Speech

Preview of next three lectures

1. Friday 25 March (MSEE 184) – Introduction to speech and short-time Fourier Analysis
2. Monday 28 March (EE 117) – Filtering of random signals and estimation of statistics for random signals (if time remains). Lecture will be given by Chun-Jung Tai
3. Wednesday 29 March (EE 117) – Continuation of short-time Fourier analysis and filter banks. Lecture will be given by Jan Allebach

Synopsis for topics to be covered on speech

1. Applications of digital processing of speech waveforms
2. Simple two component model for speech
3. Characteristics of speech waveforms
4. Short-time Fourier analysis
5. Spectrograms for speech
6. Other applications of short-time Fourier analysis: analysis of squeak noise in printers
7. Filter banks, signal synthesis, and wavelets
8. Signal approximation
9. Linear predictive coding

Comments

1. Posted legacy notes are all handwritten
2. Formulas for short-time Fourier analysis (only a few) are included with formulas document for Exam 3, which can be downloaded from the course website.

Applications for digital processing of speech waveforms

1. Analysis-only
    a. Speech recognition
        i. Siri (Apple iOS)
        ii. Customer support via phone
    b. Speaker verification
    c. Speaker identification
2. Synthesis-only
    a. Speech-based human-machine interaction
        i. Customer support via phone
    b. Text-to-speech
        i. Reading assistance for individuals with visual impairments
3. Analysis followed by synthesis
    a. Real-time speech communication systems
        i. Digital telephony
            1. VIOP (Voice over Internet Protocol)
            2. Mobile phone service
            3. Landline phone service
    b. Speech recording and play-back

Speech processing approaches

1. Waveform coding
    a. These methods do not exploit the specific characteristics of the speech waveform
    b. They are applicable to any type of audio waveform, such as music, although they may be designed to yield better results with speech than other types of waveforms, such as music
    c. Examples
        i. PCM – pulse code modulation
            1. Quantize signal
            2. Convert to binary
            3. Transmit 0s and 1s as string of pulses or other short-time waveforms
        ii. DPCM – differential pulse code modulation
            1. Same as above, but quantize difference between current signal sample value and previous signal sample value
        iii. ADPCM – adaptive differential pulse code modulation
            1. Same as above, but allow quantizer step-size to vary locally in time according to short-time signal characteristics
        iv. Filter banks and wavelets
            1. Separate signal into different frequency bands, and allocate number of bits to each frequency band according to its importance
        v. In ECE 438, we will not discuss Approaches i.-iii. above. We will discuss filter banks and maybe wavelets.

2. Speech-model-based
    a. Here the processing exploits specific characteristics of the speech waveform
    b. Our treatment of speech in ECE 438 will focus on approaches of this type

Mechanisms for Production of Speech

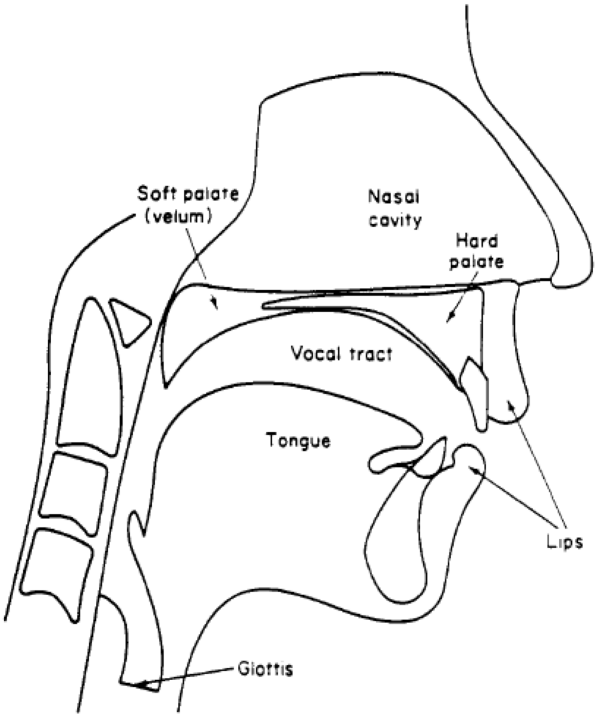


FIGURE 3-1. Cross-sectional view of the vocal mechanism, (after Markel and Gray).

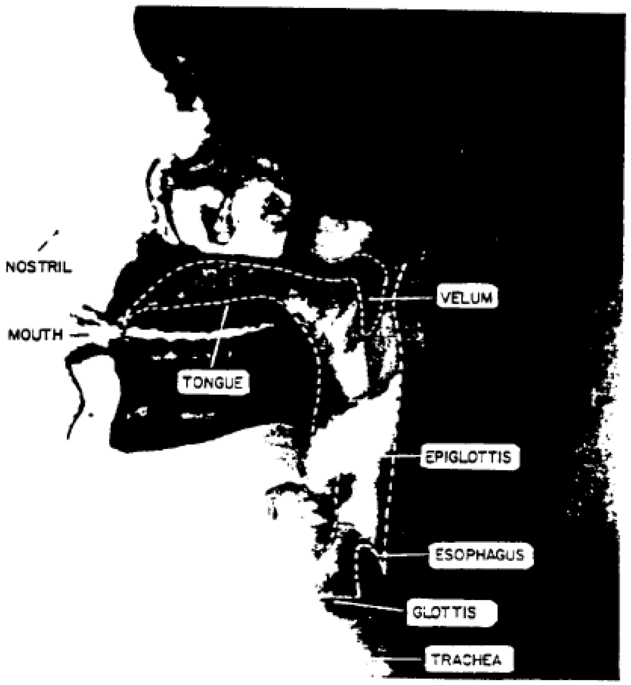


Fig. 3.1 Saggital plane X-ray of the human vocal apparatus. (After Flanagan et al. [6].)

How speech is generated

1. Air is forced out of the lungs by motion of *diaphragm*
2. The opening from the lungs to the throat is controlled by two flaps of tissue called the *glottis*.
3. The glottis operates in one of two different ways:
    a. It opens and closes periodically, thereby generating short pulses of air. This results in what is known as *voiced speech.* The interval between pulses is called the *pitch period*
    b. It opens, and stays open for an exteneded period of time, thereby generating a rush of air in a turbulent flow. This results in what is known as *unvoiced speech.*

4. The air, either pulsed for voiced speech, or a continuous (in time) turbulent flow for unvoiced speech passes up through throat and separates into two streams:
    a. One stream passes through the oral cavity, and exits through the mouth
    b. The other stream passes through the nasal cavity, and exits through the nose
5. The characteristics of the sound that results is governed by
    a. Voiced or unvoiced
    b. Position of the tongue, for example
        i. Forward and low
        ii. Back and high
    c. Position of the teeth (jaw)
        i. Open
        ii. Closed
    d. Position of the lips
        i. Open
        ii. Closed
        iii. Shape of opening
    e. The combination of b. through d. operates to form a *resonant cavity.*
        i. The resonant cavity exhibits certain *resonant frequencies* known as *formants*, much in the same manner that a pipe organ functions
        ii. The characteristics of the resonant cavity are independent of the air stream (voiced or unvoiced) that drives the cavity
6. Each distinct vocal sound is called a *phoneme*
7. The period of time during which a distinct vocal sound is made is called an *epoch*

Categorization of Phonemes

Table 3.1 Phonemes in American English.

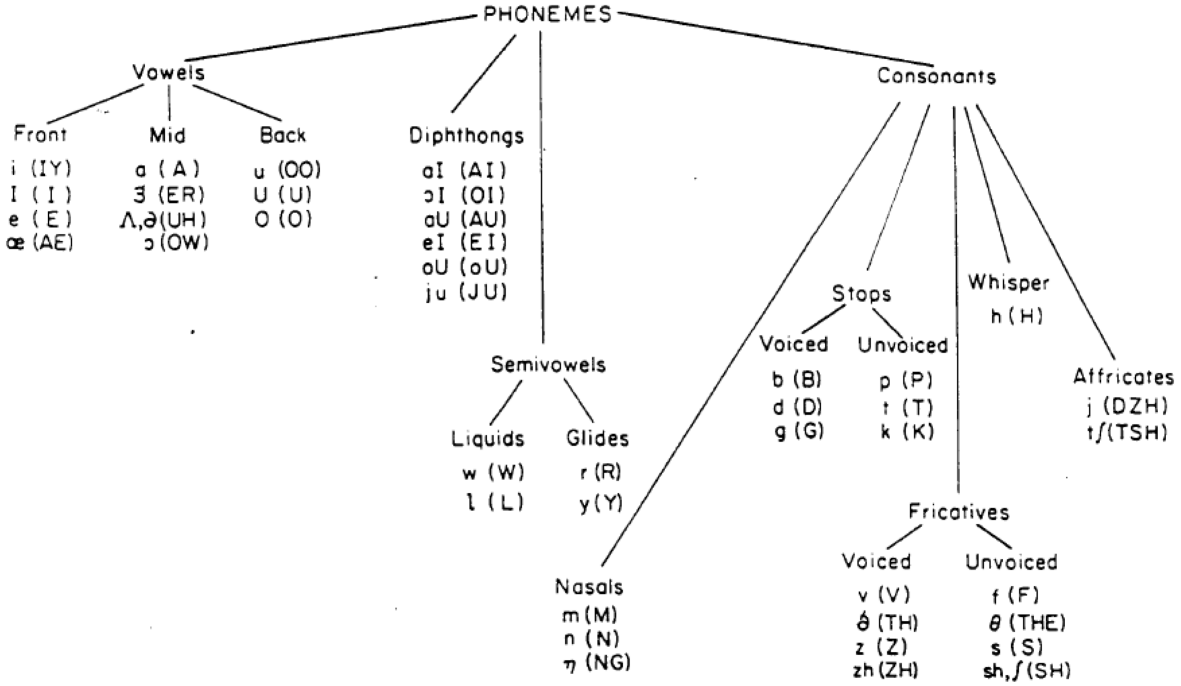
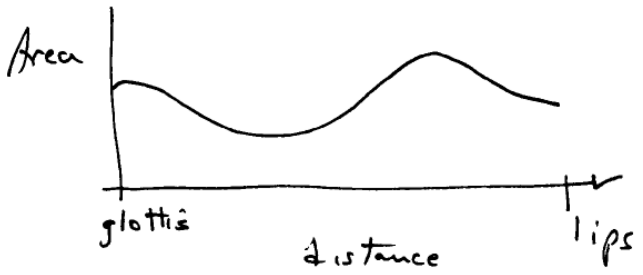


Table 3.2 Average Formant Frequencies for the Vowels. (After Peterson and Barney [11].)

FORMANT FREQUENCIES FOR THE VOWELS					
Typewritten Symbol for Vowel	IPA Symbol	Typical Word	F_1	F_2	F_3
IY	i	(beet)	270	2290	3010
I	ɪ	(bit)	390	1990	2550
E	ɛ	(bet)	530	1840	2480
AE	æ	(bat)	660	1720	2410
UH	ʌ	(but)	520	1190	2390
A	ɑ	(hot)	730	1090	2440
OW	ɔ	(bought)	570	840	2410
U	ʊ	(foot)	440	1020	2240
OO	u	(boot)	300	870	2240
ER	ɜ	(bird)	490	1350	1690

Discussion of Phonemes

1. Each phoneme is either *voiced* or *unvoiced*
    a. Voiced – vocal tract is driven by a periodic train of air pulses with period given by the pitch period
    b. Unvoiced – vocal tract is driven by a continuous turbulent flow of air
2. Each phoneme is either *continuant* or *non-continuant*
    a. Continuant – configuration of the vocal tract remains fixed during the epoch for the phoneme
    b. Non-continuant – configuration of the vocal tract changes during the epoch for the phoneme, i.e. the vocal tract is time-varying
3. The two most important classes of phonemes in the English language are *vowels* and *consonants.* In this course, we will not be concerned with the other classes.
4. The area function primarily determines the formants, and it is primarily governed by the position of the tongue



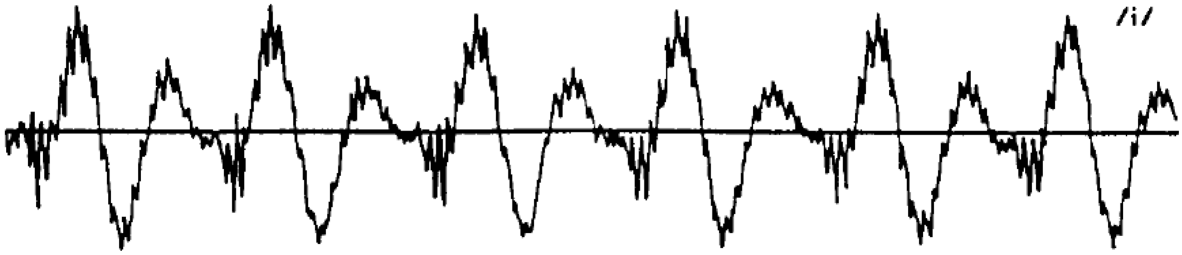
Example vocal tract configuration for phonemes

1. /a/ as in “father”
    a. open in front
    b. constricted in back
    c. tongue back

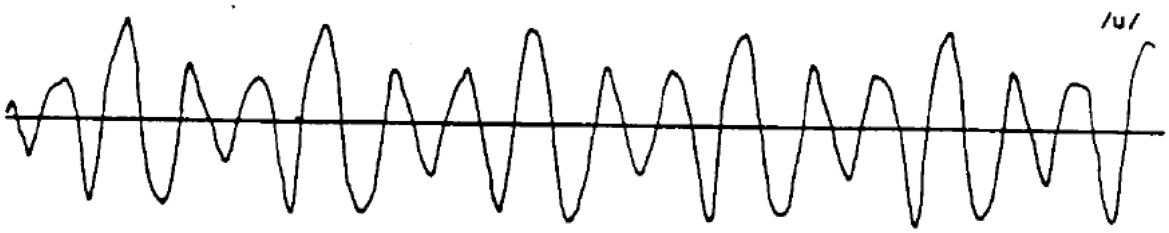
2. /i/ as in “beet”
    a. constricted in front
    b. open in back
    c. tongue forward and up

Example formants and time waveforms for phonemes

1. /i/ as in “beet”
    a. low first formant, high second formant
    b. time waveform shows slow oscillation with superimposed fast oscillation



2. /u/ as in “loom”
    a. low first and second formants
    b. time waveform is very smooth



Formants for the English language

Table 3.2 Average Formant Frequencies for the Vowels. (After Peterson and Barney [11].)

FORMANT FREQUENCIES FOR THE VOWELS					
Typewritten Symbol for Vowel	IPA Symbol	Typical Word	F_1	F_2	F_3
IY	i	(beet)	270	2290	3010
I	ɪ	(bit)	390	1990	2550
E	ɛ	(bet)	530	1840	2480
AE	æ	(bat)	660	1720	2410
UH	ʌ	(but)	520	1190	2390
A	ɑ	(hot)	730	1090	2440
OW	ɔ	(bought)	570	840	2410
U	ʊ	(foot)	440	1020	2240
OO	u	(boot)	300	870	2240
ER	ɜ	(bird)	490	1350	1690

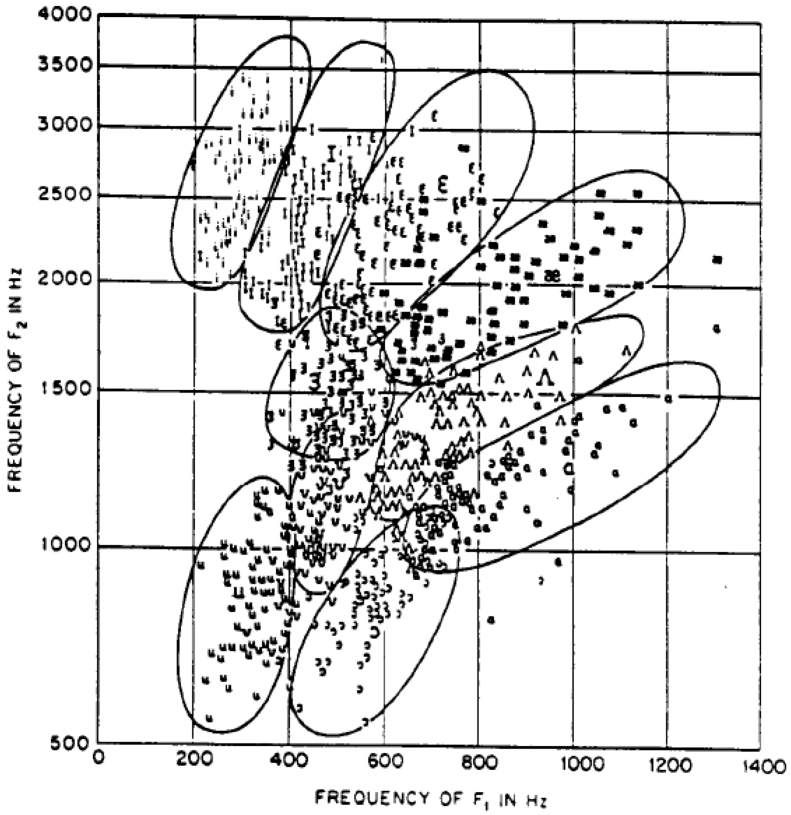


Fig. 3.4 Plot of second formant frequency versus first formant frequency for vowels by a wide range of speakers. (After Peterson and Barney [11].)

Note consistency of location of formants across different speakers

Example waveforms for phonemes – 1/2

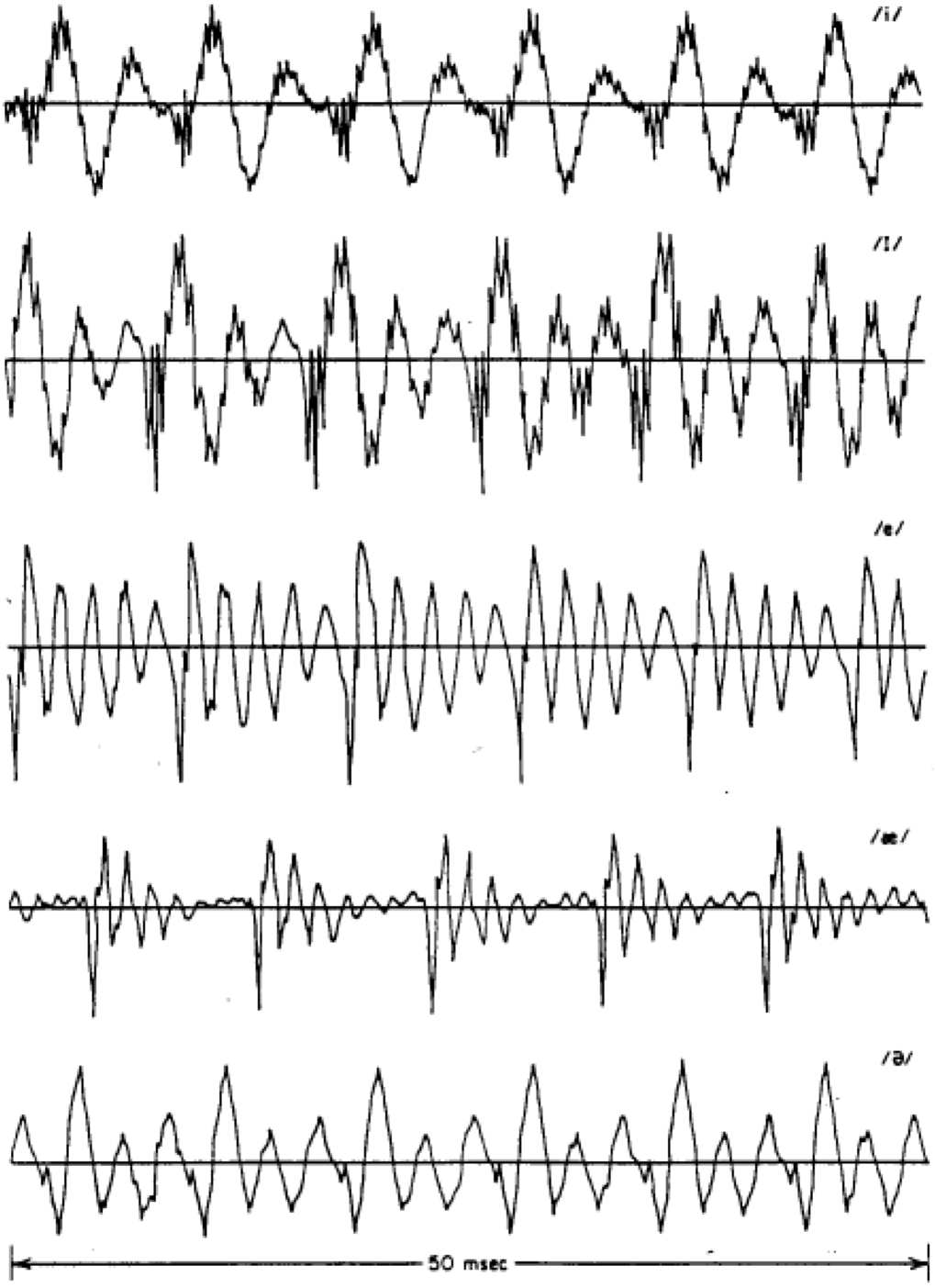


Fig. 3.6 The acoustic waveforms for several American English vowels and corresponding spectrograms.

Example waveforms for phonemes – 2/2

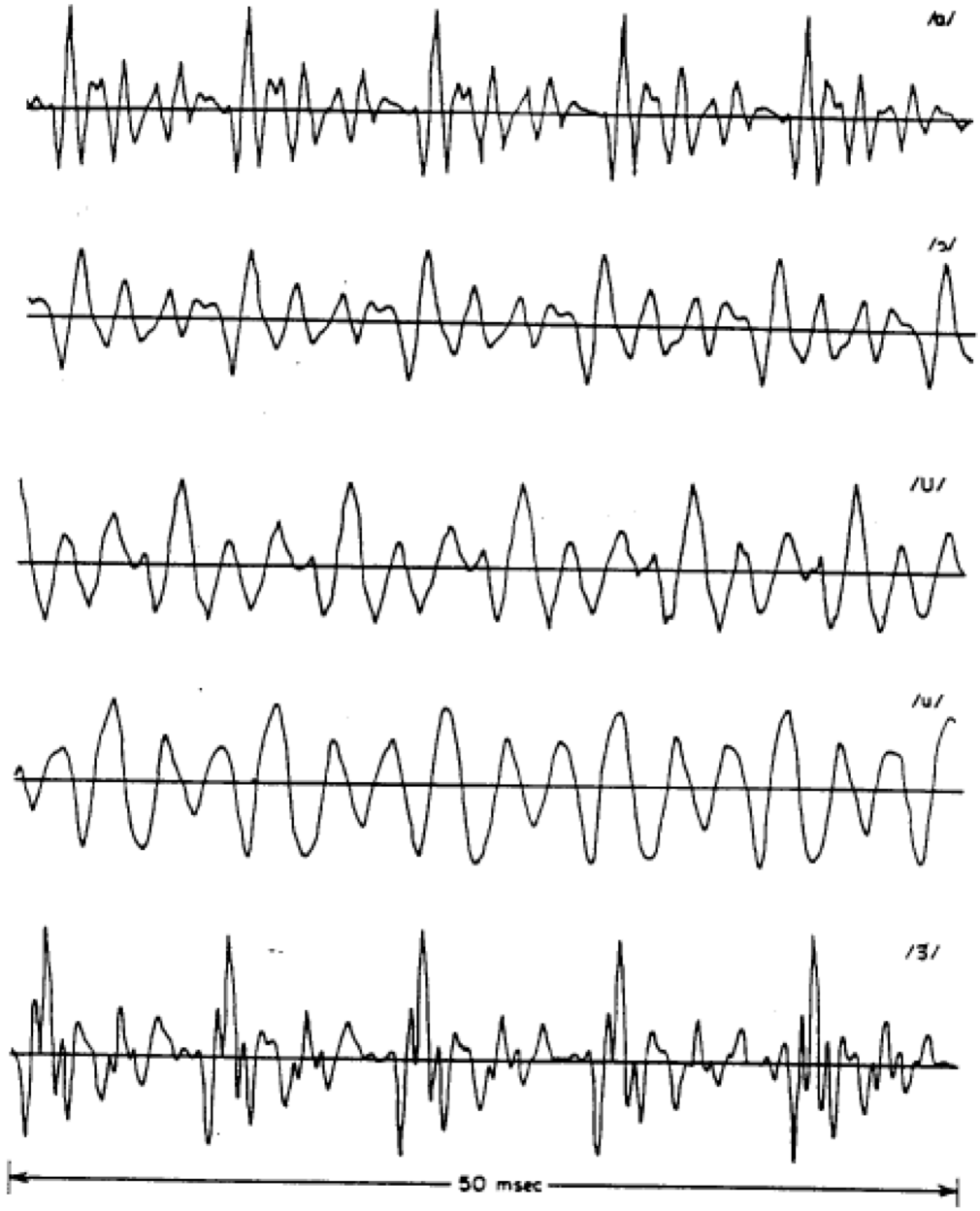


Fig. 3.6 (Continued)

Characteristics of speech waveforms

Now let's put it all together, and look at an actual speech waveform

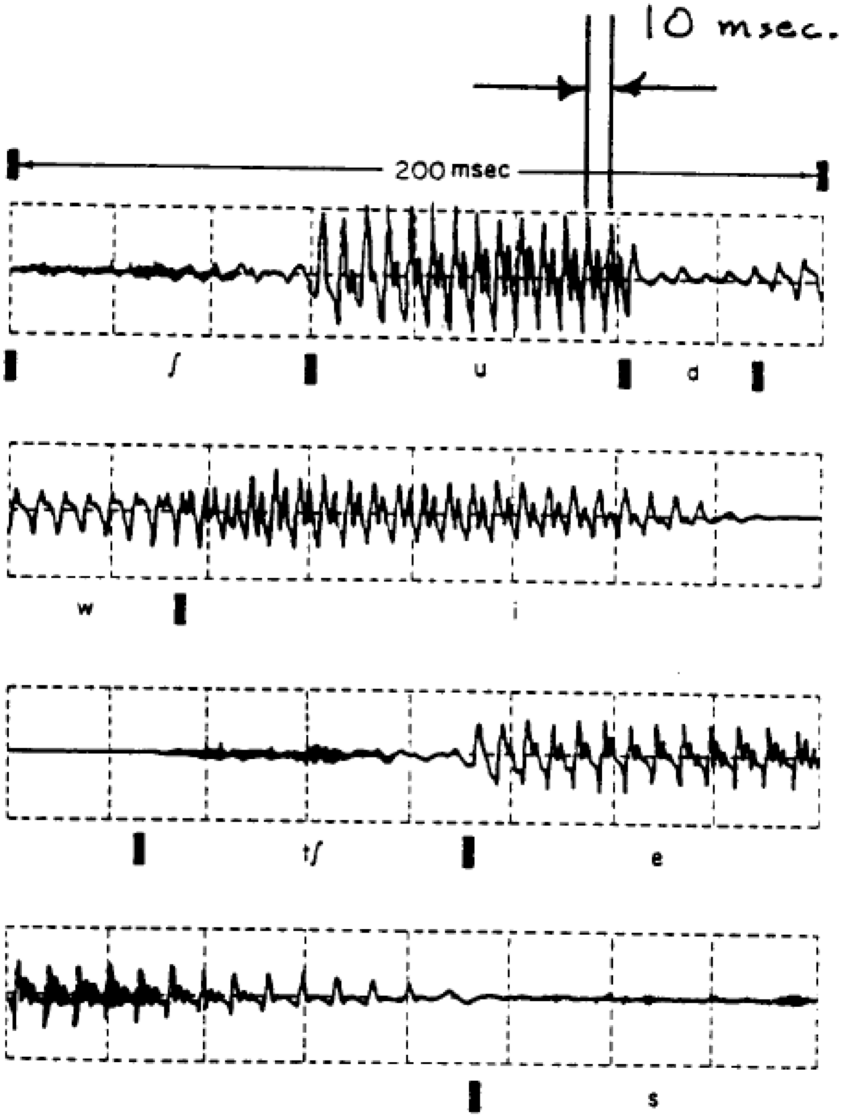
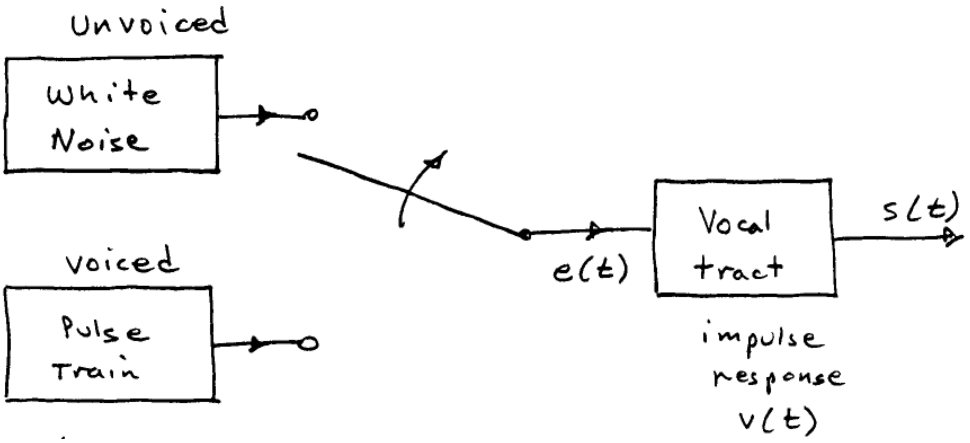


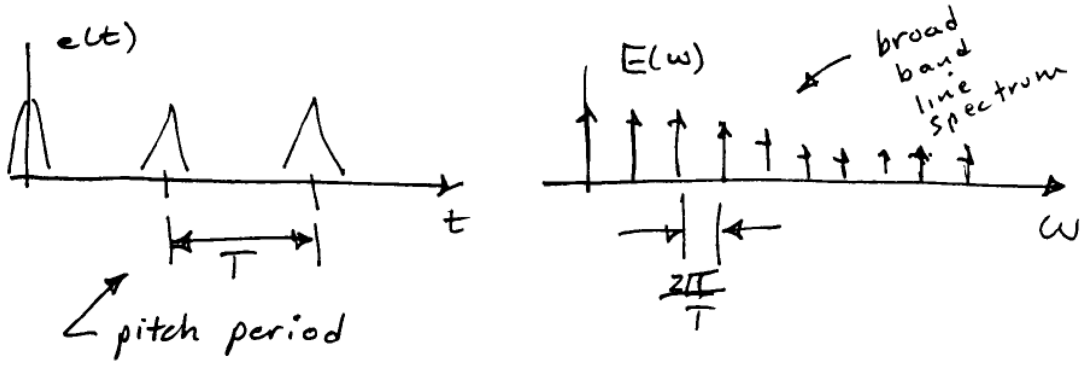
FIGURE 3-2. Example of a speech waveform illustrating different classes of sounds. The utterance is "should we chase . . .".

A simple model for speech generation



Excitation $e(t)$

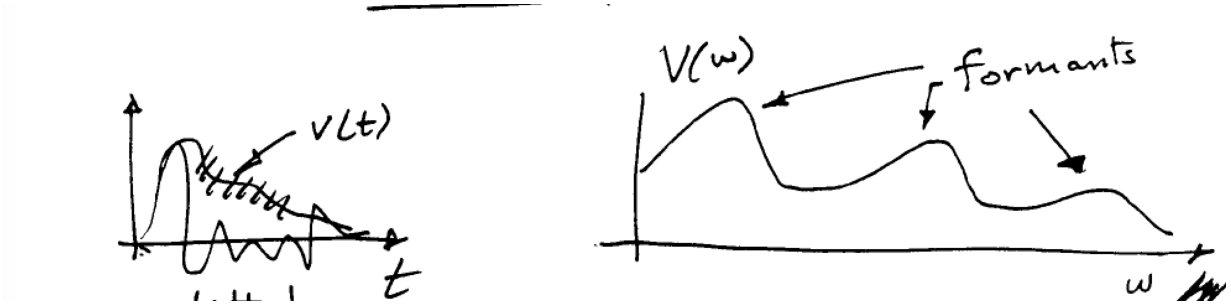
1. Voiced – vocal tract is driven by pulses of air at an interval given by the pitch period



2. Unvoiced – vocal tract is driven by continuous flow of turbulent air
    a. Model $e(t)$ as a white noise stochastic process
    b. Flat power spectrum, i.e. equal energy at all frequencies

Vocal tract

1. The configuration of the vocal tract changes with each different phoneme
    a. For *continuant* phonemes, it is constant throughout the epoch of the phoneme
    b. For *non-continuant* phonemes, it varies continuously throughout the epoch of the phoneme
2. During a short time interval, we model the vocal as a *linear time-invariant system* with vocal tract (impulse response) $v(t)$
    a. In fact, the vocal tract is time-varying.
    b. There exists theories to deal with *linear, time-varying linear systems*. However, we will not formally try to represent the time-varying nature of this system.
3. As discussed previously, the vocal tract will exhibit resonances that correspond to the formants



Speech waveform $s(t)$