Overview

1) Speech articulation and the sounds of speech.

2) The acoustic structure of speech.

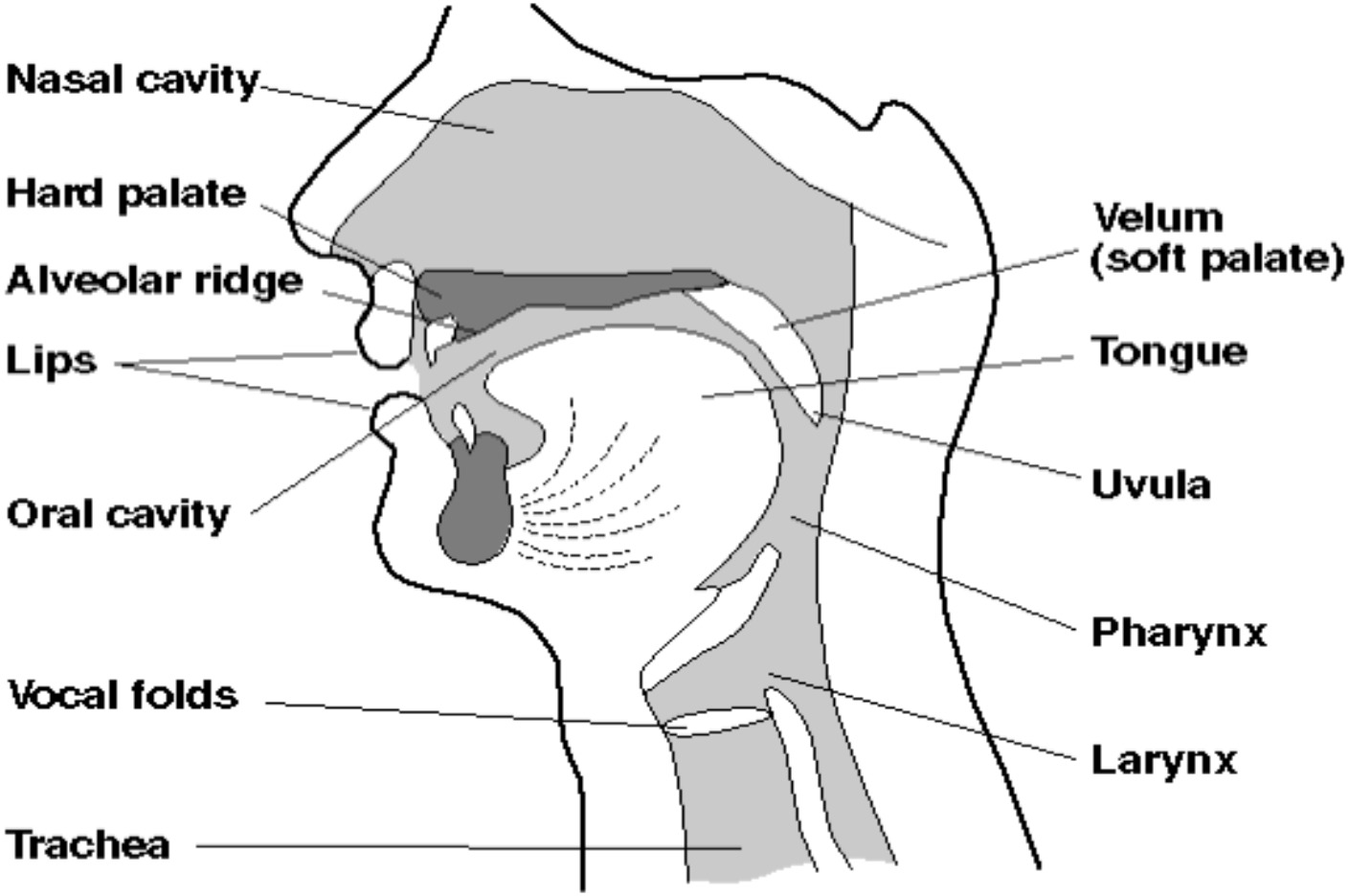
3) The classic problems in understanding speech perception: segmentation, units, and variability.

4) Basic perceptual data and the mapping of sound to phoneme.

5) Higher level influences on perception.

6) Physiology of speech perception and language.

Vocal Tract


Nasal cavity
Hard palate
Alveolar ridge
Lips
Oral cavity
Vocal folds
Trachea
Velum (soft palate)
Tongue
Uvula
Pharynx
Larynx

Articulation - 1

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 movement of the articulators takes time (they have inertia). The movements are rapid (for communication). This leads to the phenomena of coarticulation. The preceding segment alters the precise realization of the current segment (inertia or perseveration). The next segment also alters how the current segment is realized (planning or anticipation).

Articulation - 2

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 (either spread or rounded).

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/.

Articulation - 3

Basic dimensions of articulation:

1) Voicing - Vocal folds vibrate or are held open
   (voiced or voiceless)
2) Nasalization - Nasalized or not (uvula closed)
3) Place - Location in vocal tract of constriction
   bilabial, labiodental, inter-dental, alveolar, palatal, velar, glottal
4) Manner - Degree of constriction in vocal tract
   (open, moderate, constricted and closed)

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. /m/ and /b/ are different phonemes in English because /mæd/ (mad) and /bæd/ (bad) are different words.

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 - 2

Every phoneme represents the coordinated movement of the articulators that results in a different sound. 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 coarticulation.

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

Some Terms

The terms to describe the segments include allophonic, phonetic, phonemic and phonological:

Phonetic – a description of the sound segment that includes details of production. [p^h] vs [p]

Phonemic – a description of the sound segment where anything predictable is omitted. /p/

Allophonic – variations within a segment category (phonetic) that do no change the identity of the segment (phonemic).

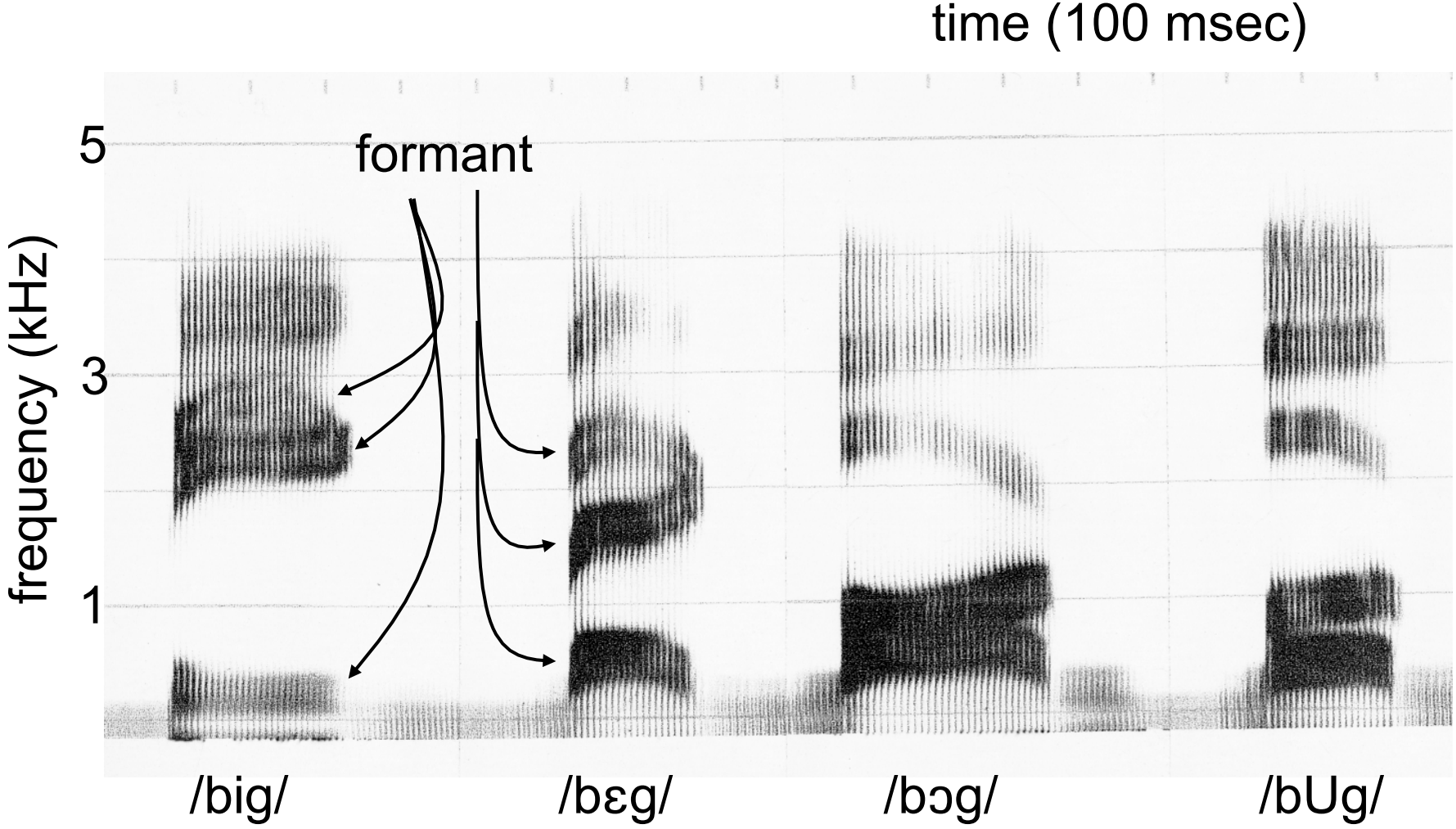
Phonological – The sound segment inventory and constraints on sequences.

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 sound for these four vowels.

bVg Examples



Speech Acoustics - 2

The formants in the speech signal vary with the position of the tongue, lips and jaw. They are a "cue" for listeners to recognize the sounds of speech.

For the vowel in /bɛg/ (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 not)
4) The presence of nasal formants
5) Presence of noise in the spectrum and the frequency/ intensity distribution of the noise
6) The duration of 1 through 5 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 coarticulation?

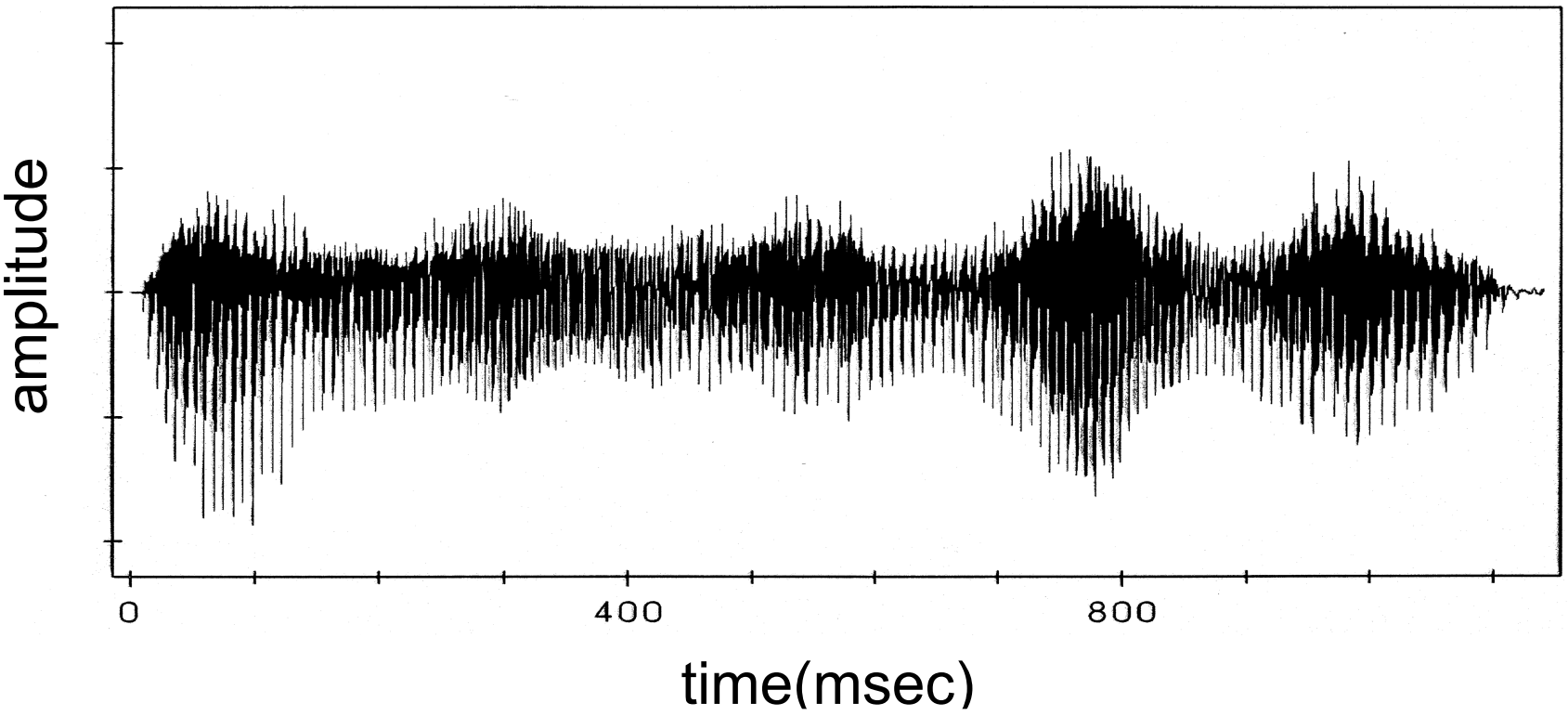
Segmentation

Speech is continuous. There are no breaks between words in fluent speech. One effect of coarticulation is to smear the boundaries between adjacent phonemes, syllables and words.

As an illustration, consider the following sentence.

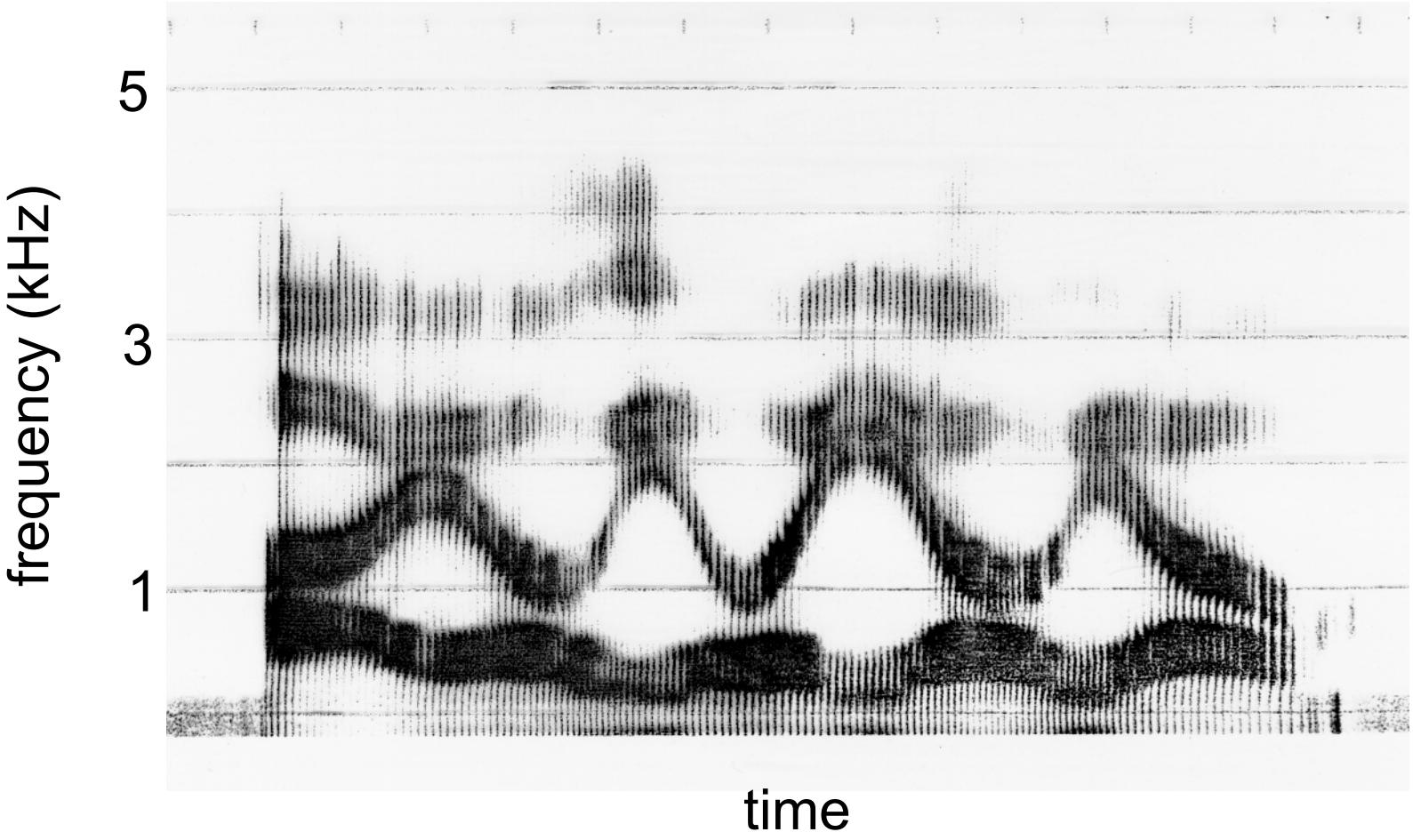
Where does one word end and the next begin? The sentence is shown first as a waveform then as a spectrogram.

Segmentation Illustration - 1



In this sentence, where are the boundaries between words?

Segmentation Illustration - 2



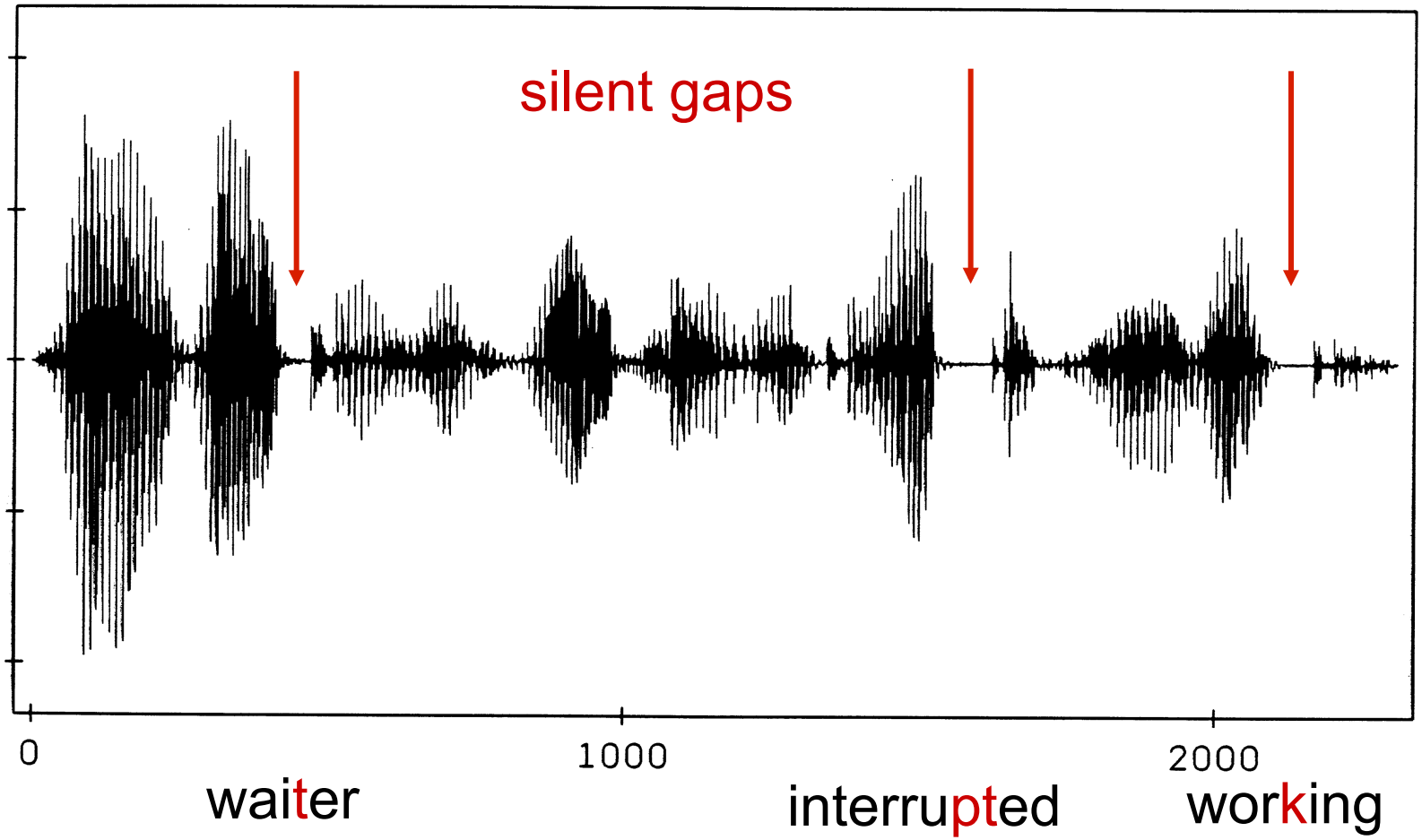
Segmentation - 2

In fluent speech, silence occurs when we take a breath, when we deliberately pause for emphasis (or to think), 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."



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 coarticulation 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.

Variability: Coarticulation

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 (an invariant). Finding such a pattern for each consonant and vowel has so far proved elusive.

Variability: Talkers

1) Individuals differ in vocal tract length and size of their larynx. 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 relatively little difficulty understanding speech across this range of variation.

Variability: 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 (in part) 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.

Variability: Environment

A person listens to speech in many different environments.

They may hear a person speaking against a quiet background, against a background of environmental noise (e.g., a busy street) or against a background of other conversations.

How does a listeners recover the intended message? How do they separate the aspects of sound that belong to the speech signal that they are trying to recognize from other sound or other speech?

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

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. That is, a listener is more accurate in their perception when they have both auditory and visual information.

This can also lead to illusions. 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 illustrates how listeners integrate information from two sensory systems in speech perception.

Articulation of Vowels

In English, vowels are voiced, non-nasal and their manner is vocalic and continuant (very little constriction).

They are classified according to the position of the tongue and the shape of the lips. The point of maximal constriction in the oral tract with the tongue can be front, mid or back. The height of the tongue in the vocal tract can be high, mid or low. The lips can be rounded (round opening) or spread (wide, shallow opening). The vowel can be tense (long duration) or lax (short duration).

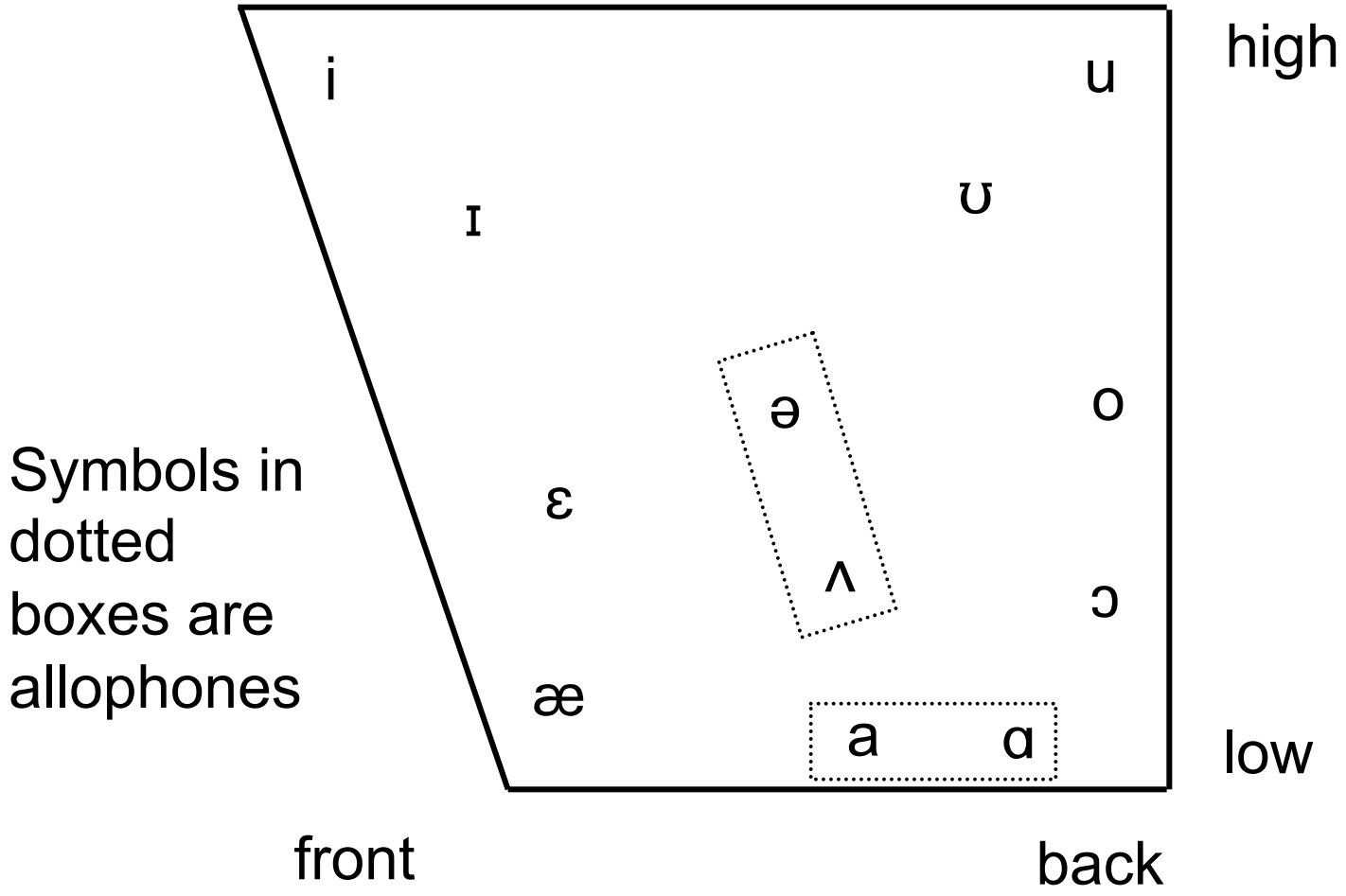
Articulation of Vowels - 2

For example, /i/ as in beat is a high (height), front (front to back), spread (lip rounding), tense (long duration), non-nasal vowel.

/ʌ/ is a low, mid, spread, lax vowel.

The next diagram show the position of the maximal constriction of the tongue in vowel production for most of the vowels of American English.

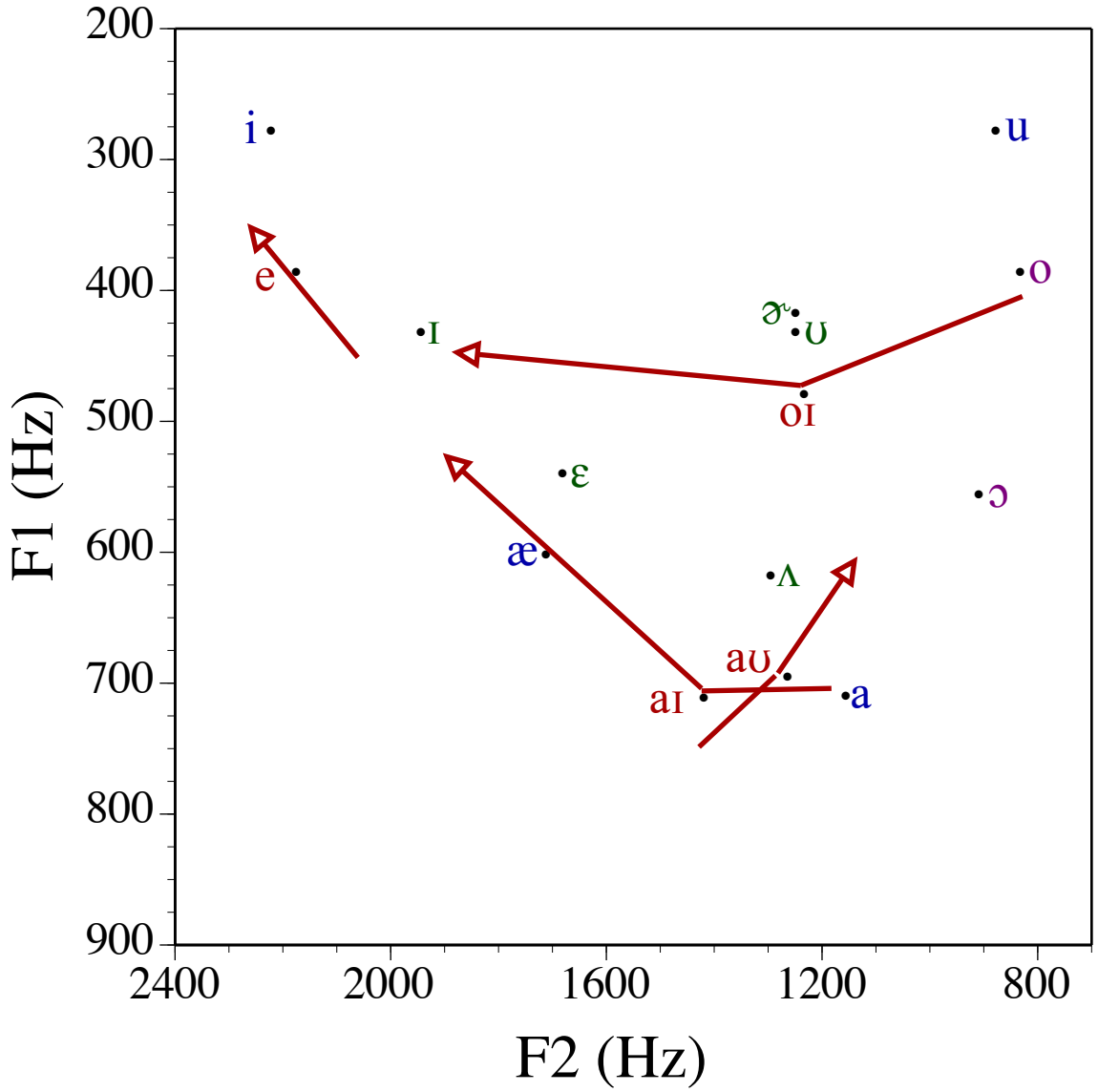
American English Vowel Space



Articulation of Vowels - 3

The vowel /ɚ/ was not show in the diagram because it is distinguished by the curvature of the tongue and not the tongue position or height, where it is similar to /ʊ/ or /ə/.

The diphthongs were not shown since they are characterized by movement of the tongue over time.



A Talker's Vowel Space

Acoustics of Vowels

Plotted as a function of the first formant (F1) and the second formant (F2) frequencies

The diphthongs are shown as an “arrow” representing their formant frequencies at the beginning, middle and end.

Other Languages

There are attributes of vowel articulation that English does not use.

1) Nasalization. A vowel can be nasal or non-nasal. French and Hindi use this distinction.

2) Long or short (duration). A vowel can be long in duration or short. Japanese uses this distinction.

Note that English speakers make long and short, nasal and non nasal vowels, but these are allophonic variations.