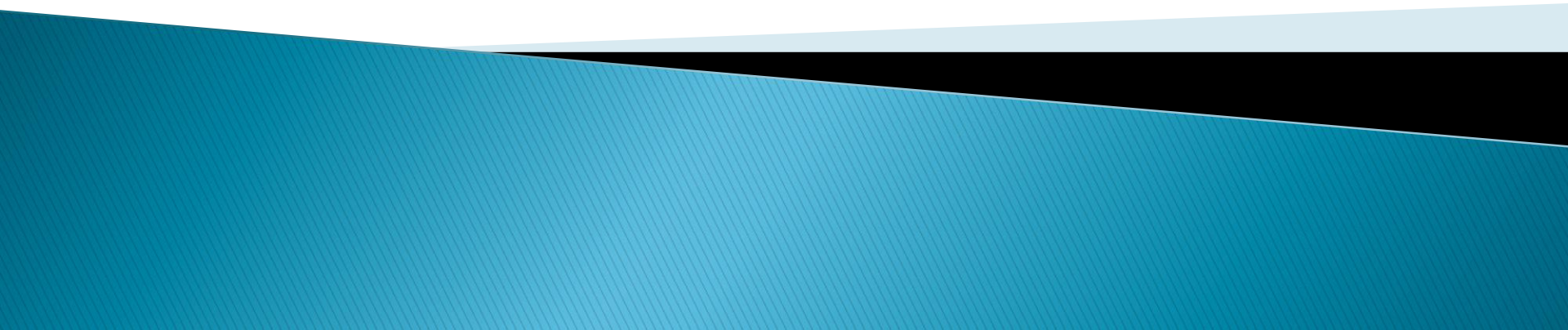
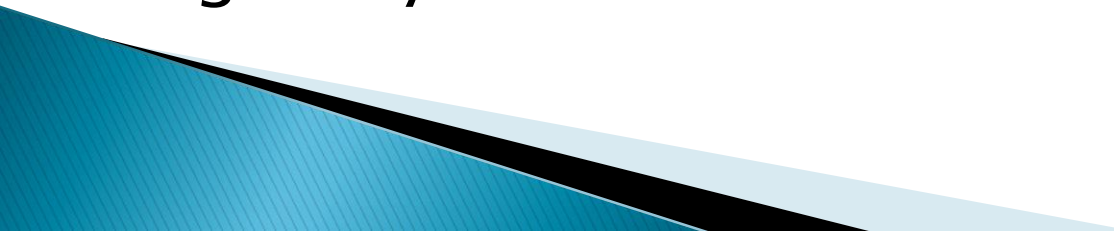


**Psycholinguistics**  
**Dr. Nesreen I. Nawwab**  
**2014–2015**  
**First Semester**  
**Lecture 3**

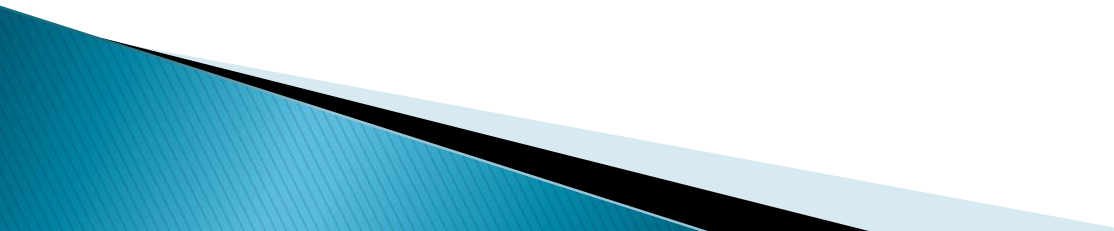


# ▶ Speech perception

Perception of phonetic segments

- ▶ Introductory remarks:
  - ▶ One of the greatest challenges facing speech perception researchers is to determine how individual sounds are isolated (segmented) from the complex speech signal and how they are identified appropriately.
  - ▶ We should remember that phonetic segments are not like beads strung on a string, one segment after another, rather, it is better to compare speech to a braid in which the properties that help us identify phonetic segments are tightly intertwined and overlap greatly.
- 

## The “Lack of Invariance” Problem

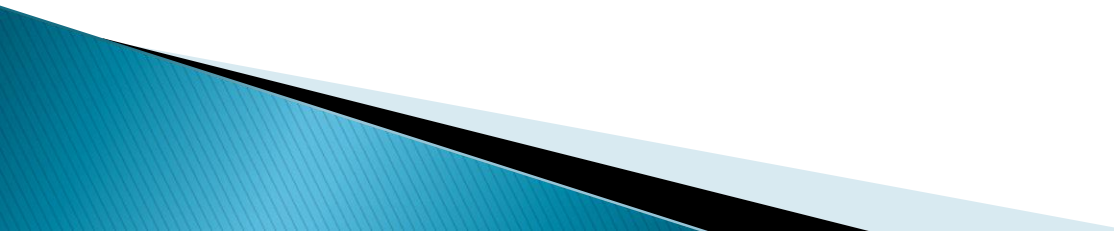
- ▶ It would be relatively easy to develop models of the speech perception process if each distinctive sound in a language was associated with a standard acoustic pattern. However, rather than displaying **invariant** (standard, unvarying) patterns, speech sounds vary considerably in their acoustic characteristics for several reasons:
- 

1. The production, and hence the acoustics, of the same phonetic segment varies depending on the context in which the segment is produced. These context effects, which result in overlapping movements for speech, are called **coarticulation effects**, e.g. **allophonic variation**.
2. The physical properties of speech sounds, especially vowels, vary according to whether they have been produced by men, women, or children, whose vocal tracts differ in size and configuration.
3. We do not pronounce the same utterance in exactly the same way twice.
4. Another factor stems from the properties of rapidly articulated conversational speech. There is a great difference between saying single words slowly and carefully and the way we actually pronounce words when we speak fluently.

- ▶ Sometimes speakers **underarticulate** (miss articulatory targets), so much so that the words lose much of their identifying information. Yet, listeners usually have little trouble understanding such speech samples.

- ▶ **In listening to others speak we appear to have no problem in dealing with these variations, and Speech perception research must explain how listeners process such “messy” samples of speech.**

# How Is Speech perceived Under Less Than Ideal Conditions?

- ▶ We will show how lexical, syntactic, and contextual information is used to interpret **ambiguous** (unclear) speech signals. Models of speech perception will need to explain how these other levels of processing contribute to the process of speech understanding.
- 

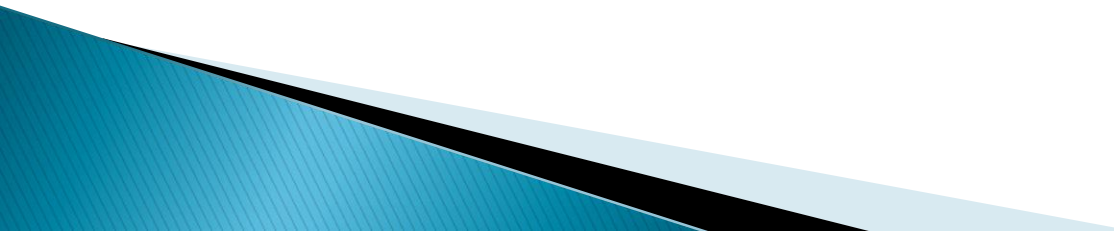


# Quick notes on how speech is produced.

## Basic Terms

- ▶ 1. Place of articulation
- ▶ 2. Manner of production
- ▶ 3. Fundamental frequency: The rate at which glottal pulsing (voicing or vibration of vocal folds) occurs during sound generation (phonation). It is about 125 glottal pulses per second for adult males, about 200 pulses per second for adult females, and about 300 pulses per second for children.

Linguists have used concepts such as the above to develop a system of distinctive features for describing speech sounds, in which sounds are described by the feature + voice or -voice, etc.

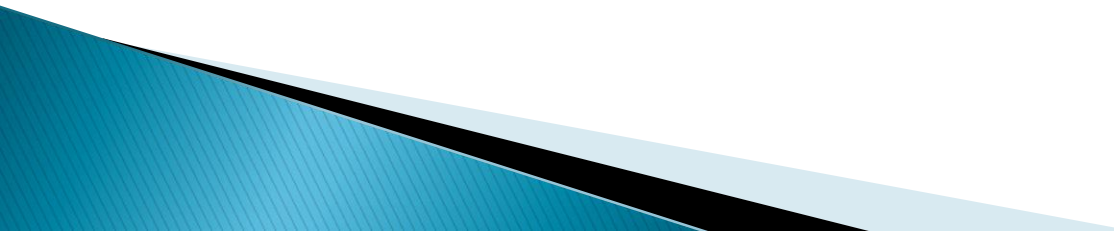
- ▶ Some speech errors suggest that **distinctive features** may be real “**building blocks**” in the speech production process. It is possible to find examples of speech errors in which a given feature, such as voicing, is misplaced, which produces the unintended output, “Baul and Peth”, for the intended sequence, “Paul and Beth.”
- 

# Acoustic Properties of Speech Sounds

- ▶ **1. Vowels:** The vowels we hear are based on a modification of the sound source in a manner that is determined by the **resonant characteristics** of the oral cavity or vocal tract during the production of that sound.
- ▶ **What are resonant characteristics?**
- ▶ Explain it through the analogy of filling an empty bottle with water. **The resonance of a relatively empty bottle is low pitched and the resonance of a bottle that is about to overflow is high pitched.**

## ▶ What are formants?

- ▶ The bands of resonant frequencies for speech change in relation to the movement of our articulators while producing speech. These bands of resonant frequencies are called **formants**.
- ▶ Spectrograms display frequency on the vertical axis, time on the horizontal axis, and amplitude in the darkness of the markings. **Formants** (bands of resonant frequencies) are easily visible on the sound spectrograms: they are the horizontal dark bands.
- ▶ Vowels are differentiated by the relative position of the first two formants. Perceptually, the first two formants are sufficient for their identification. Thus, the combination of a low frequency F1 and a high frequency F2 is characteristic of /i/, shown in figure 3.3A. The pattern that identifies /u/ consists of two low frequency formants, as in figure 3.3B.

- ▶ Diphthongs, which are two vowels produced in a smooth glide, have formants moving from one vowel to another. These movements are called **formant transitions**.
  - ▶ Single vowels produced in isolation do not have formant transitions. These relatively flat formant patterns are called **steady states**. In fluent conversational speech we hardly ever see vowels produced in steady state, and vowel formants have fairly sharp transitions going in and out of adjacent consonants.
- 

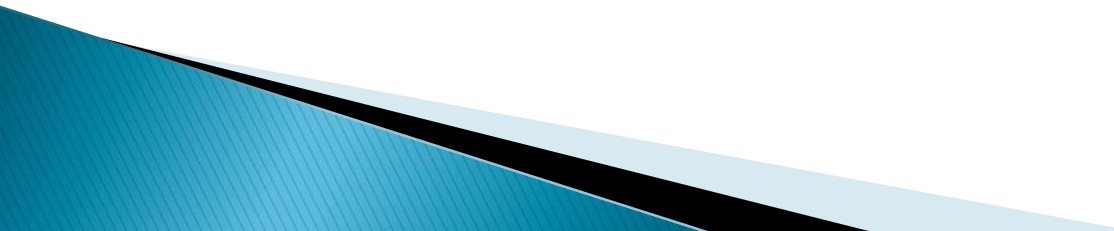
# Perception of Phonetic Segments

- ▶ One of the most important goals of research in this area has been to isolate specific aspects of the complex sound pattern necessary for the identity of a given phoneme. These critical parts of the complex sound pattern are called **acoustic cues**. Researchers required certain equipment before they could begin to work in this area. They needed speech analysis machines such as the sound spectrograph and also the capacity to synthesize according to precise specifications. The aim was to create speech stimuli that could be used to evaluate the perceptual relevance of acoustic cues.

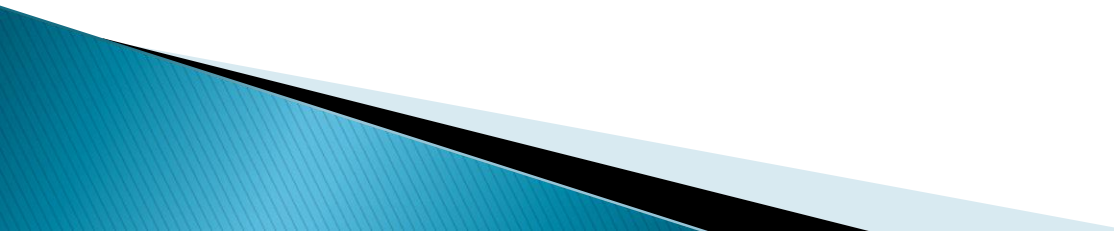
# The Role of Speech Synthesis in Perceptual Research

- ▶ In the early 1950s, Franklin Cooper, an engineer, Alvin Liberman, a psychologist, and Pierre Delattre, a linguist, joined forces to study the perception of speech. It was true then and is true now that progress in speech perception relies on interdisciplinary cooperation.
- ▶ They utilized the Pattern Playback Speech Synthesizer constructed by Cooper and his colleagues. This machine synthesizes speech sounds by converting visual pattern to complex sound waves. The pattern playback device could synthesize speech from drawn formant patterns; it could generate the sound associated with any pattern that one painted on its cellulose acetate belt. The researcher would paint a pattern on the belt, and the machine would play the pattern back to produce the desired sound. If not, they would modify and replay the pattern until they get the desired sound. In this process they discovered the acoustic cues necessary to identify that particular speech sound.

# Ways in Which Speech Perception Is Tested

- ▶ Many experiments in speech perception have made use of two tasks: discrimination and identification.
  - ▶ **Discrimination** tasks require the listener to indicate whether two stimuli are the same or different.
  - ▶ **Identification** tasks require the listener to label or determine the identity of the stimulus, e.g. write the word you hear, or choose the alternative that best matches the label (MCQ questions).
- 



- ▶ **Perception of vowels:**
  - ▶ **What** is the most important part of the vowel in establishing its identity?
  - ▶ **How** do listeners respond to one-and-two formant steady-state stimuli?
  - ▶ Listeners were able to perceive some vowels created with only a single formant. This finding suggests that the frequency information contained in one formant is enough to provide the listener with the perception of a vowel, though not its exact identity. When stimuli that contained two formants were presented, agreement across listeners was high in identifying the stimuli.
- 

## Steady States Versus Formant Transitions in Vowel Identification: An Illustrative Study.

- ▶ Vowels contained in regular words are produced in the context of consonants. Acoustically this means that vowels are marked by formant transitions going in and out of the adjacent consonants and contain steady-state segments that are either short or not present at all.
- ▶ The aim of the study was to compare the perceptual saliency of vowel steady states and formant transitions. (detailed description of the study on page 125–126).
- ▶ **Interpretation of the results:** The authors interpreted their results to indicate that formant transitions and vowel duration are more important cues to the identity of vowels than a fixed sample of the steady-state information.

- ▶ **Perception of Consonants:**
- ▶ In both conversational speech and laboratory studies, vowels are perceived more accurately than consonants.
- ▶ **Why?**
- ▶ The short duration and lower amplitude of consonants make them harder to perceive than vowels.
- ▶ **Stop consonants:** Unlike other consonants, stops lose their identity when presented in isolation. For example, in a syllable such as /ba/, it is impossible to separate the /b/ portion from the /a/ portion of the syllable. If one removes the entire vowel (transition plus steady state) from the syllable, the resultant segment sounds like a “chirp” rather than the phoneme /b/. Stop consonants in syllable initial position must contain a small piece of the transition segment in order to be perceived accurately.

- ▶ In other words, the acoustic signal from the articulation of the stop consonant plus the formant transitions into the adjacent vowel are necessary before we can hear the consonant. It is as if the consonant and the vowel information are merged together in the syllable, somewhat like the analogy of the braid mentioned earlier.
- ▶ When the acoustic information of adjacent phonetic segments merges, the phonetic segment is described as being encoded. Therefore /b/ is encoded because we can isolate no single acoustic segment that would sound like /b/. Further, the information about the consonant and the vowel is transmitted in parallel to the listener. This is called **parallel transmission**, which is highly evident in stops, while other consonants are encoded to a lesser degree. That is all sounds are affected by the characteristics of their adjacent phonemes, a phenomenon referred to as **coarticulation**.

## Phoneme Identity is Context Dependent

- ▶ Since the late 1950s, a great deal of research has been devoted to other types of perceptual studies primarily concerned with determining the acoustic cues for all phonetic consonants in English and other languages.
  1. The first finding is that acoustic cues are highly dependent on context effects. That is, not a single acoustic cue is present in all instances of a given phoneme. The acoustic cue for a phoneme changes as that phoneme is paired with other phonemes.
  2. The second finding is that more than one acoustic cue exists for differentiating a phonetic contrast, **Voice-onset-time** is considered the best single measure for differentiating voiced and voiceless stops. However, additional cues also contribute to the perception of voicing distinction.

## Voice-onset-Time: An Important Acoustic Cue

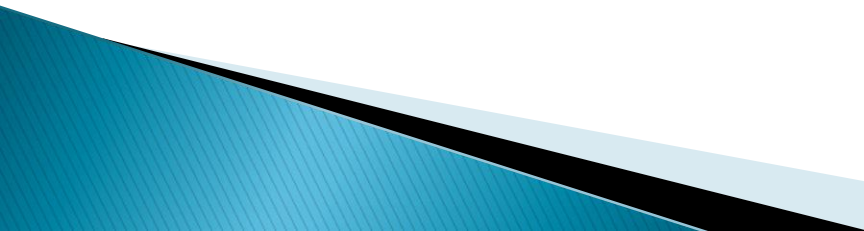
- ▶ In English, three pairs of stop consonants are identical except for the voicing feature: /b/ and /p/, /d/ and /t/, and /k/ and /g/.
- ▶ Although it might appear that the presence or absence of voicing during stop production is a rather easy cue to the discrimination of sounds, describing listeners' actual discrimination between voiced/voiceless stop **cognates** (sounds that differ only in one feature, in this case the voicing feature) turns to be more complex than anticipated. Researchers had already noted from spectrograms of speech samples that voicing distinctions were associated with different acoustic patterns depending on the phoneme and where in the word the contrast occurred. In the case of word initial contrasts, one parameter important for distinguishing between pairs such as /ba/ and /pa/, for instance, was **voice-onset-time (VOT)**.

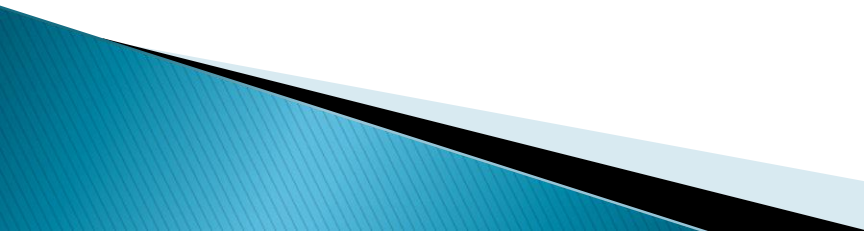
- ▶ **What is voice–onset–time?**
- ▶ In a stop–initial CV syllable, VOT represents the time between the release of air pressure (the burst) and the onset of vocal–fold vibration (voicing) for the adjacent vowel (See figure 3.8A and 3.8B, page 130).
- ▶ **Categorical Perception of Voicing Contrast:**
- ▶ Read the experiment of /ba/ versus /ta/ on page 130–131.
- ▶ We can see that stimuli 1, 2, and 3 were consistently identified as /da/ and were never labeled as /ta/. On the other hand, stimuli 5,6,and 7 were always labeled /ta/ and never as /d/. In the case of stimulus 4, the results are mixed. In 50% of the trials it was heard /da/, but the rest were heard /ta/. Stimulus 4 is called **cross–over stimulus**.

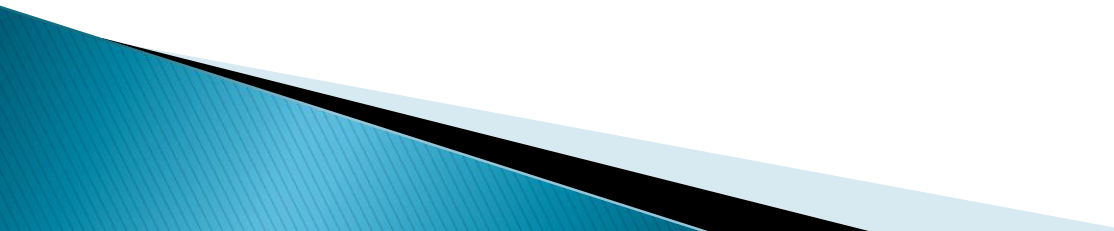
- ▶ Further, when a next stimulus in the continuum is perceived as a different phoneme, this sharp shift in perception is suggestive of a *perceptual discontinuity* across a continuously varying physical dimension. This pattern in response is characteristic of a perceptual phenomenon called **categorical perception**.



# Speech Perception Beyond A Single Segment

- ▶ Introductory remarks:
  - ▶ We will review the perception of speech signals longer than a single phoneme starting with two adjacent phonetic segments and concluding with fluent speech.
  - ▶ Almost all experiments point to the role that our expectations play in speech perception, i.e. knowledge of phonological sequences, the topic of conversation (semantic factors), and our expectations of appropriate syntactic structure to arrive at an image of what we have heard.
- 

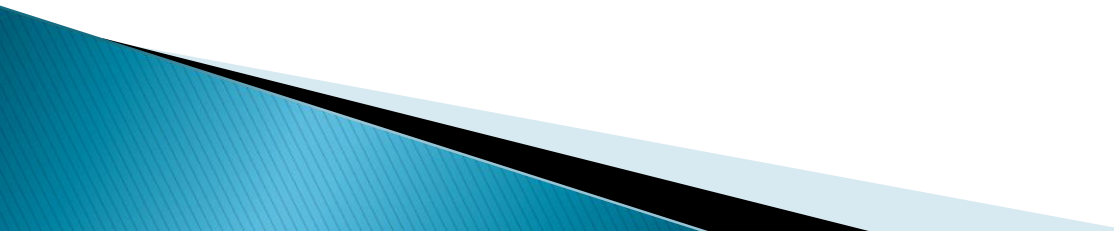
- ▶ **The perceptual outcome of coarticulation:**
  - ▶ The type of coarticulation we discuss is exemplified by the way we produce the /s/ in words such as *see* and *sue*.
  - ▶ These different articulatory gestures mean that the /s/ in these two words is produced in two different ways.
  - ▶ An experimental question is whether any information concerning the vowel in *sue* is contained in the consonant that preceded it.
  - ▶ Several researchers have investigated issues related to the perception of coarticulated segments.
- 

- ▶ See a description of task 1 on page 137, second paragraph.
  - ▶ **The results:**
  - ▶ Results indicate that the vowels /i/ and /u/ were identified reliably in the fricative segments, but not the vowel /a/. **Why? What is the role of articulatory compatibility in perception?** Page 137, paragraph 3.
  - ▶ This demonstrated that the fricative portion not only contained information about the consonant but also contained information about the identity of the following vowel.
- 

- ▶ **Lexical and Syntactic Factors in Word Perception:**
- ▶ In one study, words embedded in sentences were perceived more accurately than when the same words were excised from their sentences and presented in isolation. It is clear from the results that in a sentence context, semantics and syntax help the listener decode individual words in fluent speech. What is called **top-down processing** (the use of semantic and syntactic information) as well as phonological **bottom-up processing** (using only acoustic information to decode the speech signal) operate jointly in everyday perception of conversation.
- ▶ See a description of task 2 at the bottom of page 140 and top of page 141.

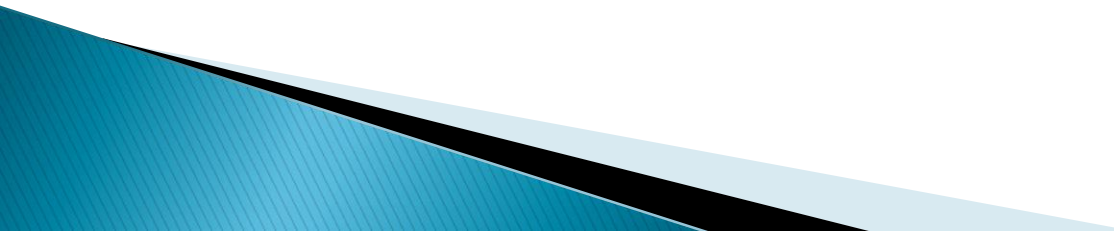
- ▶ **The results:**
- ▶ The subjects generated or *restored* a phoneme that was not part of the signal. The results were interpreted to suggest that when we listen to words, our expectations affect what we perceive. If we have most of the information necessary to specify a word, we mentally “smooth over” minor discrepancies in the speech signal.
- ▶ **The question is, then, under what circumstances do we detect irregularities or mispronunciations in words or sentences that we hear?**
- ▶ For a description of the study (**listening for mispronunciation (LM)**) that aimed at answering this question see page 141, paragraphs 2 and 3.

- ▶ **The Results:**
- ▶ Results suggested that voicing changes were detected most accurately for stops (*boot* to *poot*) (70%), followed by affricates (*chance* to *jance*) (64%), and least accurately for fricatives (*fin* to *vin*) (38%).
- ▶ The reduced ability to detect mispronunciations in fricatives may be due to their relatively weak acoustic signals. It is also true that few English words contrast minimally in the voicing characteristics of initial fricatives. Because this contrast is rare in English, listeners pay little attention to it.

- ▶ **Results of other experiments:**
  - ▶ In general, subjects were fairly accurate in detecting the mispronunciations based on place (80%–90%). Changes based on place differences (*take* to *pake*) were more perceptible than those based on voicing (*take* to *dake*). In addition, mispronunciations based on both place and voicing (*take* to *gake*) were detected no better than place changes alone.
- 

- ▶ Results of experiments comparing the perception of mispronunciations in word-initial and word-final consonants:
- ▶ The changes involved place in nasals (*made* to *nade*; *drum* to *drun*) and voicing in stops (*dish* to *tish*; *split* to *splid*). The results indicated that for all comparisons, more than twice as many correct detections were made for word-initial (72%) mispronunciations than for word-final (33%). The poor detection of word-final consonant mispronunciation may be partially explained on acoustic grounds for oral stops but not for the nasal stops. These results appear to indicate that listeners pay more attention to beginnings of words rather than ends of words. It is suggested that the listener accesses a word candidate soon after hearing the beginning of the word and “fills in” for the end of the word. This particular finding was used by Marslen-Wilson (1987) in developing his Cohort Theory.



- ▶ In conclusion, words are recognized through the interaction of sound and knowledge. Sounds in the beginning of a word are used to access word candidates, sounds in a word are recognized sequentially, and once recognized, words provide semantic and syntactic constraints used to recognize the rest of the message. This clearly demonstrates the joint influence of bottom-up and top-down processes operating when we listen to conversational speech.
- 


# Models of Speech Perception

- ▶ **Introduction:**
- ▶ The *motor theory of speech perception*, *analysis-by-synthesis*, and *fuzzy logical model* view the process of perception rising through stages from the auditory input, to a phonological level, and up to word identification. This view is called bottom-up and does not incorporate the effects of lexical and other “higher-level” cognitive knowledge into the process of speech perception.

- ▶ Models that incorporate the joint operation of multiple sources of information, including both bottom-up and top-down information, are called *interactive*. The main concern of interactive models is word recognition, whereas for bottom-up models, perception of phonetic segment is a major goal in itself. Most recent models incorporate an interactive approach. Two interactive models summarized briefly here are the *cohort model* and *TRACE model*.

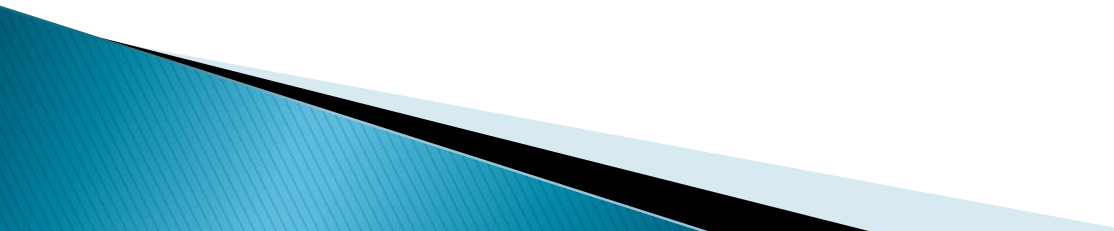
- ▶ **Motor Theory of Speech Perception:**
- ▶ This theory is advanced by Liberman and his colleagues (1967, 1970).
- ▶ The main thesis of the motor theory is that, at some point in the speech perception process, speech signals are interpreted by reference to motor speech movements. It directly links the processes of production with perception by stating that we perceive speech in terms of how we produce speech sounds. The early form of the theory hypothesized invariance at the motor articulatory level of speech production.
- ▶ Another assumption of this theory is that speech perception is phonetic and is different from auditory perception and is **species specific**(see page 144, paragraph 2).
- ▶ The evidence accumulated from research, however, has not provided strong support for the position that is necessary to engage in some form of articulatory knowledge during perceptual processing in speech.

## ▶ **Analysis-by-synthesis:**

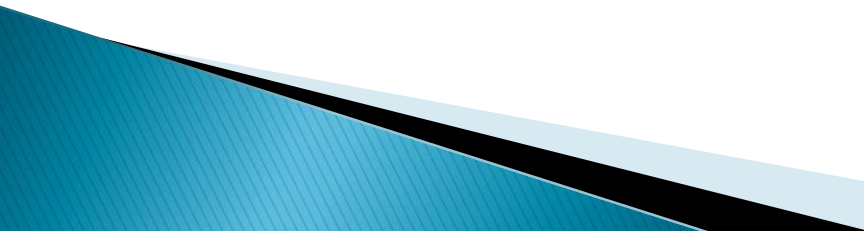
- ▶ This theory is Proposed by Stevens (1960) and Steven and Halle (1967).
  - ▶ The basic assumptions of this theory are similar to the motor theory in that speech perception and production are closely tied.
  - ▶ This model assumes that we make use of an abstract distinctive features matrix in a system of matching that is crucial to the speech perception process. The major claim of the theory is that listeners perceive (analyze) speech by implicitly generating (synthesizing) speech from what they have heard and then compare the “synthesized” speech with the auditory stimulus.(see bottom of page 144 and top of page 145).
  - ▶ However, it is an abstract model and little direct empirical evidence has been found to support it. It is vague in its evaluation of speech perception as special and different from auditory perception.
- 

## ▶ Fuzzy Logical Model:

- ▶ This theory is proposed by Massaro (1987, 1989) and Massaro & Oden (1980).
- ▶ According to this theory, speech perception is a prime example of pattern recognition. It assumes three operations in speech perception: *feature evaluation*, *feature integration*, and *decision*.
- ▶ It makes use of the idea of **prototypes**, which are summary descriptions of the perceptual units of language and contain a conjunction of various distinctive features. (see page 145, paragraphs 3 & 4).
- ▶ This model attempts to account for the difficulties of mapping acoustic attributes onto higher-level representations by viewing phonetic perception as a probabilistic process of matching features to prototype representations in memory.

- ▶ In the previous models the end results of phonetic segment identification are achieved without reference to meaning or syntax.
  - ▶ The following two models are concerned with auditory word recognition. For these models the end result is a meaningful utterance rather than a meaningless syllable, for example. These models aim to describe the interaction between the processes of phoneme recognition and word recognition.
- 

## ▶ Cohort Model:

- ▶ This theory is developed by Marslen–Wilson and his colleagues (1978, 1987).
  - ▶ It consists of two stages:
  - ▶ 1. In the first stage of word–recognition, the acoustic–phonetic information at the beginning of a target word activates all words in memory that resemble it making up the “cohort”, which is achieved on the basis of the acoustic information and is not influenced by other levels of analysis.
  - ▶ 2. In the second stage all the possible sources of information may influence the selection of the target word from the cohort. (see page 146, paragraph 3)
- 



- ▶ **TRACE Model:**
  - ▶ This model is developed by Elman and McClelland (1984, 1986).
  - ▶ This is a neural network model. It states that processing occurs through excitatory and inhibitory connections among numerous processing units called *nodes*. (see bottom of page 146 and first two paragraphs on page 147).
  - ▶ This theory is still actively undergoing development, refinement, and evaluation.
- 