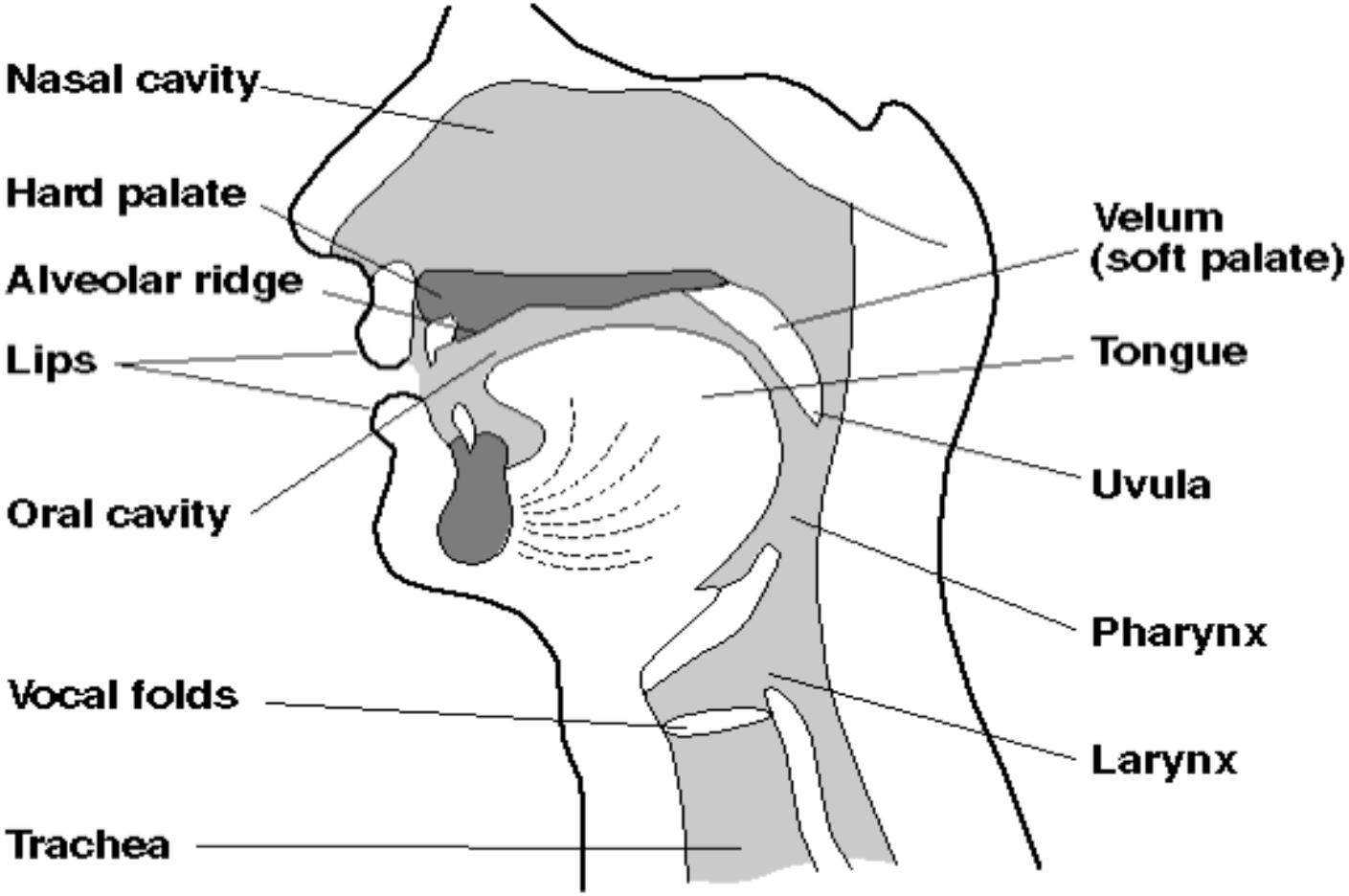
Speech Perception

- The speech signal is the result of the movement of the tongue, lips, jaw, and vocal cords in modifying the air stream from the lungs.
- The sounds of speech (phonemes) are a basic set of building blocks from which words are built.
- Perception involves both a complex mapping of basic auditory features onto phonemes and influences of linguistic knowledge. (You know the words of your language.)
- The role of individual neural cells in speech perception is largely unknown. Data from human studies reveal parts of the cortex that have specialized roles in language processing.

Overview

1. Speech articulation and the sounds of speech.
2. The acoustic structure of speech.
3. The classic problems in understanding speech perception: segmentation and variability.
4. Basic perceptual data and the mapping of sound to phoneme.
5. Higher level influences on perception.
6. Physiology of speech perception and language.

Articulation



Articulation - 2

The sounds of speech are a result of the coordinated movement of the articulators in modifying the air-flow from the lungs.

The units of language include sentences and phrases, words, syllables, and phonemes. The phonemes, in turn, can be described in terms of the roles of the articulators in producing them.

When we write the sounds of speech (phonemes), we put the symbols between //s to indicate that these are the sounds (*not spelling*).

Articulation - 3

For example, when a vowel sound is produced, the vocal cords vibrate and the tongue is in a particular position within the oral cavity. The lips are open.

For a nasal consonant, such as /m/, the uvula is pulled down and sound is allowed to resonate (flow) through the nasal cavity. During /m/ production, the lips are closed for a brief interval.

For a fricative consonant, such as /s/, the vocal folds are held open (they do not vibrate) and air is forced through a narrow opening between the tongue and the alveolar ridge. This produces the noise quality of /s/.

Phonemes

Phonemes are the smallest segment of the signal that, if changed, would produce a different word with a different meaning. Thus, while words carry meaning, phonemes are the units from which words are built.

Different languages have different numbers of phonemes (Hawaiian has 11, Midwestern American English has 39), but all come from a universal set.

All languages divide their inventory of phonemes into vowels and consonants. All languages group phonemes into sequences to form syllables.

Phonemes – Midwestern American English Vowels

- /i/ - heed
- /ɪ/ - hid
- /e/ - fade (diphthong)
- /ɛ/ - head
- /æ/ - had
- /aɪ/ - hide (diphthong)
- /oɪ/ - void (diphthong)
- /a/ - sod
- /ɔ/ - hawed
- /o/ - hoed
- /ʊ/ - hood
- /u/ - who'd
- /aʊ/ - how'd (diphthong)
- /ʌ/ - thud
- /ɚ/ - herd

America English Consonants

stops & nasals	fricatives

- /b/ - **b**at
- /p/ - **p**at
- /m/ - **m**at
- /d/ - **d**ot
- /t/ - **t**ot
- /n/ - **n**ot
- /g/ - **g**ap
- /k/ - **c**ap
- /ŋ/ - sa**ng**

- /f/ - **f**at
- /v/ - **v**at
- /θ/ - **th**in
- /ð/ - **th**is
- /s/ - **s**ip
- /z/ - **z**ip
- /ʃ/ - a**ss**ure (**sh**ip)
- /ʒ/ - a**z**ure
- /h/ - **h**ip

America English Consonants (cont.)

affricates

- /tʃ/ - **ch**ug
- /dʒ/ - **j**ug

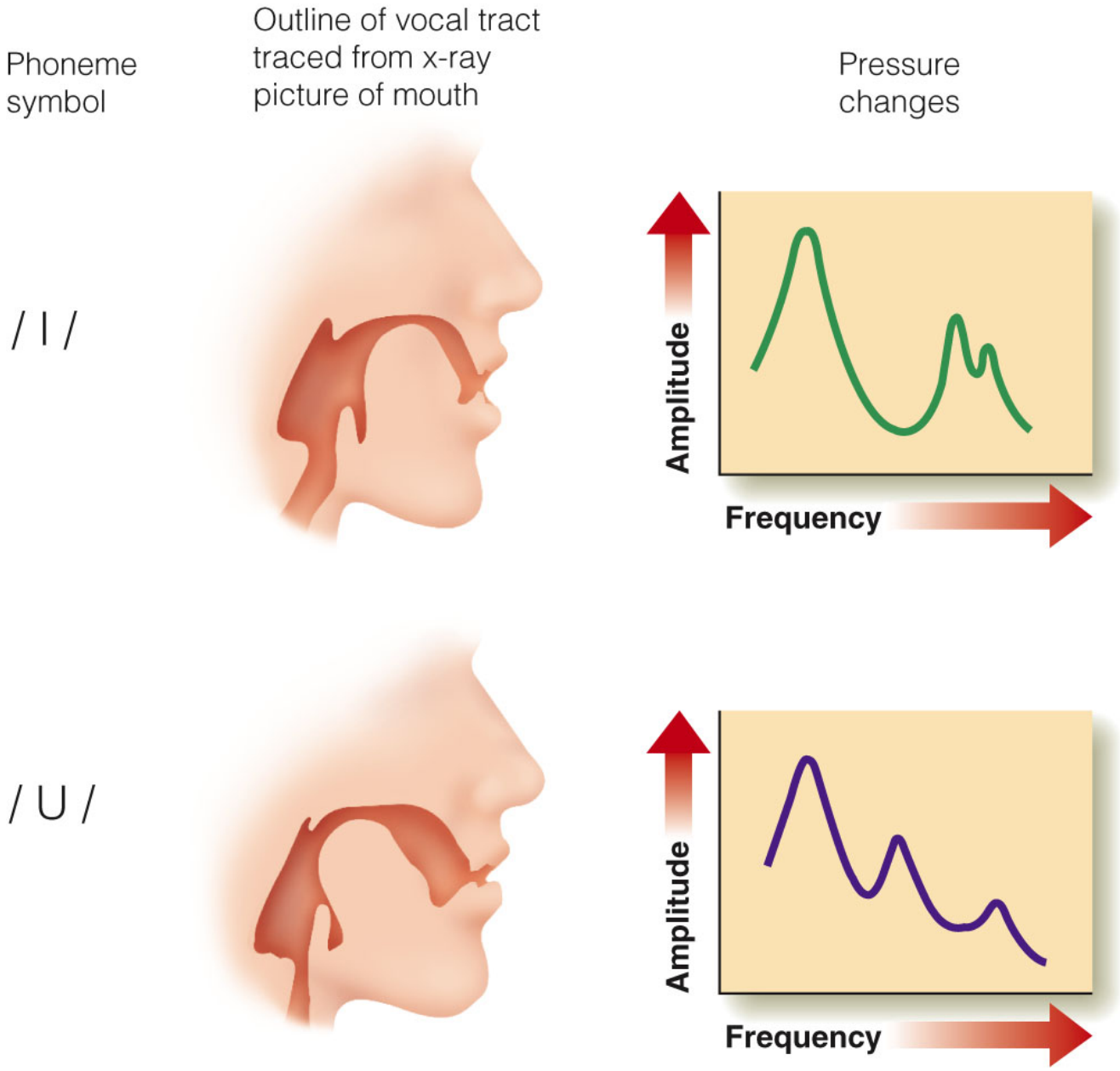
approximates

- /w/ - **w**att
- /l/ - **l**ot
- /r/ - **r**ot
- /y/ - **y**acht

Phonemes - 2

Every phoneme is the result of the coordinated movement of the articulators. Because the movement from one phoneme to the next is continuous, the precise sound that represents a phoneme varies with the nature of what precedes and follows it. This phenomenon is called co-articulation.

The position of the tongue, lips and jaw produces a set of resonators (tubes with a particular length and area, like a pipe organ). These amplify some frequencies and attenuate others. The pattern of this frequency information, over time, is speech.

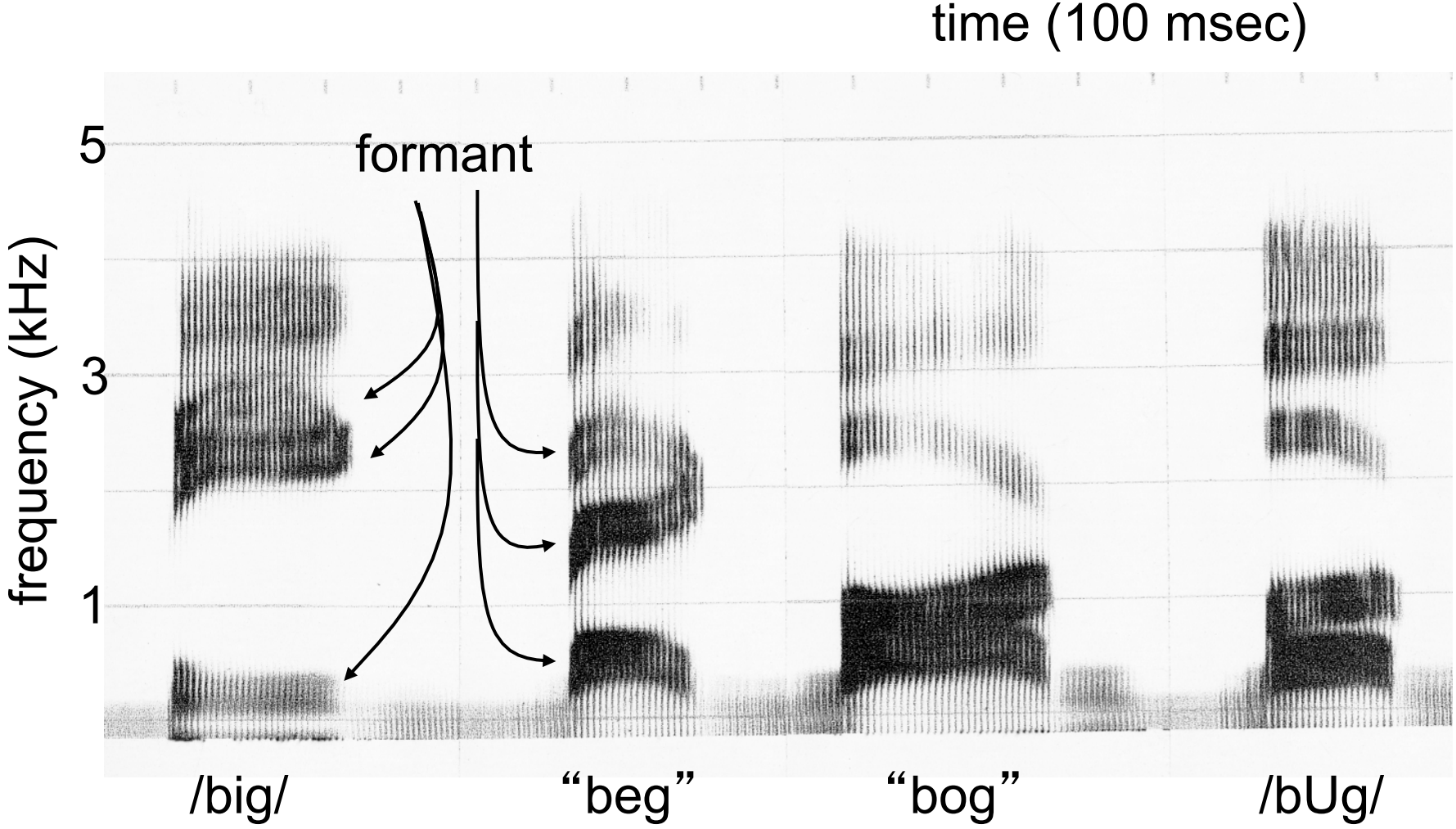

Phoneme symbol
Outline of vocal tract traced from x-ray picture of mouth
Pressure changes
/ I /
Amplitude
Frequency
/ U /
Amplitude
Frequency

The Acoustics of Speech

The speech signal is broken down by the ear into a representation of the intensity at each frequency over time. The sound spectrogram is a similar representation of the acoustic information in speech.

Dark areas are concentrations of energy at a particular frequency. When such a concentration occurs over time, it is called a **formant**. In the next graph, the energy in four syllables that start with the consonant /b/ and end in the consonant /g/ are shown. These syllables differ in the vowel and illustrate the different sounds for these four vowels.

Vowel Examples



Speech Acoustics - 2

The formants in the speech signal are critical information for listeners to recognize the sounds of speech.

For the vowel in “beg”, spoken by this male talker, the center frequencies of the first three formants at the middle of the syllable are approximately 560 Hz, 1750 Hz and 2400 Hz.

For the consonants /b/ and /g/ at the beginning and end of each syllable, the formants change rapidly over time. These changes are called formant transitions and are critical to our ability to recognize consonants and vowels.

Consonants and Vowels

The sounds of speech vary in:

1. The frequencies and intensities of the formants
2. The pattern of change in the formants, over time
3. Voicing (whether the vocal folds vibrate or do not)
4. The presence of nasal formants
5. The duration of 1 through 4 above

The Challenge of Speech Perception

The question of how humans perceive speech is complex because of two classical problems:

1. How is a continuous signal divided up into phonemes, syllables and words?

2. How does the listener recognize the sequence of phonemes, syllables and words when the speech signal changes because of differences in speaker, speaking rate, dialect, and co-articulation?

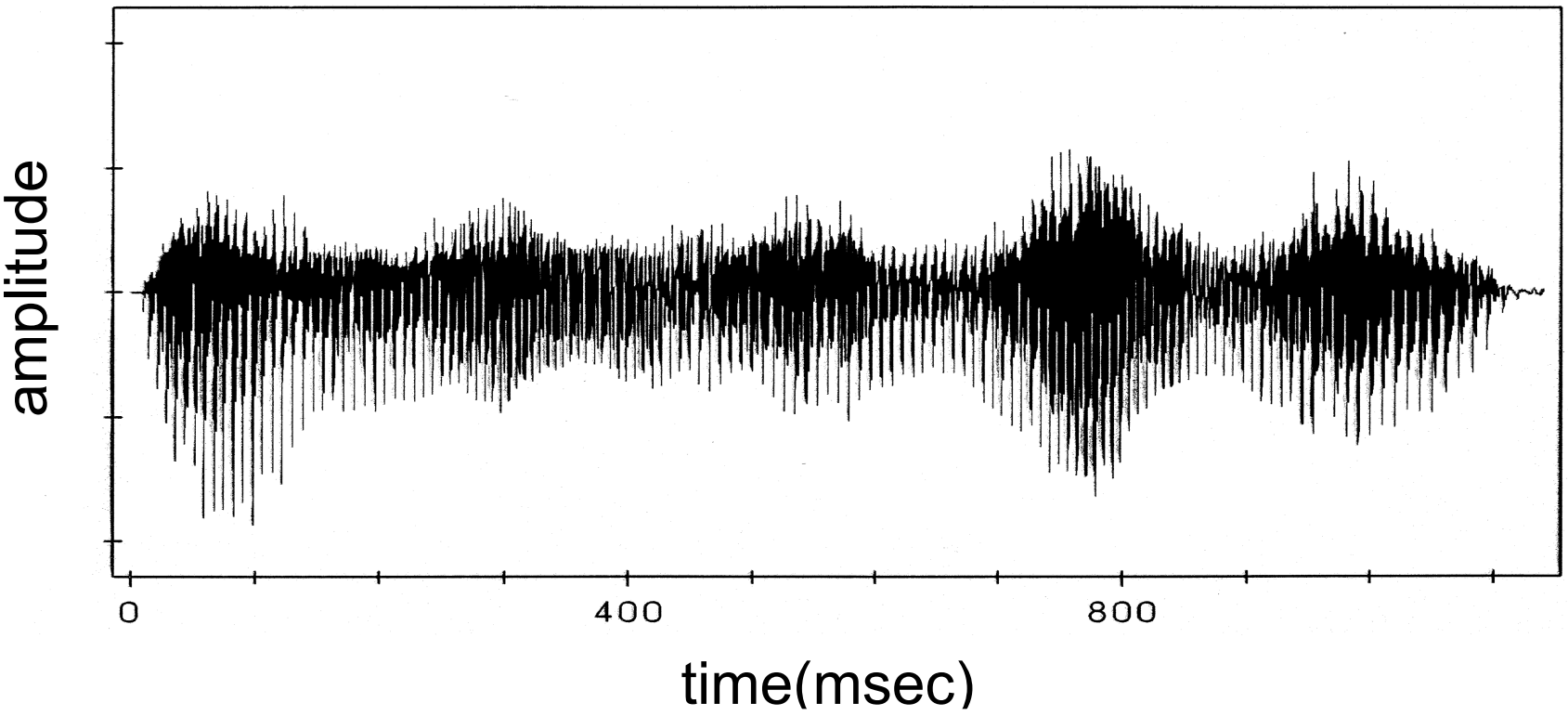
Segmentation

The speech sound is continuous. There are no breaks between words in fluent speech. The effect of co-articulation is to smear the boundaries between adjacent phonemes, syllables and words.

As an illustration, consider the sentence in the following waveform and spectrogram.

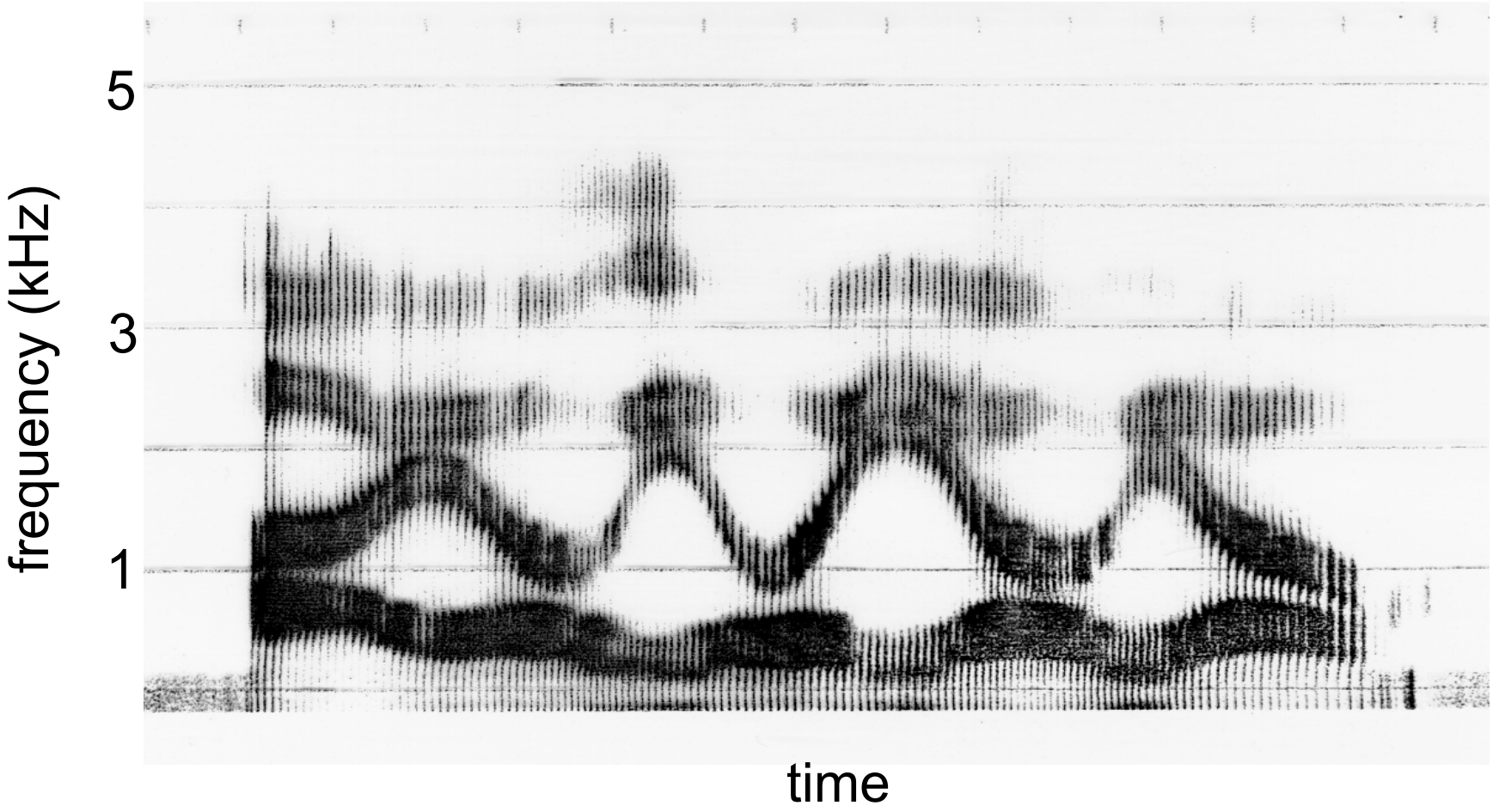
Where does one word end and the next begin?

Segmentation Illustration - 1



In the sound, where are the boundaries between the five words?

Segmentation Illustration - 2



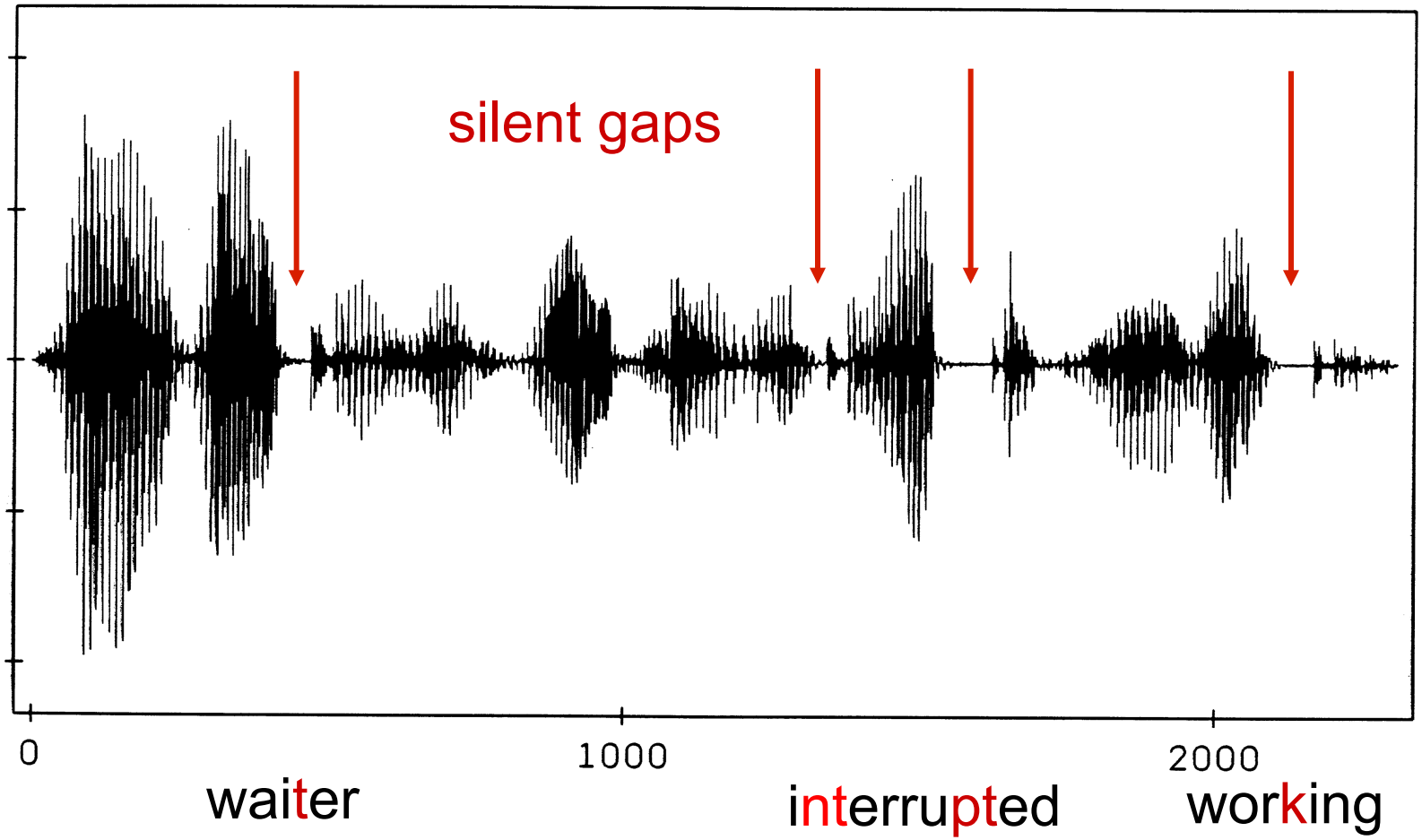
Segmentation - 2

In fluent speech, silence occurs when we take a breath or when a particular type of phoneme occurs (stops) in which the vocal tract is briefly closed.

These silence intervals that occur with stops can be within words or between words.

Segmentation Illustration - 3

"Our waiter was rudely interrupted while working."



Ambiguity in Segmentation

There are sequences of words whose phonetic structure is the same. Consider “ice cream” and “I scream” or “gray tie” and “great eye”.

Our ability to correctly segment these is based on:

1. Context. These words sequences are typically used in conversations about different topics.

2. Details in the sound. The precise acoustic details of a phoneme depend upon whether it is at the end of one syllable or the beginning of the next syllable.

Variability

Variability refers to the changes that occur in phonemes, syllables, and words because:

1. They occur with different other sounds before and after them. This influence of co-articulation alters the sound for a phoneme based on what came before and after.

2. Different talkers have different length vocal tracts and speak with different dialects and idiolects.

3. Talkers vary their rate of speech and the accuracy (carefulness) of their articulation.

Co-articulation

The formant transitions that characterize a /b/ or a /g/ change with the vowel. That is, the acoustic details of /b/ in the words beat, bit, bet, bat, box, bought, boat, book, boot, but, bird, bite, bout, and boy are different.

One of the primary goals of research in speech is to find a way to characterize a pattern of change in the sound which is the same for all examples of a particular phoneme (e. g. /b/s) and distinguishes it from other phonemes. Finding such a pattern (an invariant) for each consonant and vowel has so far proved elusive.

Talkers

1. Individuals have different length vocal tracts and voice pitch (e.g. typical male versus female voice). They speak different dialects with different accents. This leads to different physical sounds that correspond to the same phonemes and words.

2. Individuals vary in how careful or "sloppy" they are in their articulation. In saying "Did you get to know him well", "Did you" is often said as "dija", "get to" becomes "geta", and the "h" in "him" is omitted. In spite of this, listeners have little difficulty understanding speech across this range of variation.

Speaking Rate

A person may intrinsically speak rapidly or slowly. Each individual also varies their rate of speech. This causes segments (phonemes and syllables) to vary in duration.

However, some phonemes such as /b/ and /w/ ("beat" and "wheat") are differentiated by their duration. This implies that listeners adjust their perception for the rate at which the person is speaking.

Like size-distance scaling in vision, this implies that certain properties of the sound must be extracted first to properly perceive the distal object (phoneme and word).

Perception

In speech perception, listeners show evidence of phonetic constancy. They hear the same speech sounds in spite of variation in who is talking, how fast they talk, or other variation in the sound.

From an ecological perspective, this has led to a search for *invariant* properties or features in the sound. When the feature is present, it would signal the listener that a particular phoneme has been spoken.

While some invariant properties have been identified, there are also many phonemes for which no invariant properties have been found yet.

Categorical Perception

One form of perceptual constancy found with speech is illustrated by the phenomenon of categorical perception.

Here, listeners are asked to identify a series of speech sounds. They are also asked to try to discriminate among physically different speech sounds. The interesting result is that in certain circumstances, listeners can not tell physically different sounds apart unless they have different labels.

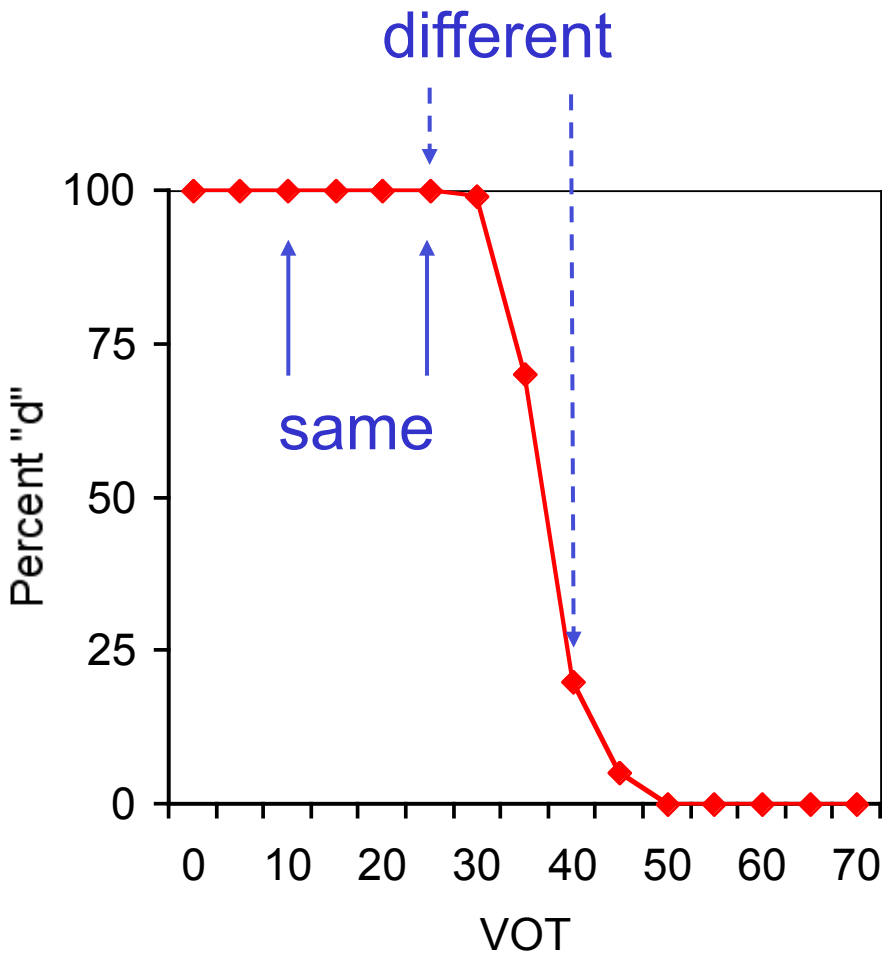
Categorical Perception - 2

For example, a series of speech sounds is made by gradually lengthening the voice onset time (VOT) of the consonant /d/ in /da/ until it is the same duration as that of a /t/ as in /ta/.

When these syllables are presented to listeners for identification, all of the shorter VOT consonants are labeled “d” and the longer as “t”.

If listeners are asked to discriminate between different sounds labeled as “d”, they are at chance. Similarly, discrimination between different sounds labeled “t” is at chance. When sounds have different labels, discrimination is excellent.

Categorical Perception Data



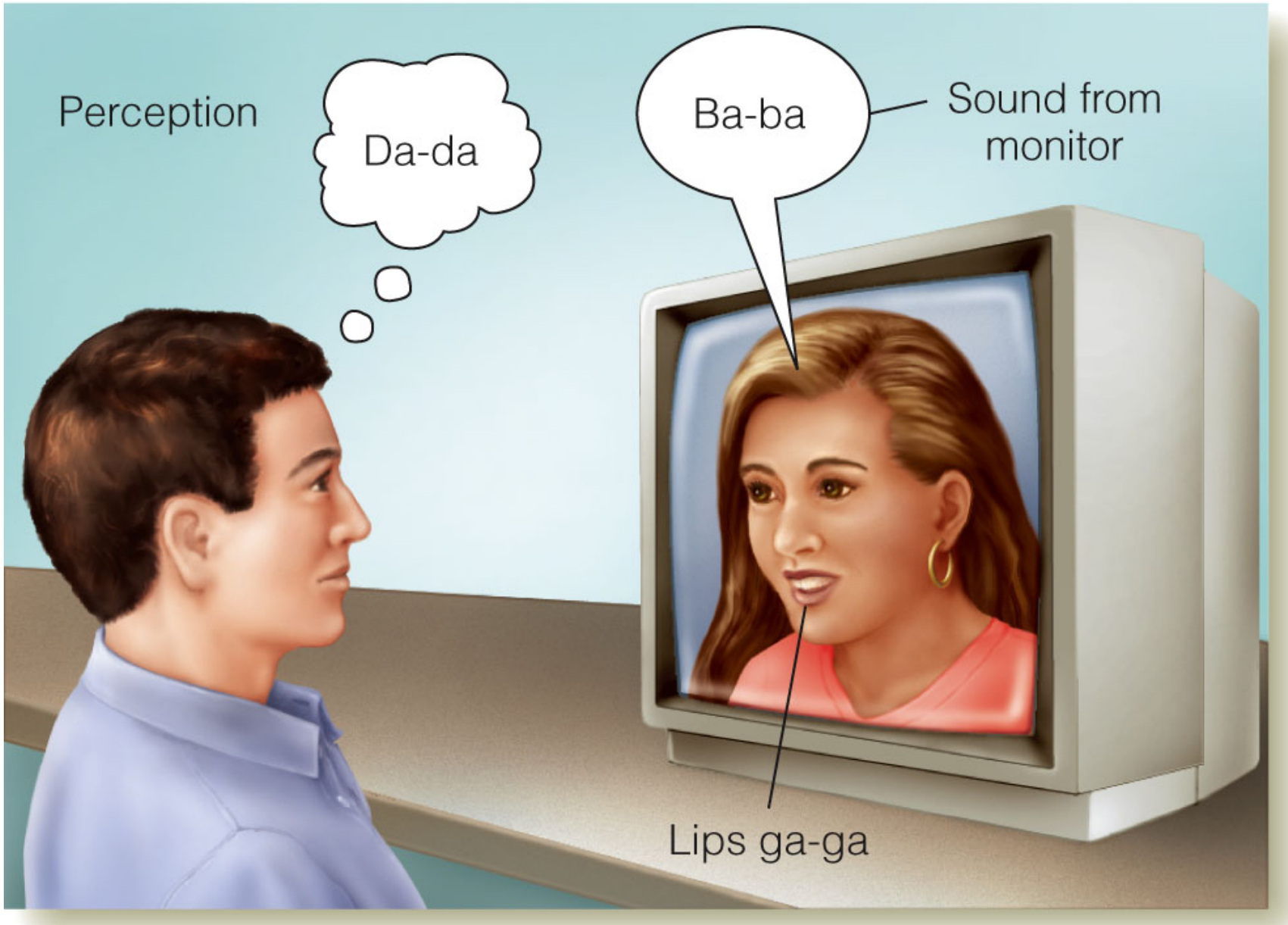
The consonants with VOTs of 10 and 25 msec are both labeled “d” and can not be discriminated accurately.

The consonants with VOTs of 25 and 40 msec are labeled “d” and “t” respectively and can be discriminated accurately.

Speech by Ear and by Eye

The perception of speech can take advantage of visual information (from looking at the face of the speaker) in addition to the sound.

If we edit a video to show a speaker saying /ga/ while the audio track plays /ba/, an observer will report that they hear “da”. If the observer closes her/his eyes, they hear “ba”. Known as the McGurk Effect, this illusion illustrates how listeners integrate information from two sensory systems in speech perception.


Perception
Da-da
Ba-ba
Sound from monitor
Lips ga-ga

Higher Level Influences on Perception

There are many demonstrations of the influence of higher level knowledge on speech perception.

In an early study, Polleck and Pickett recorded sentences such as "The time is five to nine". These were played to listeners who were to report each word. Then, they removed words from the sentences (e.g., "nine") and played them to listeners who were to report what they heard. Listeners were much more accurate for the words in sentences than in isolation.

Higher Level - 2

Another example is “phoneme restoration”. In a recording of “legislature”, the /s/ is removed. If the word with the missing sound is played, listeners report that a sound is missing (the /s/).

If a recording of a cough is overlayed over the missing /s/ word and this is played to listeners, they report a normal word. That is, they report hearing the /s/, even though it is not present. They also have difficulty correctly locating where, in the word, the cough occurred.

Phoneme Restoration (cont.)

In a sense, this illusion is like the Gestalt law of good continuation. However, it is based on knowledge since the effect is strong with real words and weak or non-existent for non-words.

Talker Familiarity

Suppose we train a group of listeners to identify the voices of a set of talkers (Nygaard et al.). They hear words, spoken one at a time, by one of 10 talkers. They are asked to identify the talker.

Later, they are given a word perception task. Some listeners hear new words spoken by the 10 familiar talkers. Other listeners hear these same new words spoken by different (new) talkers.

Accuracy is higher for the words spoken by the familiar talkers.

Higher Level Influences - Summary

The perception of speech is a result of both bottom-up processing of the sound and the top-down influence of our knowledge of the words of the language, our knowledge of meaning and language structure, and our experience with the voice that we are listening to.

As a last example, have you ever listened to someone speak a language that you do not know (e.g. Navaho)? Can you identify where one word ends and the next begins?

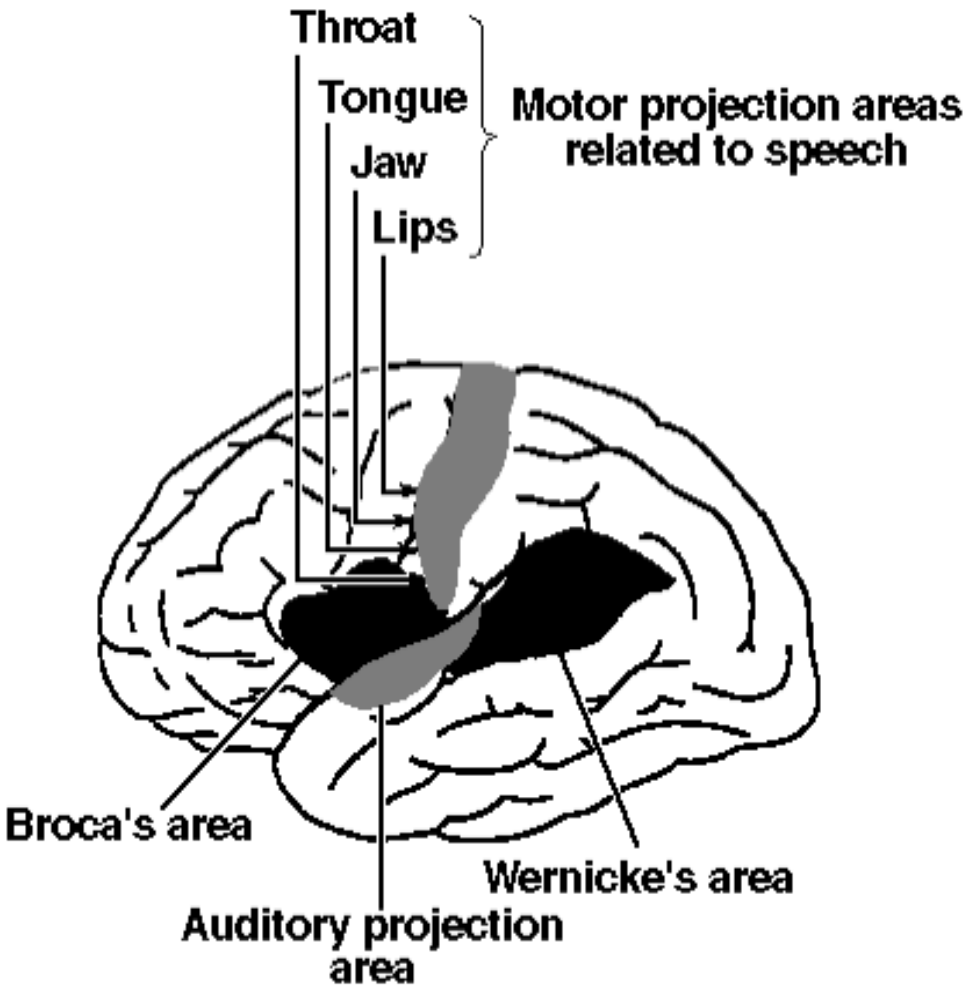
Physiology of Speech Perception

Human speech is unique to humans. Consequently, we can not study single-cell recording in animals of speech perception *as it occurs in humans*.

We do find cells in the auditory cortex of animals that respond to "features" of sound similar to the formant transitions and formant combinations that occur in speech.

There is also an area of monkey cortex that contains cells that respond to "monkey calls". The analogous area in humans is involved in spoken language.

Localization of Language Function



Using brain imaging techniques and studying patients with brain damage, we find that certain brain areas play specialized roles in language.

Localization of Language - 2

Damage to Brocca's area leads to individuals who have difficulty producing speech, particularly complex sentences. They show little or no problem with listening.

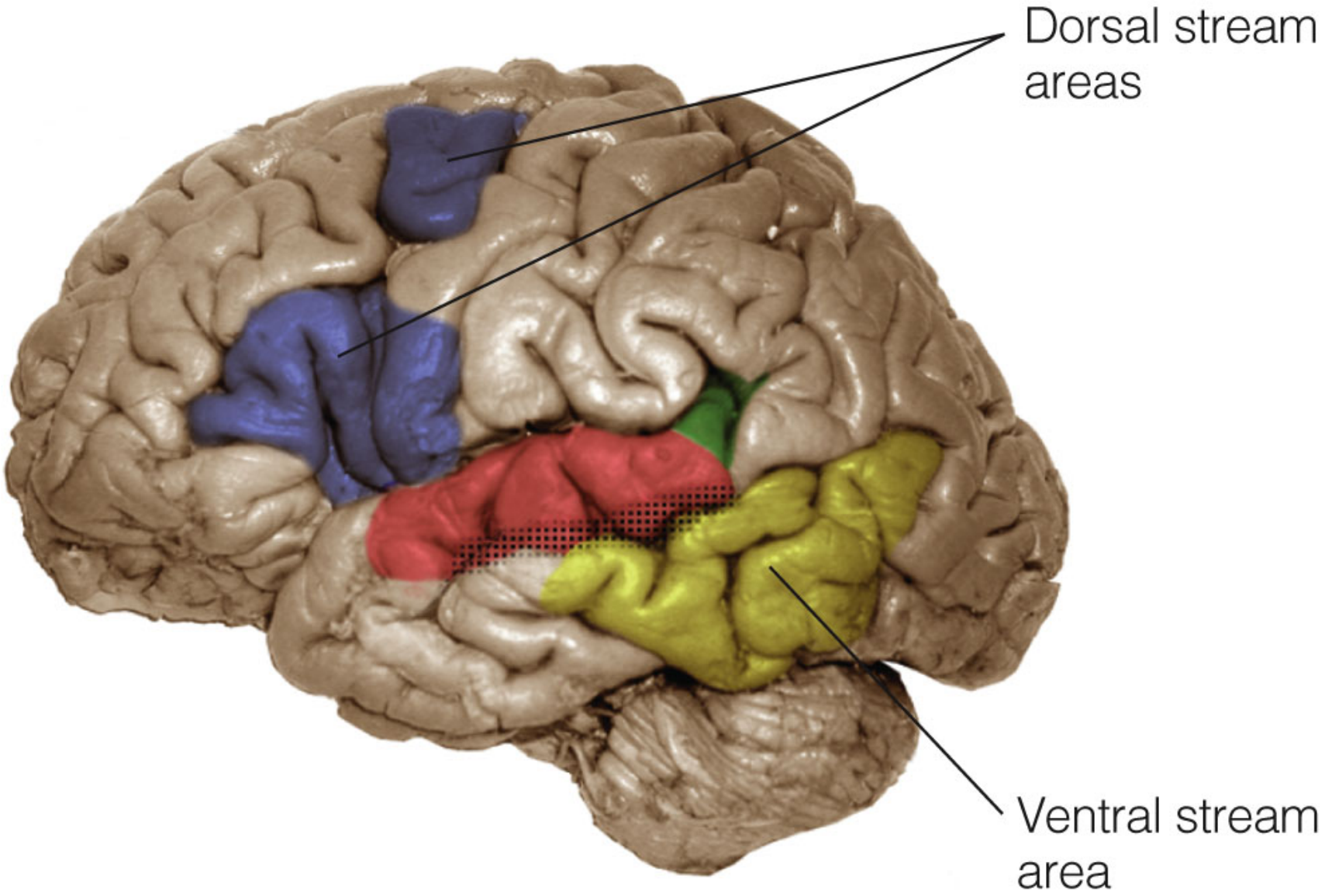
Damage to Wernicke's area leads to individuals who produce fluent, but often meaningless speech and who have problems with language comprehension.

Both of these areas are on the left side of the brain.

What and Where Pathways

There is an area in the superior temporal sulcus (STS) that is more activated by human voices than by other sounds (based on fMRI imaging). The STS is a part of the ventral stream in hearing that seems to be specialized for identifying sounds.

In the dorsal (where) stream, we find areas that integrate auditory and visual information, perhaps including mirror neurons that respond to both how something is said and how it is heard.


Dorsal stream areas
Ventral stream area

Plasticity and Learning

Human infants can distinguish speech sounds from one another (e.g. /d/ from /t/). This includes the ability to distinguish /r/ and /l/ (as in “lake” and “rake”) for infants in both the U.S. and Japan. By 12 months of age, the U.S. children can still do this distinction but the Japanese children can no longer do this distinction.

There are correlated changes in EEG activity. that occur with language experience. These changes may help to explain the difficulty that adults have with learning the sounds of a new language.

Physiology and Speech Summary

There is brain specialization for language.

Does the perception of speech involve processes and/or brain mechanisms that are specially adapted to deal with speech?

The data from brain imaging studies and lesion studies are not conclusive. The processing of speech and language is widely distributed in the cortex. However, it does appear to involve two different “pathways” that handle different aspects of our perception.