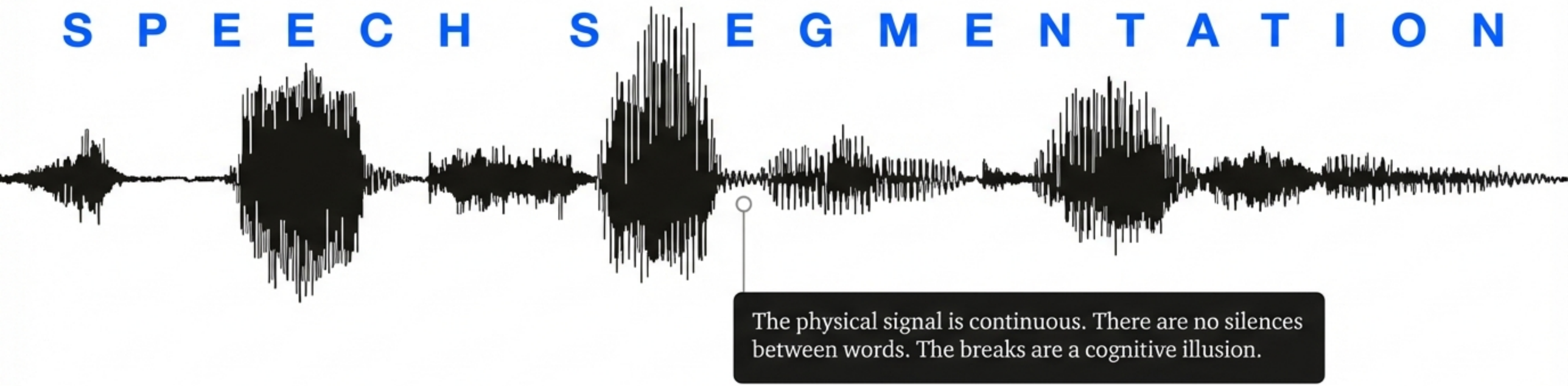


Processing involves the Superior Temporal Sulcus (STS) and the Fusiform Face Area (FFA). When we hear a familiar voice, the brain activates face-processing regions even without visual input.

Solution 3: Top-Down Processing

Perception is prediction. We use context to hallucinate missing data and insert breaks where none exist.

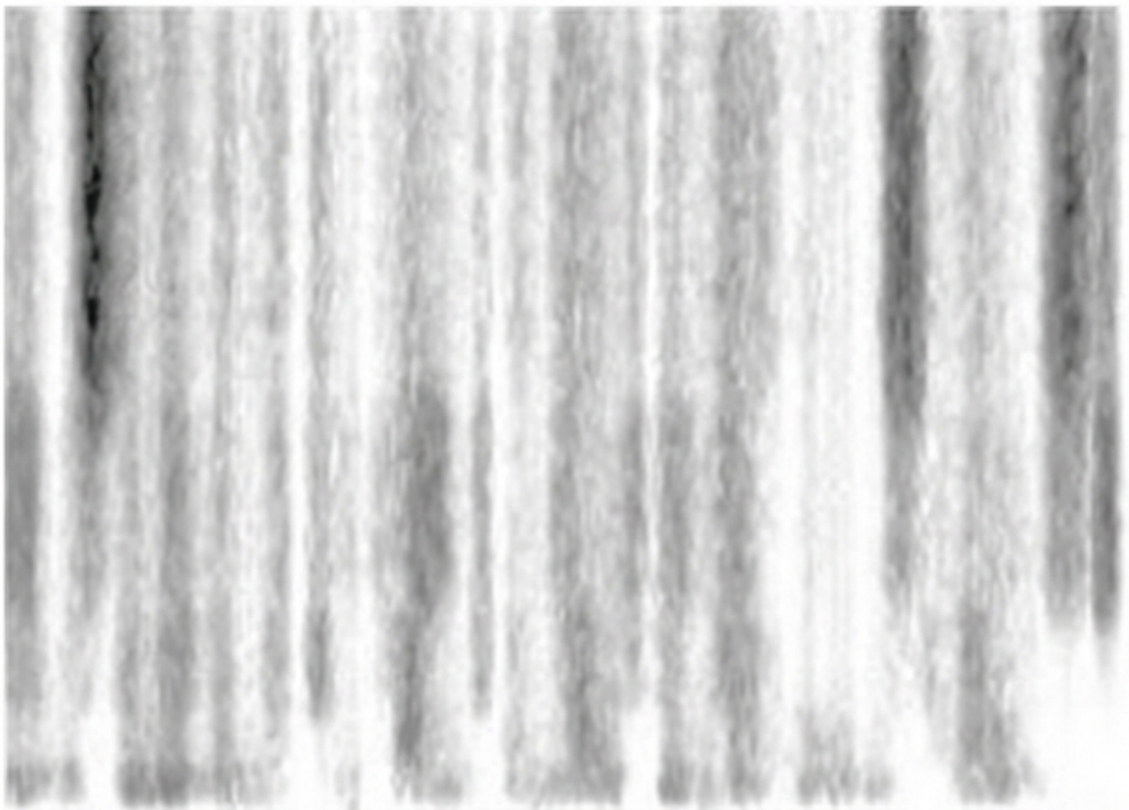
- Phonemic Restoration Effect: Filling in missing sounds based on context.
- Speech Segmentation: Mentally inserting word boundaries.



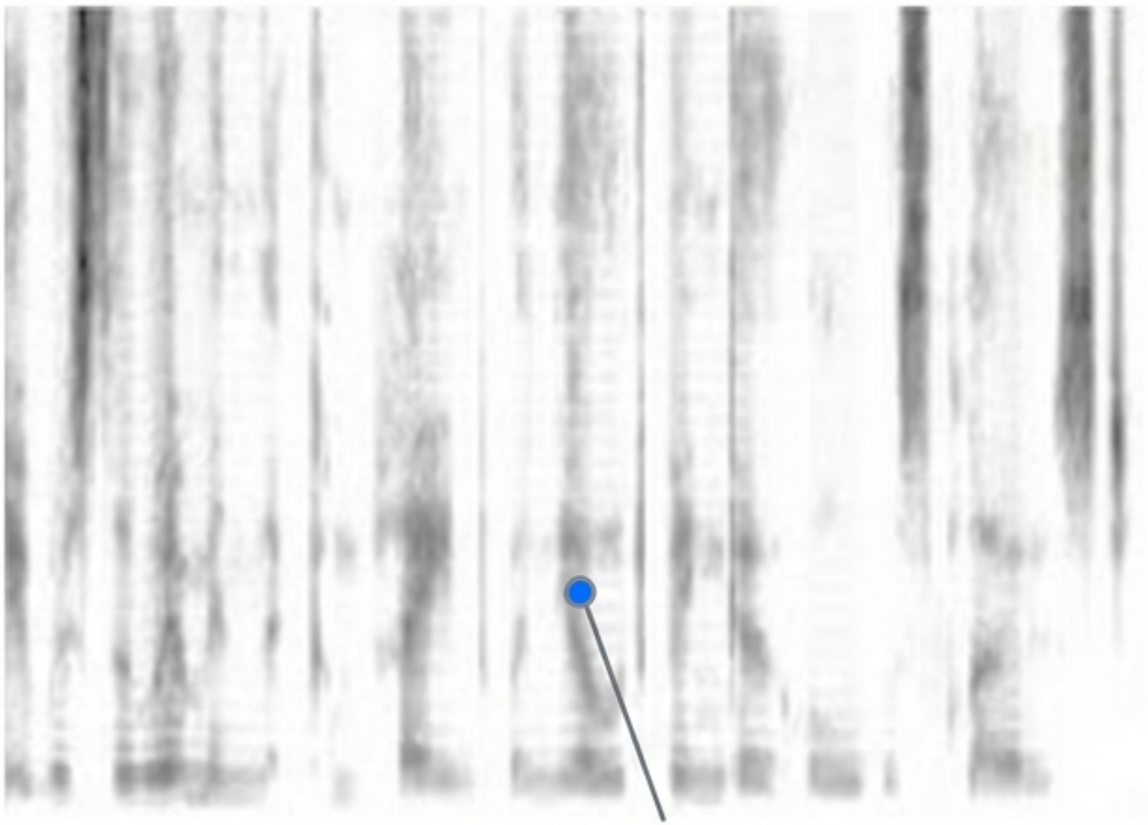
Solution 4: Statistical Learning

The brain calculates “Transitional Probabilities” to guess where words end.
We can even learn to decode degraded signals.

Clear, Original Speech

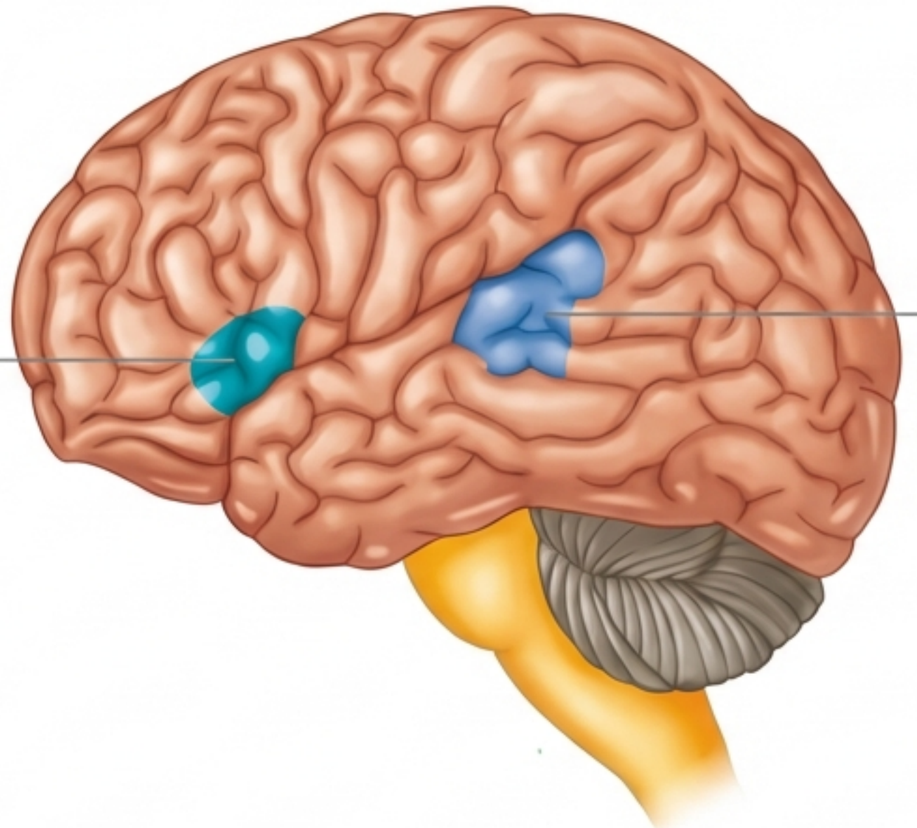


Noise-Vocoded (Degraded) Speech



Noise-Vocoded Speech. Even with this loss of frequency detail, the brain can learn to decode the temporal patterns over time.

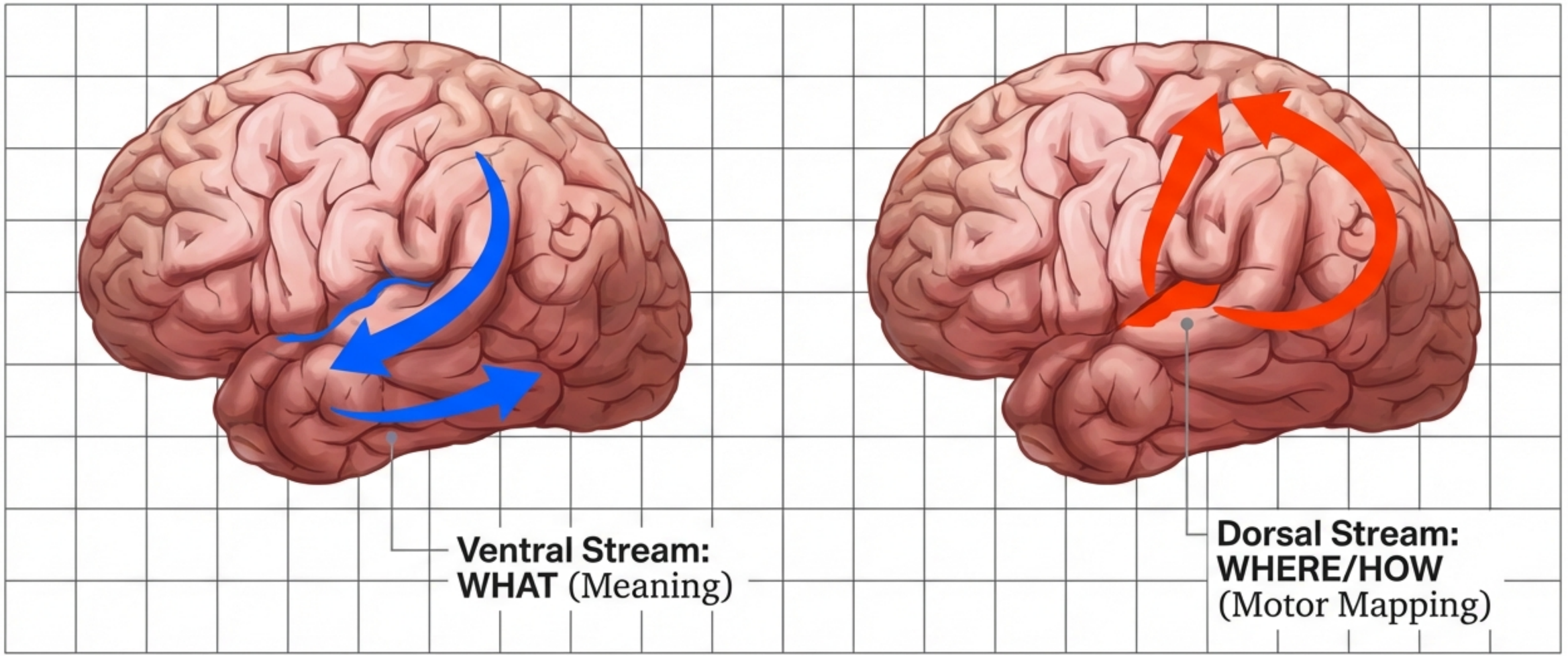
The Neural Hardware: Classic Models



Broca’s Aphasia: Damage leads to slow, labored speech.

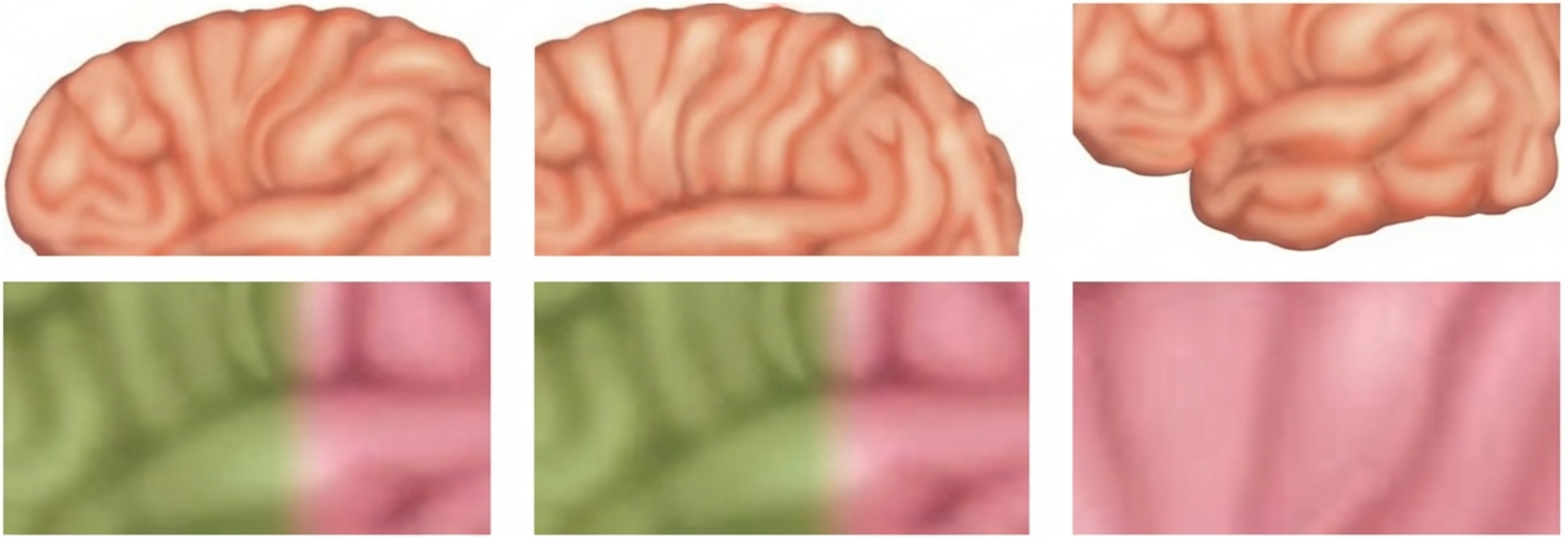
Wernicke’s Aphasia: Damage leads to fluent but nonsensical “word salad”.

The Dual-Stream Model



Coupling Production and Perception

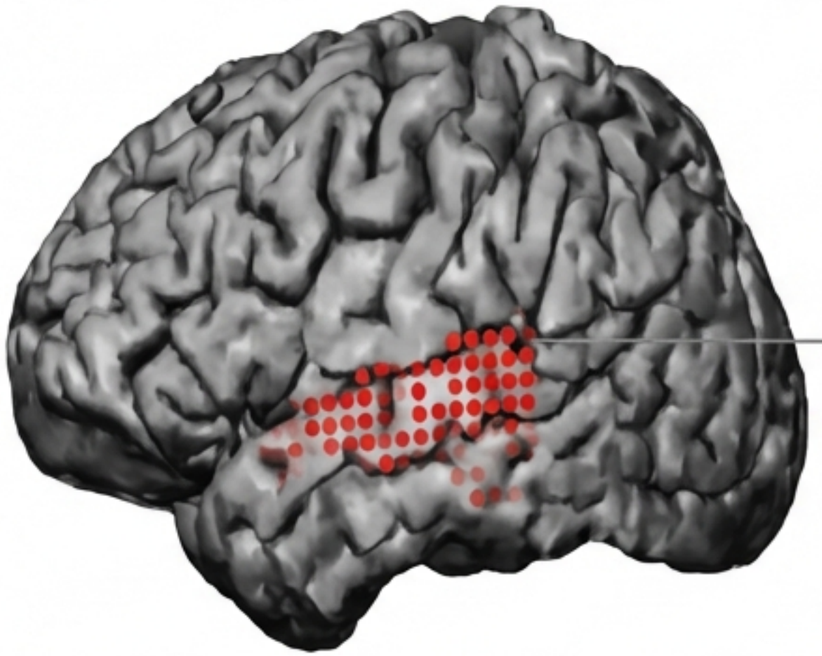
Understanding speech recruits the same machinery used to create it.



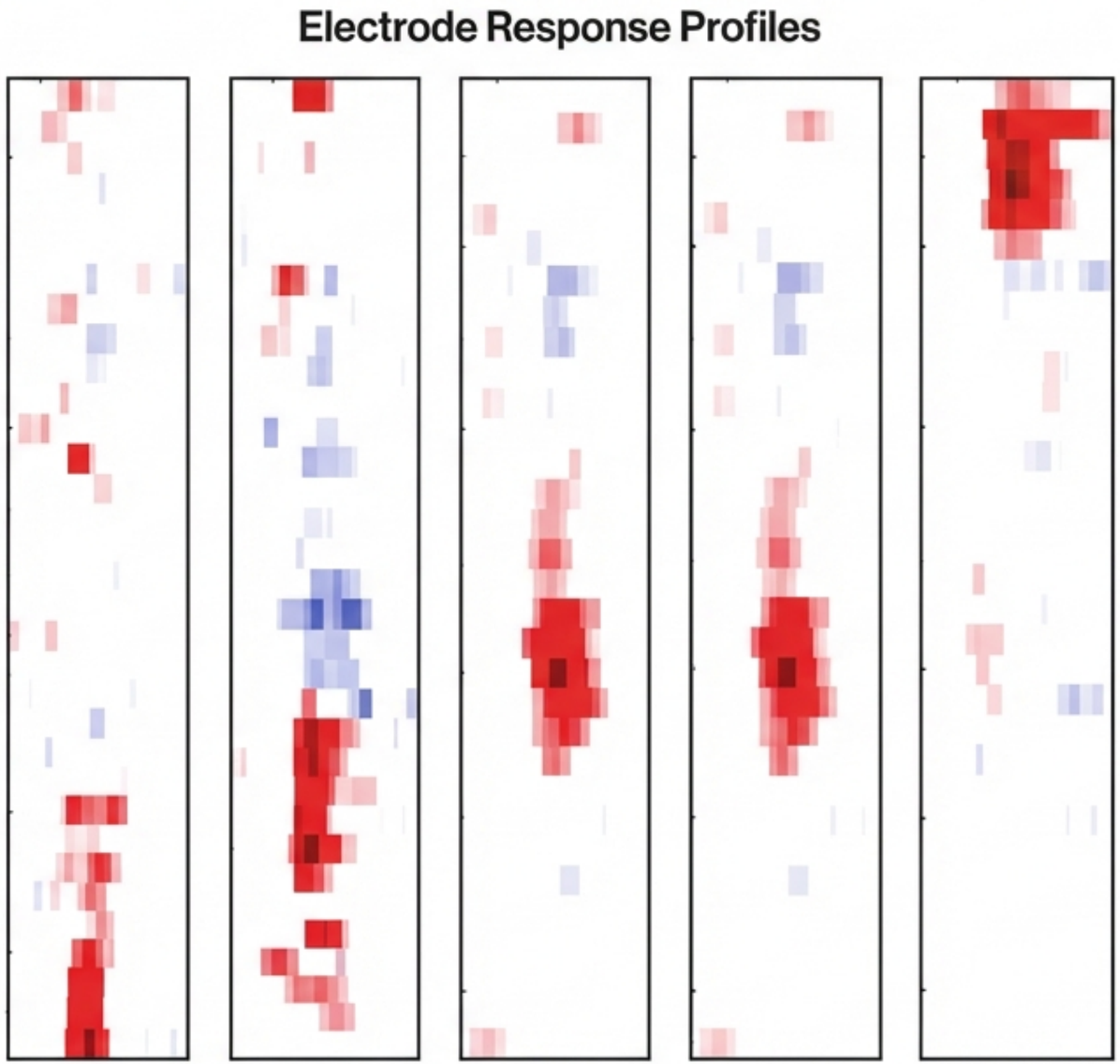
Cortical Surface Activation (Silbert et al., 2014)

fMRI evidence shows coupled neural responses. The areas active when telling a story (**Production**) significantly overlap with areas active when listening to a story (**Comprehension**).

Neural Decoding at the Phoneme Level

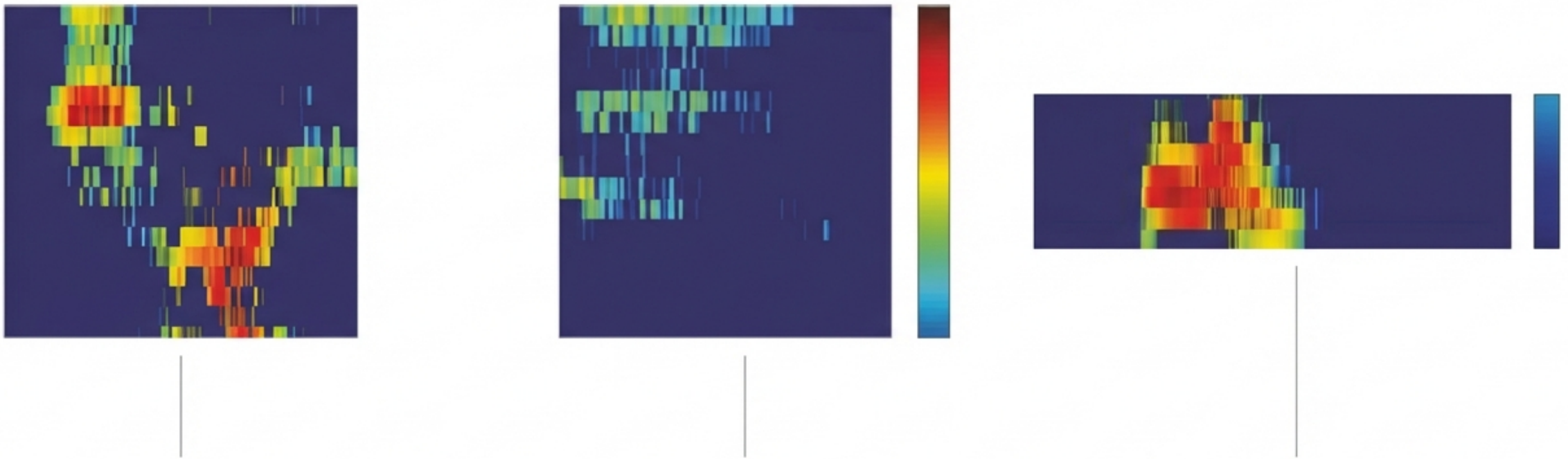


Electrodes placed directly on the temporal lobe reveal that neurons are tuned to specific Phonetic Features (e.g., plosives vs. fricatives) rather than just whole words.



Bio-Hacking: The Cochlear Implant

Bypassing damaged hair cells to stimulate the auditory nerve directly.

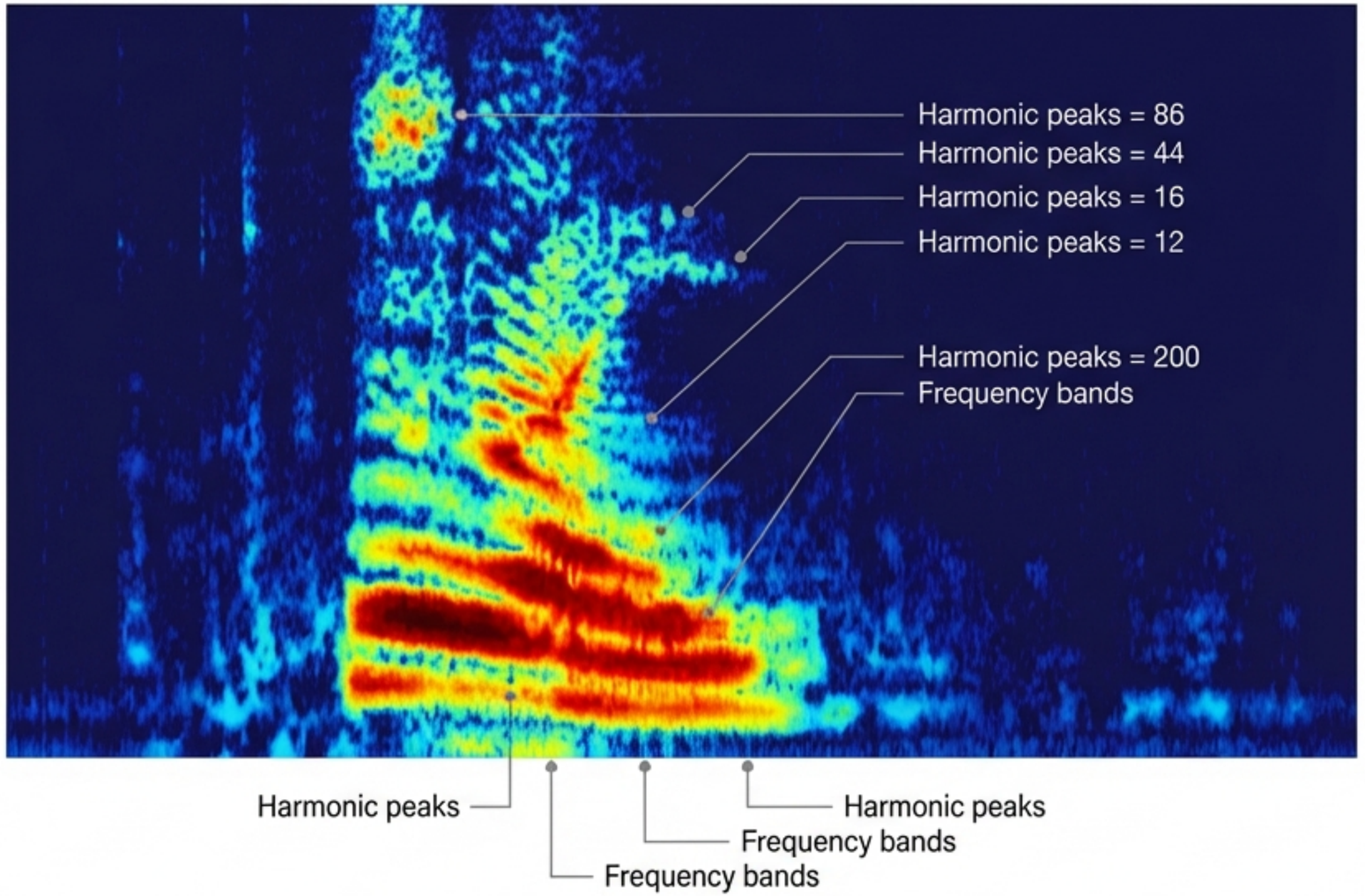


The Input Signal: CIs split sound into coarse frequency bands. The resulting signal is low-resolution, often described as sounding like a "radio out of tune".

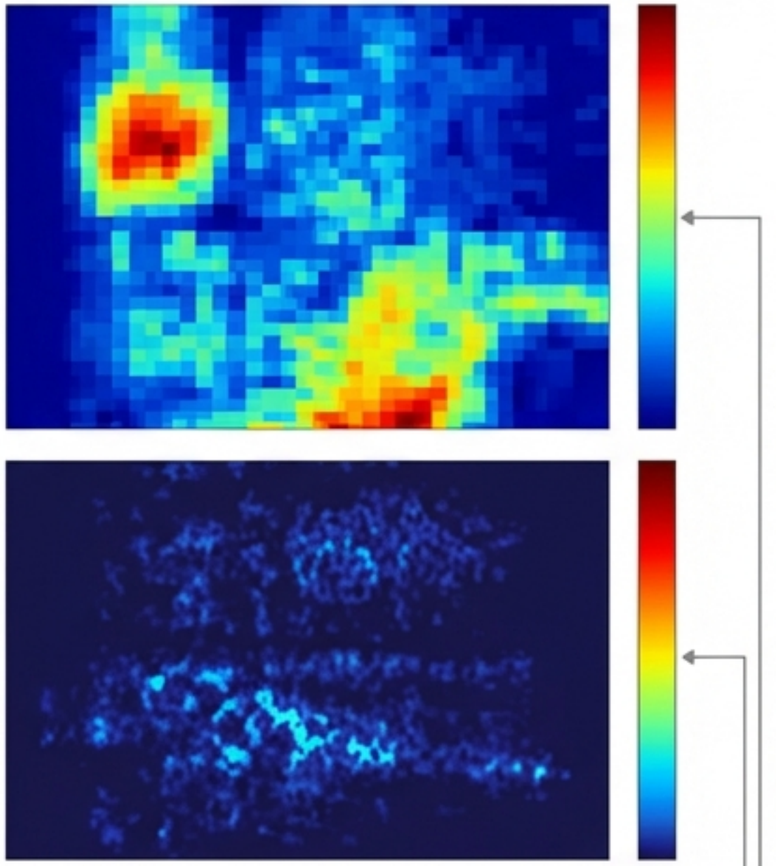
Natural vs. Artificial Resolution

Comparing the original spectrogram (Left) with the CI simulation (Right) reveals a massive loss of harmonic detail. Yet, thanks to plasticity, the brain eventually learns to map these coarse signals to meaning.

Natural Hearing



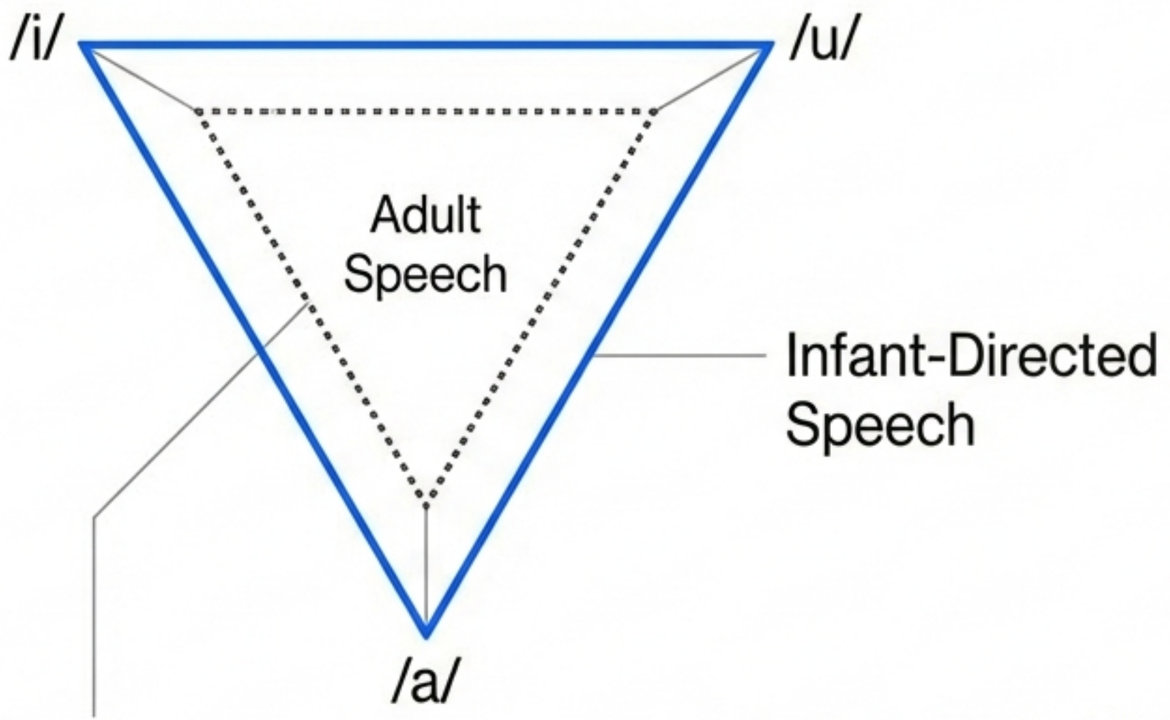
Implant Stimulation



These simulations represent the electrical signals delivered by the implant, lacking the fine structure and harmonic richness of natural sound.

Development: Infant-Directed Speech

‘Parentese’ is a teaching tool. It uses higher pitch and exaggerated vowels to help the infant brain learn categories.



The Expanded Vowel Triangle: Exaggerating formants makes phonetic distinctions clearer.

The Resilient Decoder

Speech is a collaborative act between the speaker's motor system and the listener's memory. We have traced the signal from air pressure to neural firing. Next, we move from pressure waves in the air to pressure on the skin.

End of Chapter 14: Perceiving Speech