CS578- Speech Signal Processing

Lecture 2: Production and Classification of Speech Sounds

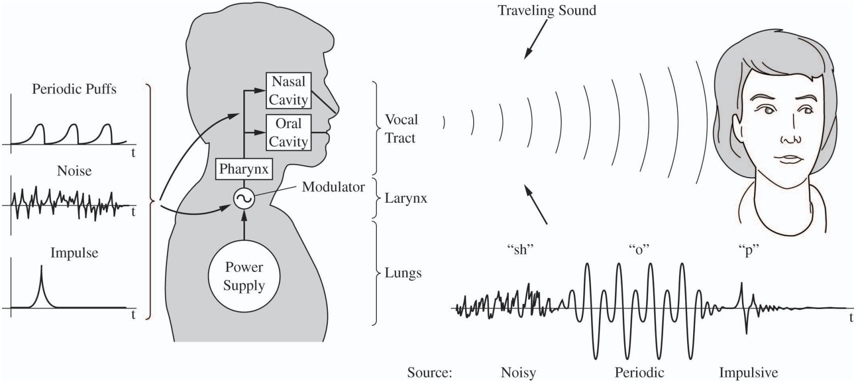
Yannis Stylianou

University of Crete, Computer Science Dept., Multimedia Informatics Lab
yannis@csd.uoc.gr

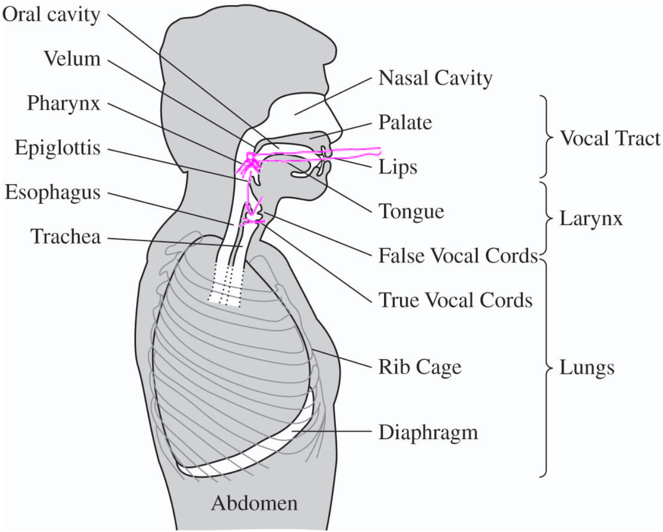
Univ. of Crete

Outline

A Simple view

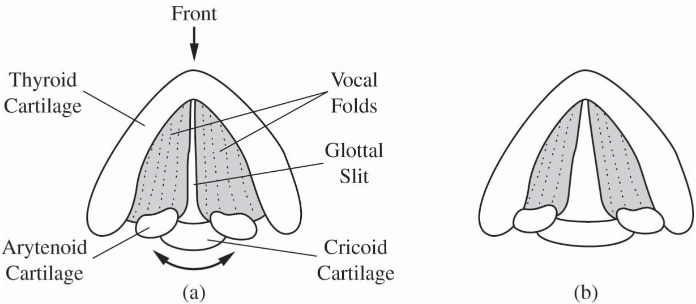


Cross sectional view

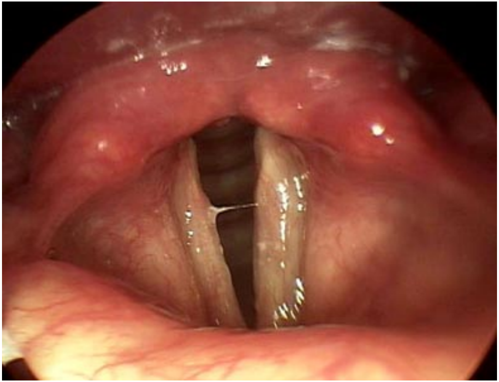


Downward-looking into the larynx: Vocal folds

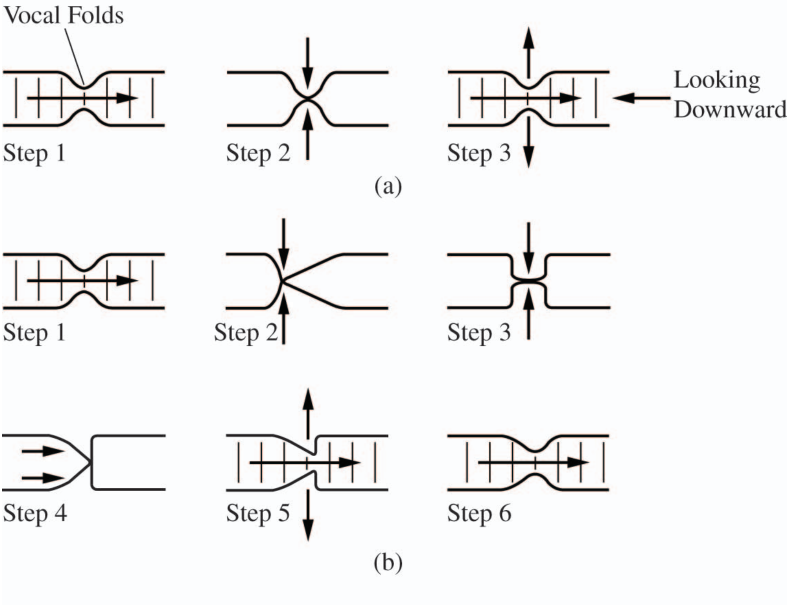
Left: Voicing, **Right:** Breathing



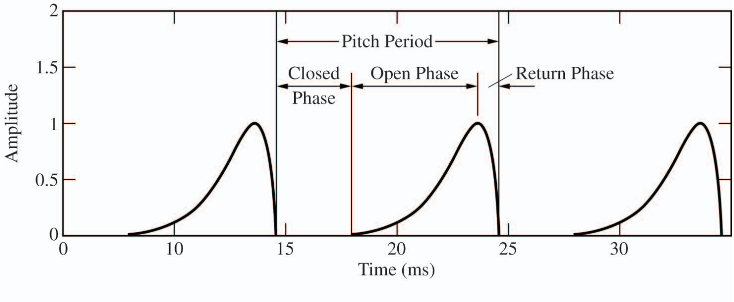
Vocal Folds vibration



Bernoulli's principle in the glottis

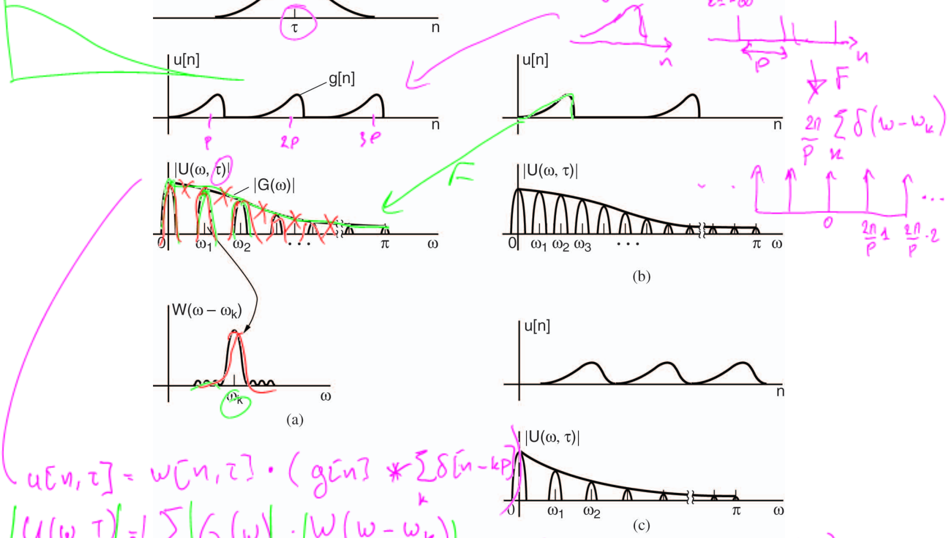


Glottal airflow velocity



Softer, typical, and relaxed glottal flow

$$u[n] = g[n] * \sum_{k=-\infty}^{\infty} \delta[n-kP]$$



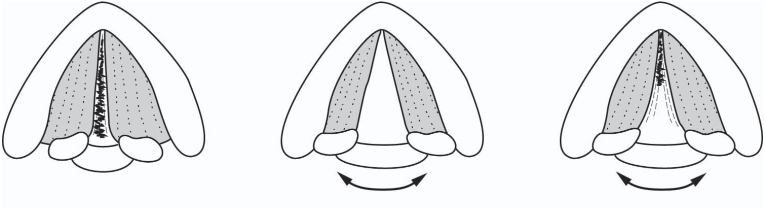
$$u[n,\tau] = w[n,\tau]\cdot\left(g[n] * \sum_k \delta[n-kP]\right)$$

$$|U(\omega,\tau)| = \frac{1}{P}\sum_k |G(\omega)|\cdot|W(\omega-\omega_k)|$$

$$\omega_k = \frac{2\pi}{P}\cdot k \quad (k = 0, 1, 2, \ldots P-1)$$

Other vocal folds configurations

Left: Whispering, **Middle:** Voicing **Right:** Whispering voicing


(a) (b) (c)

Other forms of vibration

- Creaky voice:

  vocal folds very tense
  only a portion of them in oscillation
  harsh-sounding voice
  high and irregular pitch

- Vocal fry

  folds are massy and relaxed
  abnormally low and irregular pitch
  secondary pulses during open phase

- Diplophonia

  extra flaps
  secondary pulses during the closed phase

Other forms of vibration

- Creaky voice:

  vocal folds very tense
  only a portion of them in oscillation
  harsh-sounding voice
  high and irregular pitch

- Vocal fry

  folds are massy and relaxed
  abnormally low and irregular pitch
  secondary pulses during open phase

- Diplophonia

  extra flaps
  secondary pulses during the closed phase

Other forms of vibration

- Creaky voice:

  vocal folds very tense
  only a portion of them in oscillation
  harsh-sounding voice
  high and irregular pitch

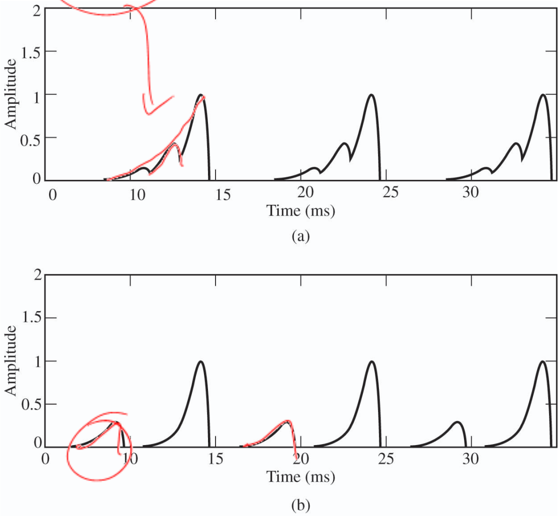
- Vocal fry

  folds are massy and relaxed
  abnormally low and irregular pitch
  secondary pulses during open phase

- Diplophonia

  extra flaps
  secondary pulses during the closed phase

Upper panel: vocal fry, **Lower panel:** diplophonia

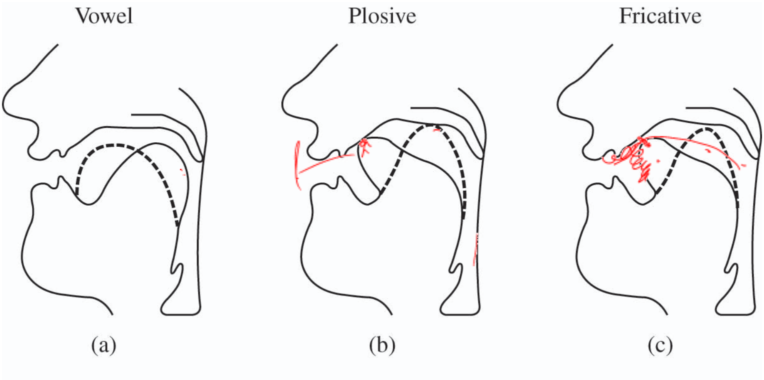


Vocal Tract

By saying Vocal Tract we mean:

- Oral cavity: from the larynx to the lips, and the Nasal cavity
- Oral tract: 17cm for male voice, shorter for females
- Its purpose is to spectrally "color" the source and generate new sources for sound production

Vocat tract shapes



Spectral Shaping

Vocal tract is often approximated by a linear filter with:

- Formant frequencies
- Formant amplitude
- Formant bandwidth

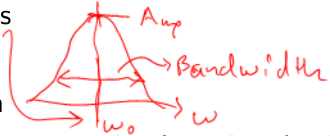
Assuming a stable vocal tract and only with poles filter:

$$
\begin{aligned}
H(z) &= \frac{A}{\prod_{k=1}^{N_i}(1 - c_k z^{-1})(1 - c_k^* z^{-1})} \\
&= \sum_{k=1}^{N_i} \frac{A_k}{(1 - c_k z^{-1})(1 - c_k^* z^{-1})}
\end{aligned}
$$

Spectral Shaping

Vocal tract is often approximated by a linear filter with:

- Formant frequencies
- Formant amplitude
- Formant bandwidth

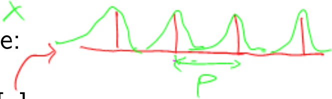


Assuming a stable vocal tract and only with poles filter:

$$
\begin{aligned}
H(z) &= \frac{A}{\prod_{k=1}^{N_i}(1 - c_k z^{-1})(1 - c_k^* z^{-1})} \qquad \text{παραγοντα} \\
&= \sum_{k=1}^{N_i} \frac{A_k}{(1 - c_k z^{-1})(1 - c_k^* z^{-1})} \qquad \text{σειρά}
\end{aligned}
$$

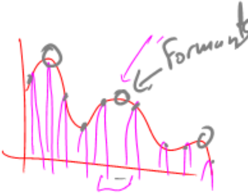
Example

Let the excitation of vocal tract, $h[n]$, be:



$$u[n] = g[n] \star p[n]$$

then, the output speech, $x[n,\tau]$, is given by:

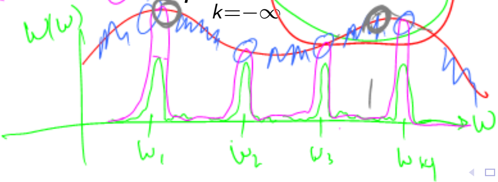


$$x[n,\tau] = w[n,\tau]\{h[n] \star (g[n] \star p[n])\}$$

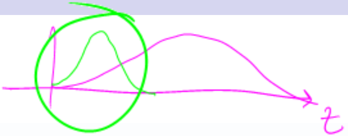
and

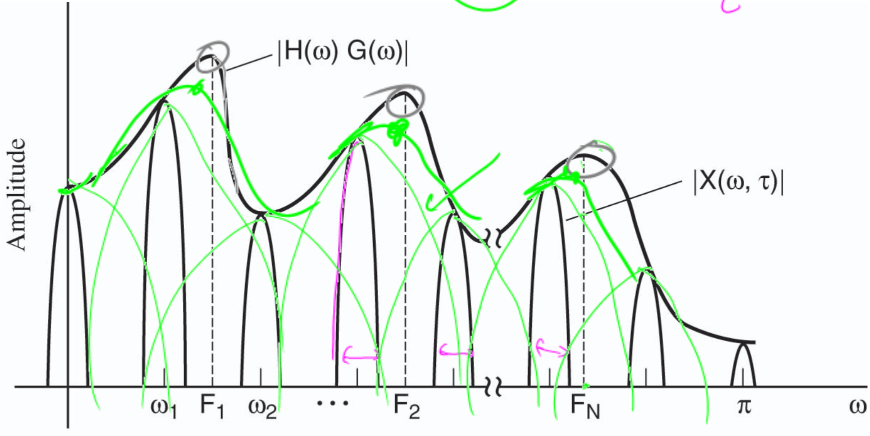
$$X(\omega,\tau) = \frac{1}{P}\sum_{k=-\infty}^{\infty} H(\omega_k)G(\omega_k)W(\omega-\omega_k,\tau)$$

$\omega_k = \frac{2\pi}{P}\cdot k$



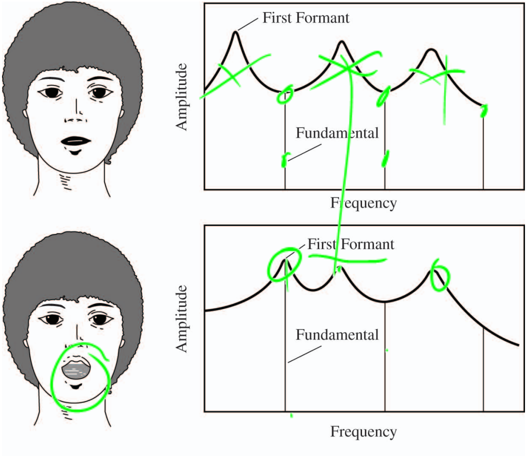
Harmonics and Formants





$W_2 \approx 2W_1$

Sopranos



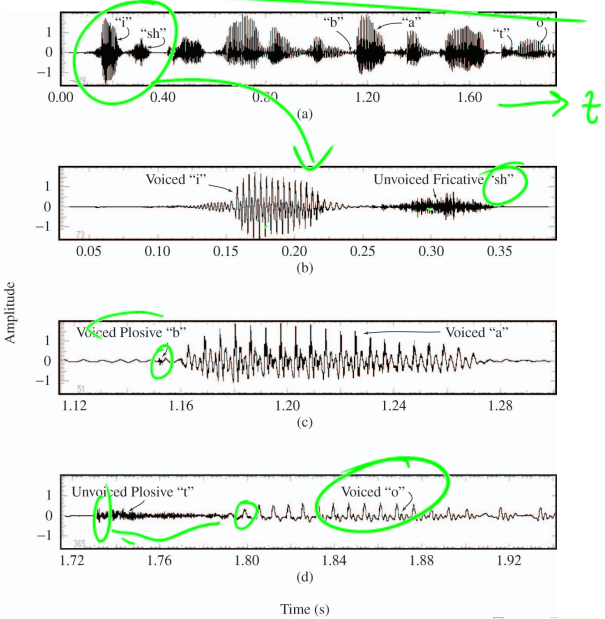
First Formant
Amplitude
Fundamental
Frequency
First Formant
Amplitude
Fundamental
Frequency

Ways to categorize speech sounds

- Vocal fold state:
  - Voiced
  - Unvoiced
- Oral tract state:
  - Plosives
  - Fricatives

Also: voiced and unvoiced plosives (/b/,/t/), voiced and unvoiced fricatives (/z/,/f/), whispered unvoiced

"Which tea party did baker go to?"



Outline

?

Short Time Fourier Transform, STFT

STFT:

$$X(\omega, \tau) = \sum_{n=-\infty}^{\infty} x[n, \tau] e^{-j\omega n}$$

where

$$x[n, \tau] = w[n, \tau] x[n]$$

Spectrogram:

$$S(\omega, \tau) = |X(\omega, \tau)|^2$$

Short Time Fourier Transform, STFT

STFT:

$$X(\omega, \tau) = \sum_{n=-\infty}^{\infty} x[n, \tau] e^{-j\omega n}$$

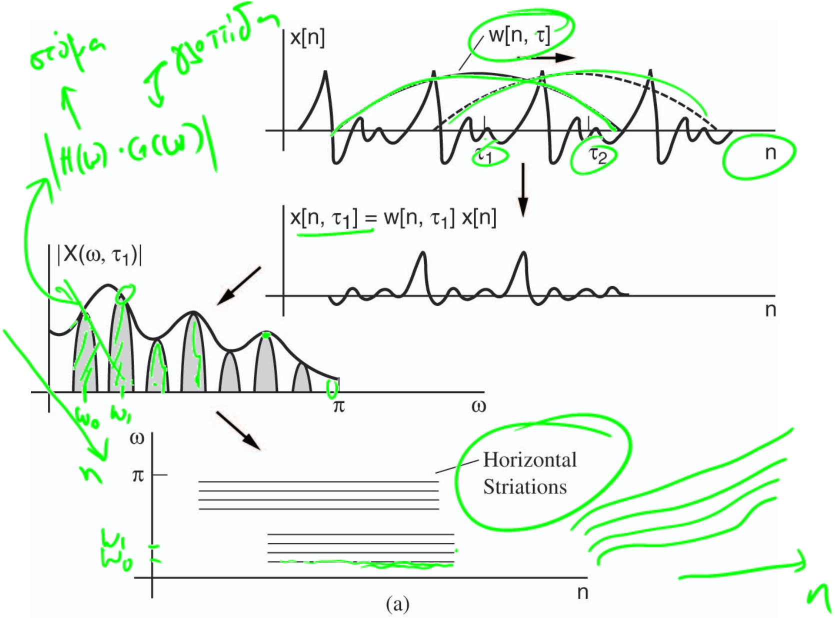
where

$$x[n, \tau] = w[n, \tau] x[n]$$

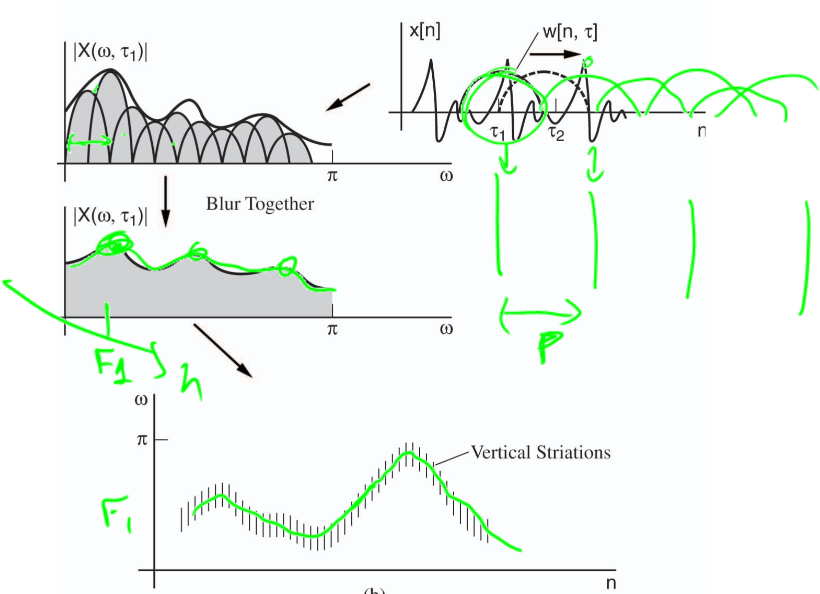
Spectrogram:

$$S(\omega, \tau) = |X(\omega, \tau)|^2$$

Narrowband Spectrogram

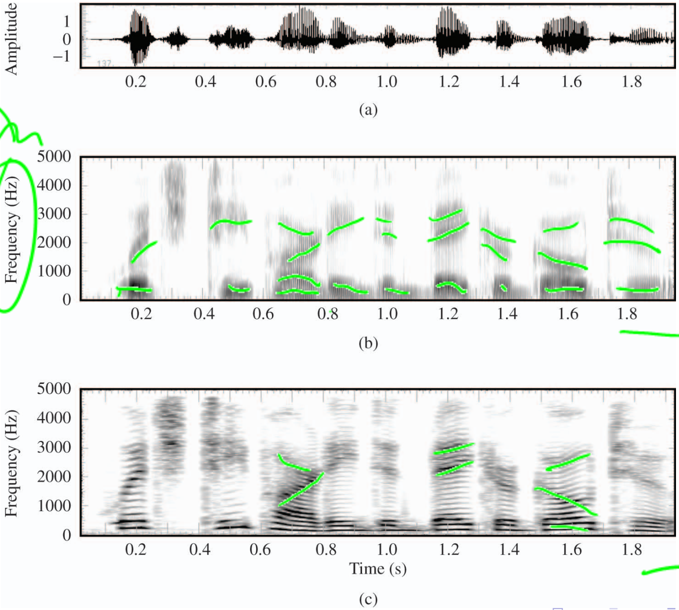


Wideband Spectrogram



(b)

Spectrogram on speech

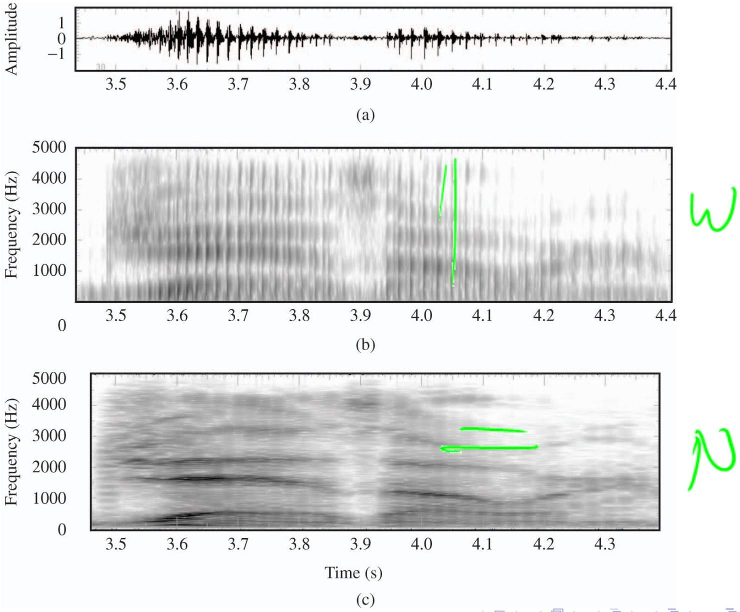


? N
W

? N
W

n

Spectrogram on speech; another example



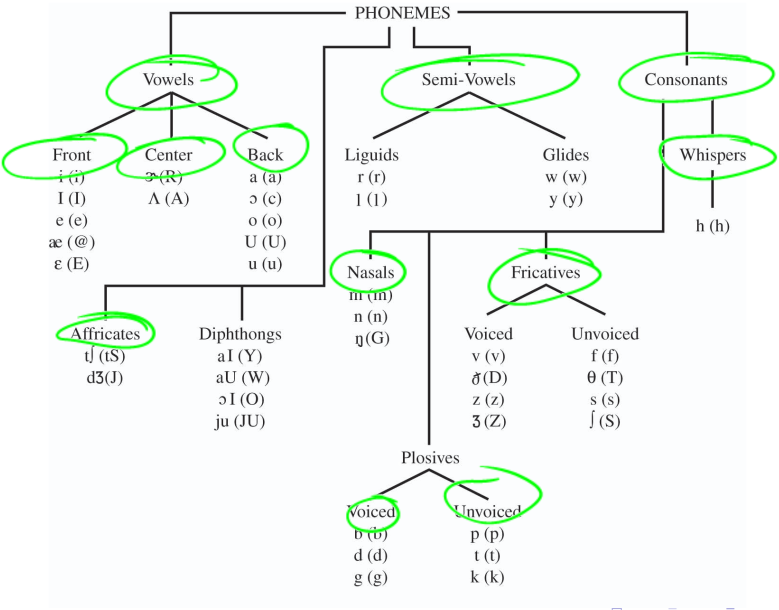
Do we know better now?

to classify sounds by looking in time of in frequency domain for

- periodic, noisy, impulsive sources?
- shape of vocal tract?

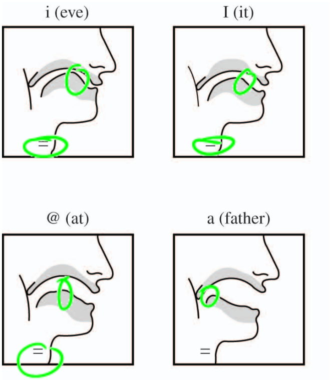
Outline

Phonemes' map

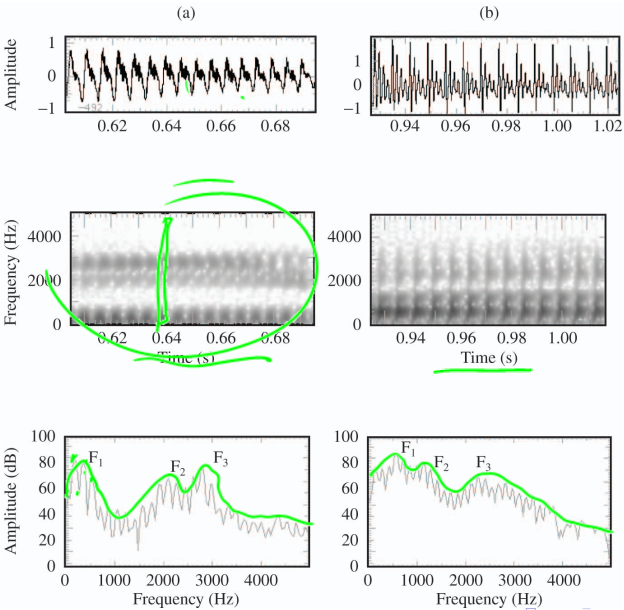


Vowels

- **Source:** Quasi-periodic puffs of airflow
- **System:** Each vowel phoneme corresponds to a different vocal tract configuration.

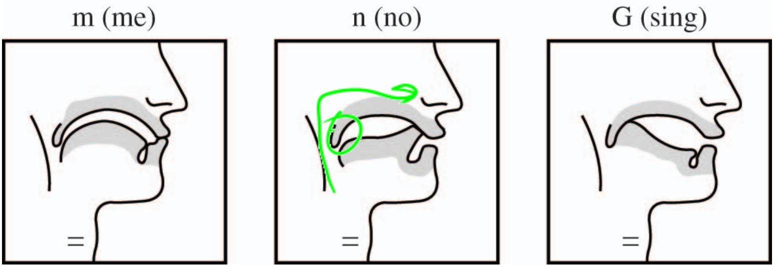


Vowels: time and spectrogram

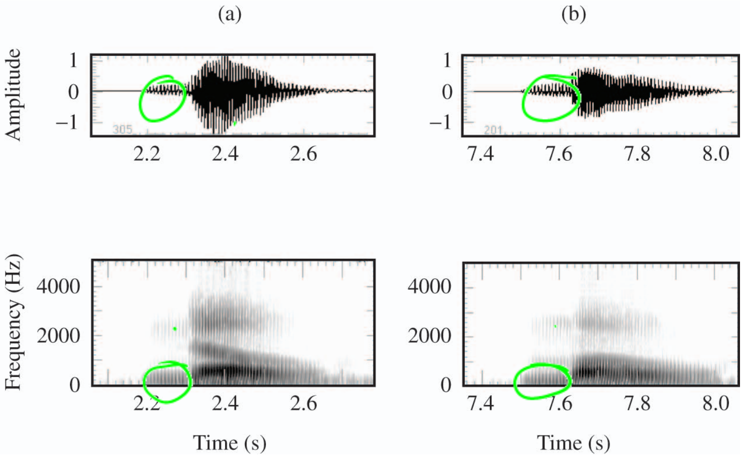


Nasals

- **Source:** Quasi-periodic puffs of airflow
- **System:** Air flows mainly through the nasal cavity and oral tract being constricted



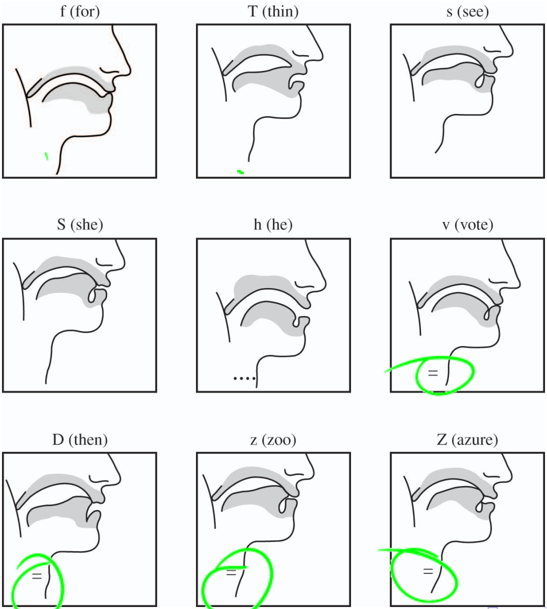
Nasals: time and spectrogram



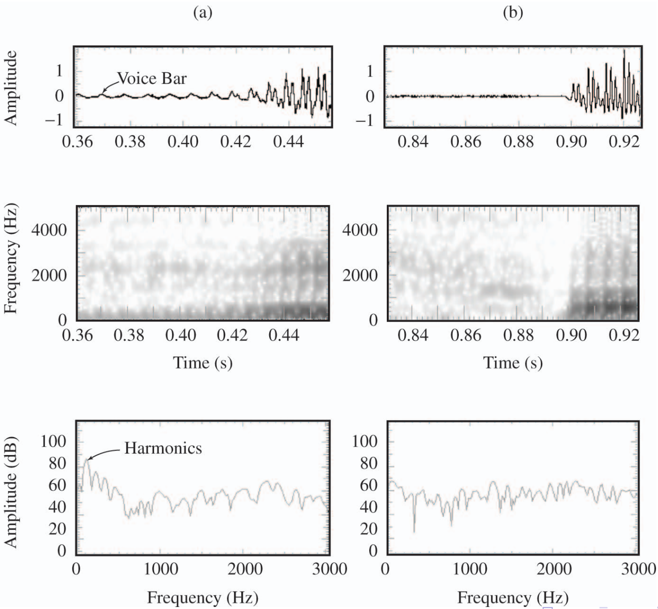
Fricatives

- **Source:**
  - *Voiced:* vocal-folds vibrate
  - *Unvoiced:* vocal-folds are relaxed and not vibrating
- **System:** Oral tract being constricted by tongue at the back, center, or front of the oral tract, or at the teeth or lips

FRICATIVES' PROFILE



Fricatives: time and spectrogram



Plosives, or "burst" signals

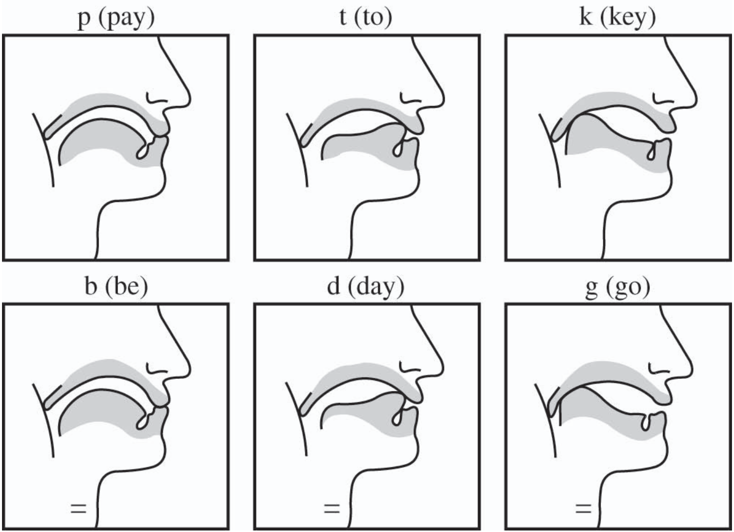
Voiced:

- **Source:** vocal folds are vibrating ("voice bar")
- **System:** Oral tract being constricted by tongue at the back, center, or front of the oral tract, or at the teeth or lips

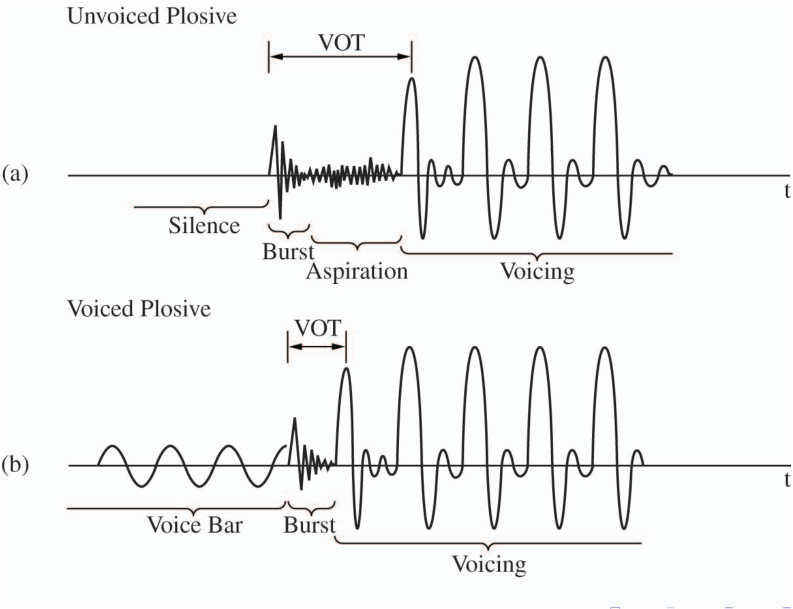
Unvoiced:

- **Source:** vocal folds are not vibrating
- **System:** Oral tract being constricted by tongue at the back, center, or front of the oral tract, or at the teeth or lips

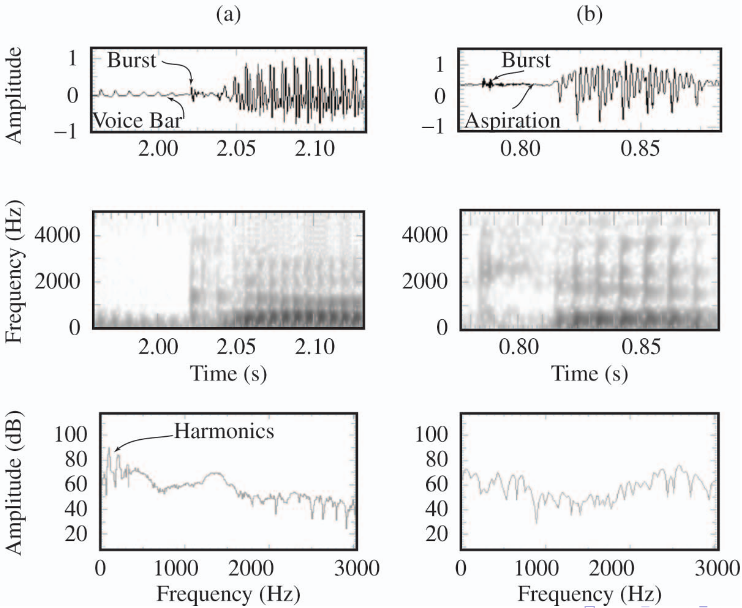
Plosives' profile



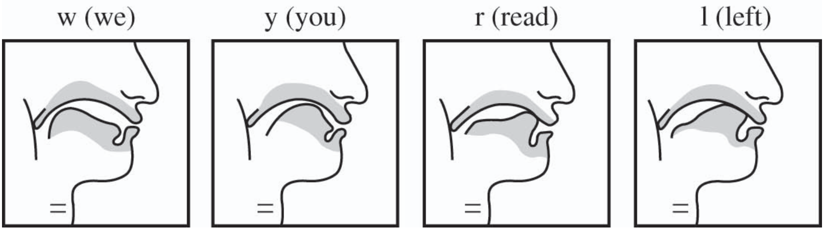
Voice Onset Time



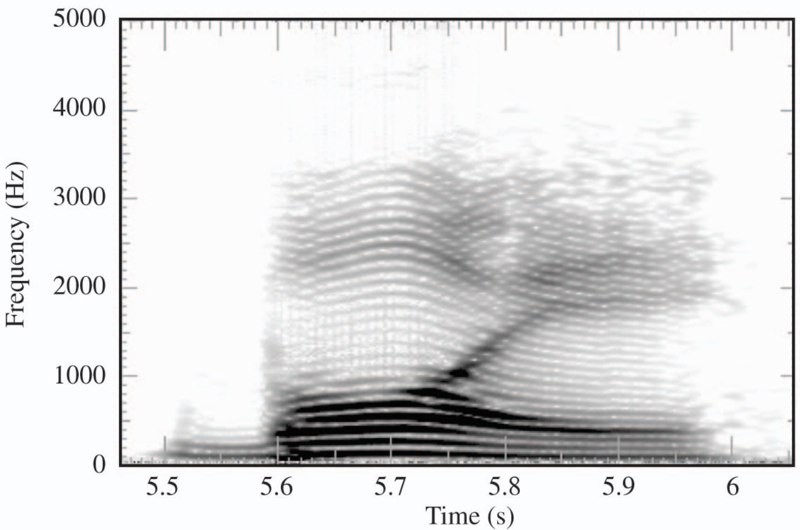
Plosives: time and spectrogram



Semi-vowels



Transitional Speech Sounds: “boy”



Outline

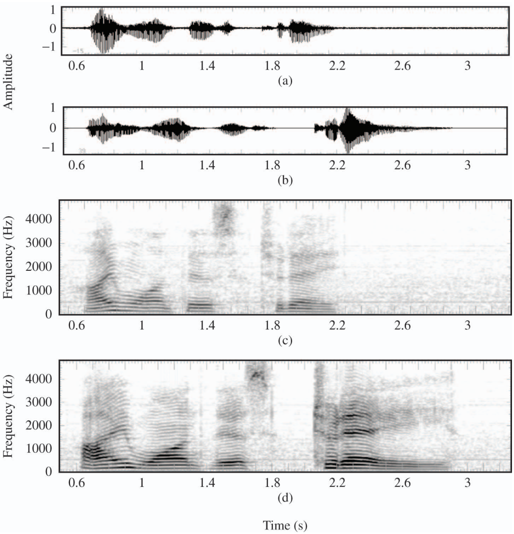
Prosody of speech

As prosody of speech we refer to:

- Rhythm
- Fundamental frequency contour (pitch)
- Loudness

Stressed speech

"Please do this today":



Outline

Perception of Speech

?

Outline

Acknowledgments

Most, if not all, figures in this lecture are coming from the book:

T. F. Quatieri: Discrete-Time Speech Signal Processing, principles and practice
2002, Prentice Hall

and have been used after permission from Prentice Hall