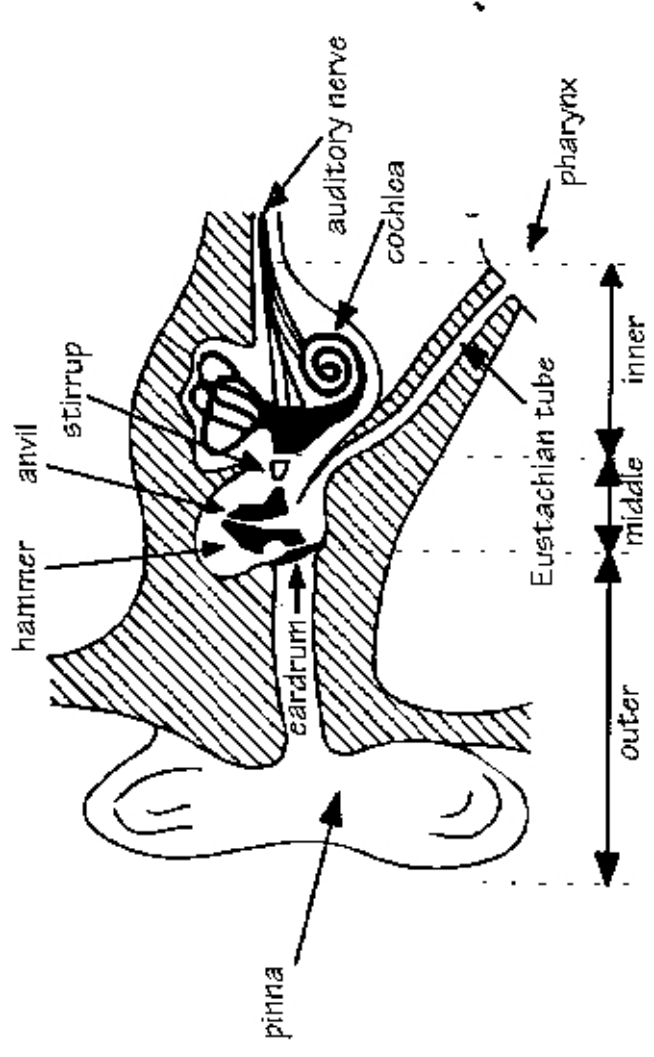


CECS401
Fundamentals of Spoken Language Processing

Note-3
Tuesday 8/31/99

D. Hearing sound
Illustration of auditory system



after Boite and Kunt, 1987.

In outer ear, sound wave is collected by pinna and vibrates eardrum.

In middle ear, the vibration is amplified and propagated by the ossicles (hammer-anvil-stirrup) to the inner ear.

In inner ear,

- cochlea contains a basilar membrane that is partially covered with hair-cells
- mechanical vibration of basilar membrane is transformed to neural firing of hair cells which are transmitted to the brain by auditory nerves
- sinusoidal sounds with different frequencies produce different amounts of displacement along basilar membrane, causing each hair-cell to be most sensitive to a specific frequency band

Therefore the inner ear approximately performs spectrum analysis on sounds.

E. Spectrogram of speech signals

Speech signal is stationary over 10 ms to 50 ms time windows.

Spectral analysis of speech needs to be performed over windowed analysis frames.

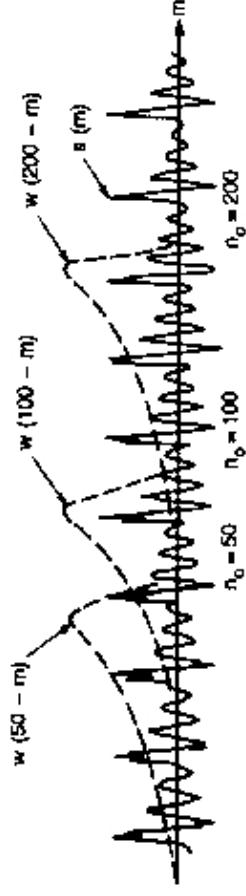


Figure 3.10 The signals $s(m)$ and $w(r - m)$ used in evaluation of the short-time Fourier transform.

Short-time Fourier transform (FT):

$$S_{n_0}(e^{j\omega_i}) = FT[s(m)w(n - n_0)]|_{\omega=\omega_i}$$

time: n_0

frequency: ω_i

energy distribution in time and frequency: $|S_{n_0}(e^{j\omega_i})|^2$

Spectrogram is a time-frequency representation of speech signal and is widely used in speech analysis.

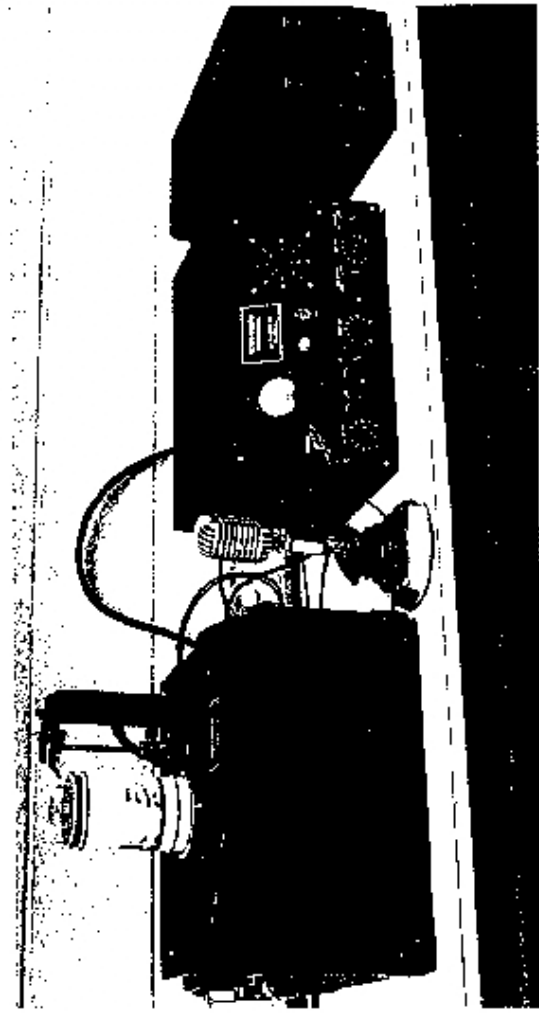
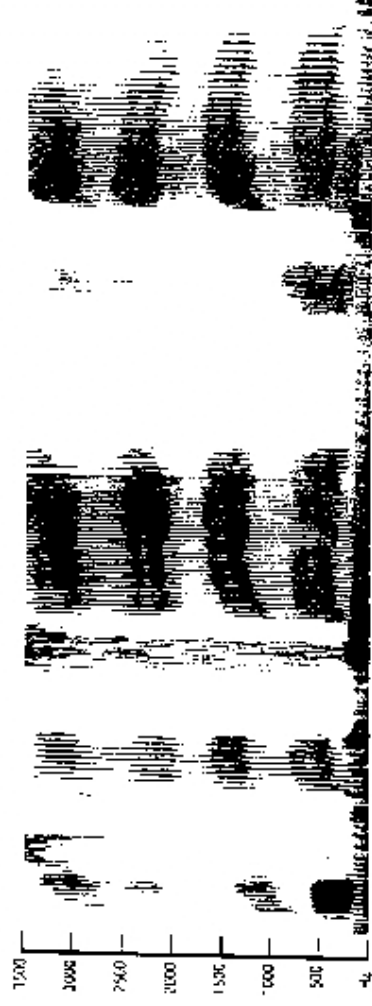


FIG. 10.—Sona-Graph Sound Spectrograph
Manufactured by Kay Elemetrics Corporation, ca. 1949

Time-frequency resolution tradeoff in spectrogram

Wideband spectrogram (15 ms — 20ms window)



- temporal resolution is good, frequency resolution is poor
- spectral envelopes are clear
- pitch harmonics are unresolved

The peaks of spectral envelopes correspond to the resonant frequencies of vocal tract and are called **formants**.

Narrowband spectrogram (40 ms — 50ms window)

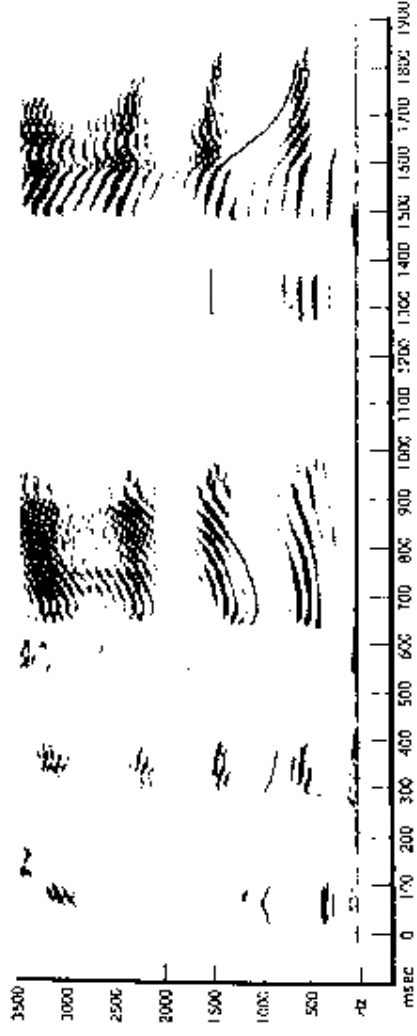


Figure 8.17 Wide-band (upper part of the figure) and narrow band (lower part) spectrogram of the question "Is Pat sad, or mad?" A line has been drawn through the tenth harmonic in the narrow-band spectrogram.

- temporal resolution is poor, frequency resolution is good
- spectral envelopes are unclear
- pitch harmonics are resolved

Formants are ranked in frequency from low to high as $F_1, F_2, F_3, F_4, \dots$ (F_0 is reserved for fundamental frequency).

The first three formants F_1, F_2, F_3 are important for discriminating vowels.

Formant trajectory changes with surrounding phoneme contexts.

Fundamental frequency can be estimated from narrowband spectrogram as:

$$F_0 = \frac{\text{nth harmonic frequency}}{n}$$

Vowels

Spectrogram of eight English vowels in the h-d context
(from Peter-Ladefoged, *A course in Phonetics*, 1993)

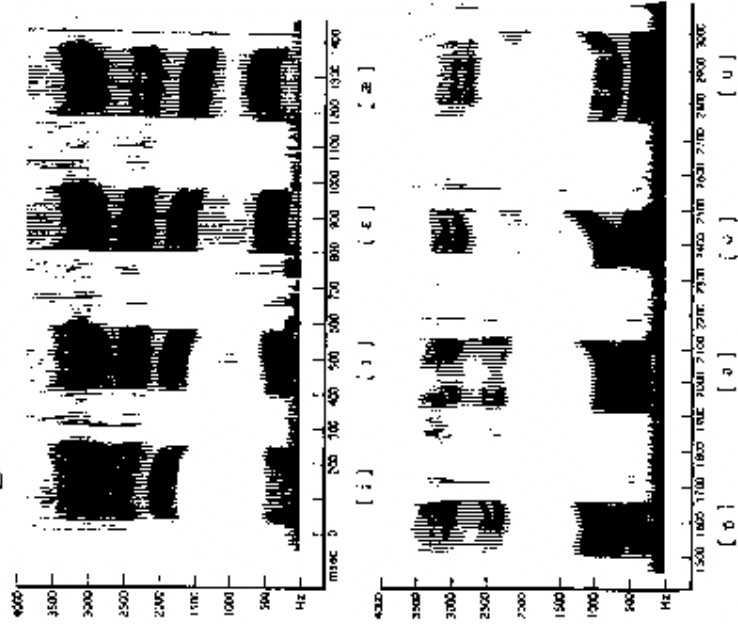


Figure 8.6 A spectrogram of the words "hood, hid, heed, had, hood, hood, hood" as spoken in a British accent.

Average formant frequencies for the eight vowels

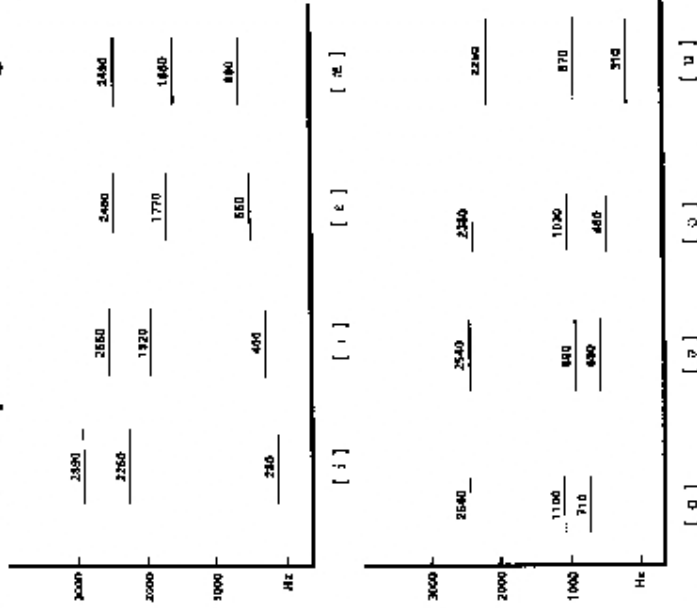


Figure 8.5 The frequencies of the first three formants in eight American English vowels.

- F_1 increases with tongue-hump level going from high to low
- F_2 decreases with tongue-hump position going from front to back

F_1 and F_2 distributions among talkers:
Data were collected from 76 speakers with 33 men, 28 women,
and 15 children, each reading twice ten words in h-d context.

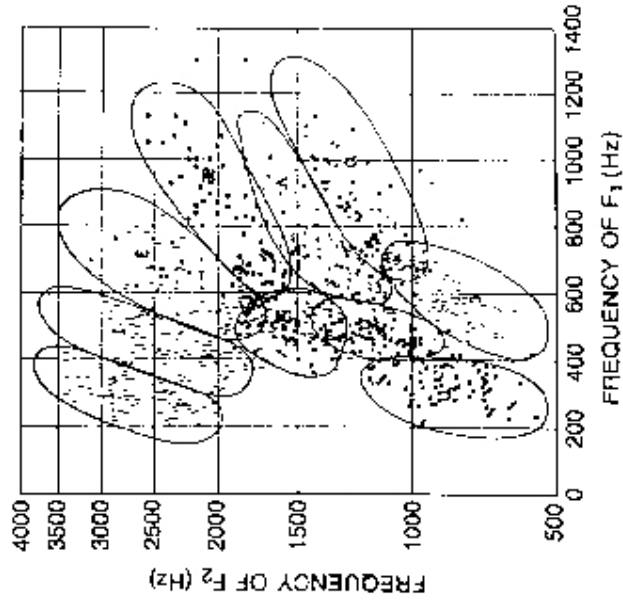
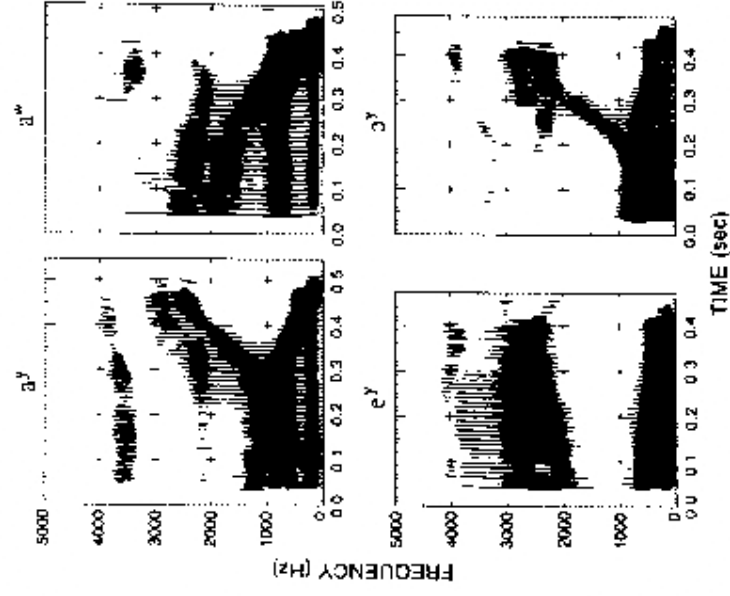


Figure 2.16 Measured frequencies of first and second formants for a wide range of talkers for several vowels (after Peterson & Barney [7]).

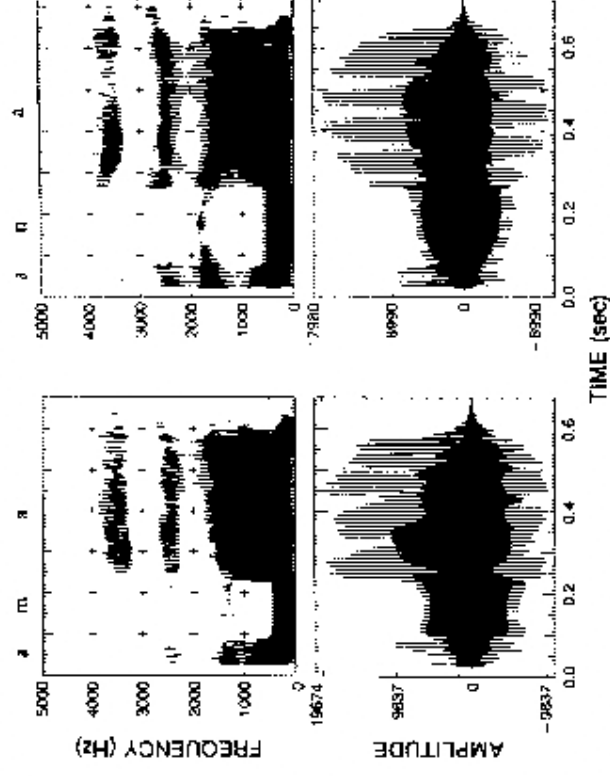
Diphthongs

Characterized by time-varying formant trajectories moving from the first target vowel to the second target vowel.



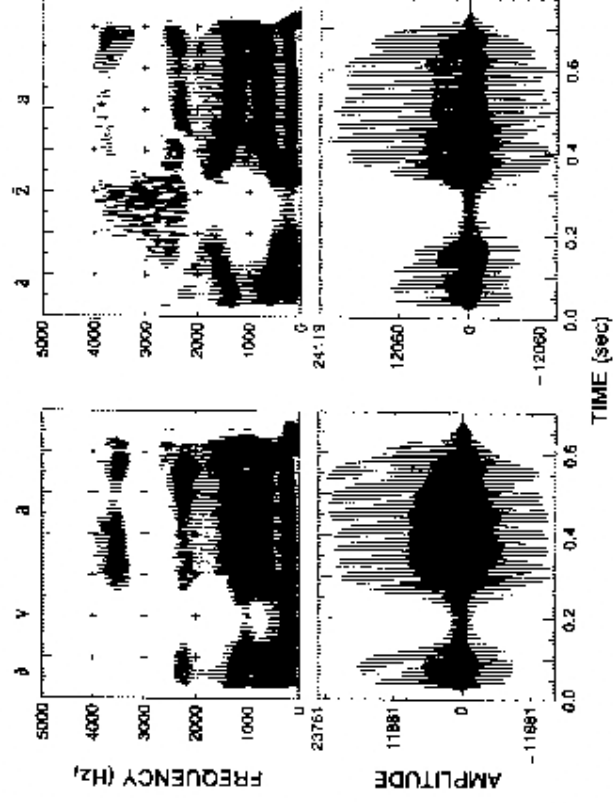
Nasal consonants

Characterized by energy concentration at low frequency band with broad bandwidth as well as spectral zeros due to oral cavity antiresonance.



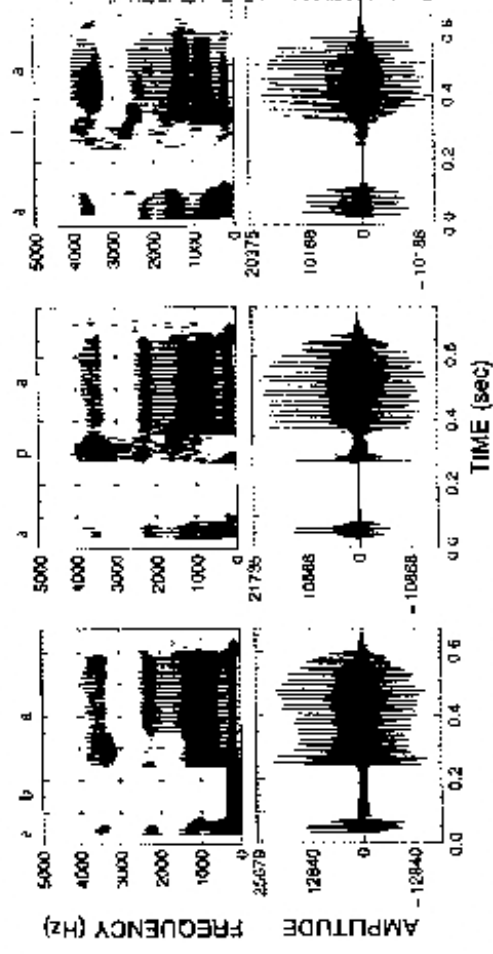
Voiced fricatives

Characterized by noise due to constriction of oral tract and vertical striation due to vocal cords vibration.



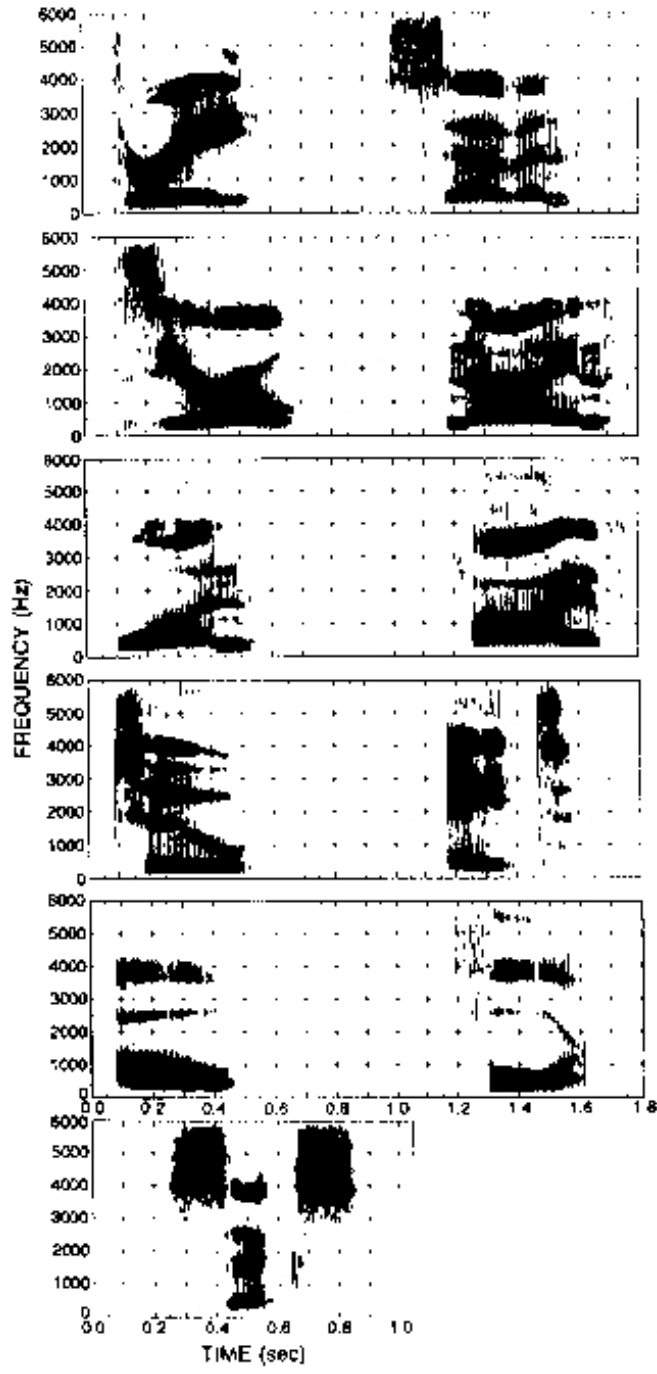
Stop consonants

Characterized by a gap (unvoiced stop) or voice bar (voiced stop) due to closure of oral tract (and vibration of vocal cords in voiced stop) followed by a release.



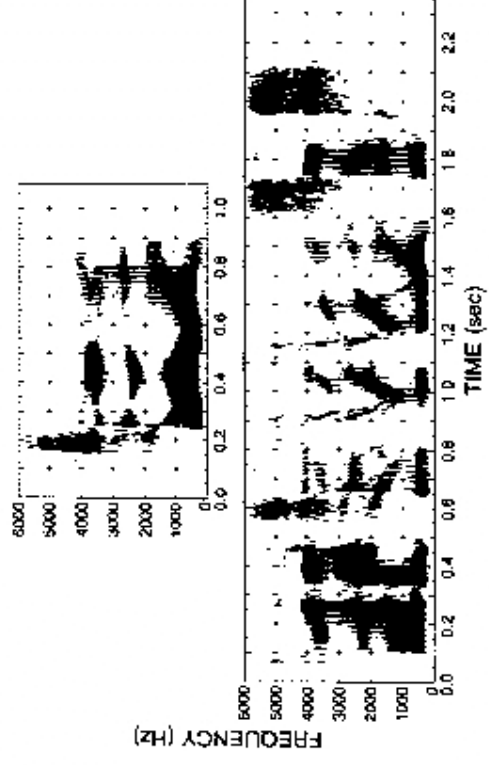
Example

Sound lexicon and spectrogram of 11 isolated digits: 0 through 9 plus oh (spectrogram in random sequence).



Example

Spectrogram of a 10-digit string



Example

Sound lexicon and spectrogram of 11 isolated digits: 0 through 9 plus oh (spectrogram in random sequence).

TABLE 2.3. Sound Lexicon of Digits

Word	Sounds	ARPABET
Zero	/z ɪ r o/	Z-IH-R-OW
One	/w ʌ n/	W-AH-N
Two	/t u/	T-UW
Three	/θ r i/	TH-R-IY
Four	/f o r/	F-OW-R
Five	/f aʲ v/	F-AY-V
Six	/s ɪ k s/	S-III-K-S
Seven	/s eː v ə n/	S-EH-V-AX N
Eight	/eʲ t/	EY-T
Nine	/n aʲ n/	N-AY-N
Oh	/o/	OW